SEQUENCING AND CHARACTERIZATION OF THREE PLASMIDS FROM *HELICOBACTER PYLORI* STRAIN AL236

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

BY

JESSE REEVES, B.A.

DENTON, TEXAS

MAY 2010

TEXAS WOMAN'S UNIVERSITY
DENTON, TEXAS

April 7, 2010

To the Dean of the Graduate School:

I am submitting herewith a thesis written by Jesse Reeves entitled "Sequencing and Characterization of Three Plasmids from *Helicobacter pylori* Strain AL236." I have examined this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Biology.

Dr. Sarah A. McIntire, Major Professor

We have read this thesis and recommend its acceptance:

Department Chair

Accepted:

Dean of the Graduate School

DEDICATION

To my parents, Virginia and Jesse Reeves.
Thanks for all of your love, support, and encouragement.

To my siblings, Bridget, Theresa, and Jimmy.
Thanks for your love and support.

To my grandmother, Juanita Belle Parr Reeves.
Thanks for all of your love, support, and prayers.

To my grandfather, Billy Ray Reeves.
September 23, 1928 – October 27, 1993.
Thank you.

To my grandfather, Dr. Paul Lawrence Flynn.
February 28, 1916 – November 9, 1982.
Thank you.

To my grandmother, Virginia Elizabeth Horsefield Flynn.
August 31, 1919 – October 20, 1981.
Thank you.

To my first advisor at TWU, Dr. Sterling Smith.
1940 – 2006.
Thank you.

ACKNOWLEDGMENTS

I would like to acknowledge several individuals who have contributed to the completion of this thesis. I gratefully acknowledge Dr. Sarah McIntire, my mentor, for her continuous guidance, encouragement, and for not giving up on me. A special appreciation is also extended to my committee members, Dr. Heather Conrad-Webb and Dr. DiAnna Hynds. Thanks for everything. I would like to acknowledge Dr. John Knesek and Nani Cooper for their support and guidance with the sequencing strategy, which allowed me to obtain the complete DNA sequence of all three plasmids. Thanks to Dr. Brian Beck for his guidance with the phylogenetic analysis. Thanks also to Dr. Ilse Rickets for providing me with the plasmids utilized in this study.

A special thanks to Reta Foreman and Dr. Shazia Ahmed for their expertise and guidance during my time as a teaching assistant. Even though he is no longer with us I will always be grateful for having a chance to know Dr. Sterling Smith. When I decided that I wanted to change my career and go into biology Dr. Smith was the first person I met at Texas Woman's University. He took time out of his schedule and mapped out what I needed to do to obtain a degree in biology.

Other individuals that I would like to acknowledge who have had an influence on my life and have helped me reach my goals include: Coach Don Mauro; Coach David Bolstorff; Coach Michael Skoglund; Paulette Church; Dr. Donna Cain; Jeff Meyerholz; and numerous other lab-mates, friends, and family. Thanks for all of your support.

ABSTRACT

JESSE REEVES

SEQUENCING AND CHARACTERIZATION OF THREE PLASMIDS FROM *HELICOBACTER PYLORI* STRAIN AL236

MAY 2010

Helicobacter pylori, a microaerophilic, spiral-shaped, Gram negative bacterium is a gastrointestinal pathogen. Plasmids from *H. pylori* are currently being sequenced and characterized in our laboratory to determine their role in *H. pylori.* A previous study indicated unexpectedly that *H. pylori* strain AL236 might contain three separate plasmid species. This observation was based on the insertion of the EZ-Tn5™ <R6Kγ*ori*/KAN-2> transposon into the total AL236 plasmid DNA, which yielded recombinant plasmids pAL236-2, pAL236-5, and pAL236-11. In this study, all three plasmids were sequenced and characterized.

Utilizing forward and reverse primers specific for the EZ-Tn5™ <R6Kγ*ori*/KAN-2> transposon, initial plasmid DNA sequence was obtained. To determine the remaining DNA sequences, unlabeled M13 forward and reverse tailing primers were designed for each recombinant plasmid and utilized to produce PCR products containing M13 primer sites. Sequencing of these PCR products provided the remaining DNA sequence for each plasmid from strain AL236. Results indicated that pAL236-2 is 1448 bp, pAL236-5 is 1216 bp, and pAL236-11 is 3148 bp.

Each plasmid sequence was submitted to BLAST at NCBI, and was reported to have identity with other *H. pylori* plasmids. Results also indicated the replication proteins of plasmid pAL236-2 and pAL236-5 had identity with *H. pylori* plasmids that replicate via the rolling-circle mechanism. For plasmid pAL236-11, results indicated the replication protein, RepB, had identity with *H. pylori* plasmids that replicate via the theta mechanism. Computer analysis of the plasmid sequences was used to determine open reading frames (ORF), promoter, and ribosomal binding sites (RBS).

Phylogenetic trees were constructed through ClustalW2 of known *H. pylori* plasmid replication proteins. Results showed the existence of three classes of *H. pylori* replication proteins which included RepA, RepB, and RepH. Both RepA and RepB are involved with theta replication while RepH is involved with rolling-circle replication.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures Page

CHAPTER I

INTRODUCTION

In 1979 Dr. Robin Warren observed curved bacteria in gastric biopsy specimens submitted for histological examination (Warren and Marshall, 1983). Dr. Barry Marshall became interested in Dr. Warren's observations and together they utilized methods for the isolation of the organism based on the growth requirements of *Campylobacter* species (Marshall and Warren, 1984). The isolated organism was identified as a new *Campylobacter* species and given the name *C. pylori*. Further analysis based on the sequencing of 16S rRNA genes, enzymatic activities, fatty acid profiles, growth characteristics, nucleic acid hybridization profiles, and the observation that *C. pylori* was urease positive, unlike the other *Campylobacter* species, provided evidence that *C. pylori* belonged in a different genus (Goodwin *et al.*, 1989; Paster *et al.*, 1991; Vandamme *et al.*, 1991). As a result, *C. pylori* was placed in a new genus, *Helicobacter,* and renamed *Helicobacter pylori* (Goodwin *et al.*, 1989). Many historical aspects of *Helicobacter pylori* have been reviewed (Versalovic *et al.*, 2001).

H. pylori, a Gram negative, motile, microaerophilic, spiral-shaped bacterium colonizes the human stomach (Dunn *et al.*, 1997; Karianne *et al.*, 2005). With over 20 years of research since Warren and Marshall first reported *H. pylori* in 1983, no reservoir has been identified except the human stomach and little is known about the modes of transmission or how this organism is spread (Dunn *et al.*, 1997; Schwarz *et al.*, 2008).

Research indicates that 50% of the population is colonized with *H. pylori* worldwide (Torres *et al.*, 2000). Various reports show that infection with *H. pylori* can vary according to geographic areas, socioeconomic status, and can persist for a lifetime of the host despite immune and inflammatory responses, turnover of the gastric epithelial cells and the overlying mucus layer where *H. pylori* is found (Brown, 2000; Blaser *et al.*, 2001; Kersulyte *et al.*, 2004; Kusters *et al.*, 2006). Indeed, infection with *H. pylori* is higher in underdeveloped than in developed countries (Dubois, 1995; Dunn *et al.*, 1997; Brown, 2000; Schwartz *et al.*, 2008). Even though the exact route of *H. pylori* transmission is not clear, there have been reports of a possible link between direct person-to-person transmission by either the fecal-oral or oral-oral routes or both (Dubois, 1995; Dunn *et al.*, 1997; Brown, 2000; Kusters *et al.*, 2006; Schwarz *et al.*, 2008).

H. pylori infection has been linked to gastritis, duodenal and gastric ulcers, and to certain types of gastric cancers, including gastric adenocarcinoma and mucosa-associated lymphoid tissue (MALT) lymphoma (Dubois, 1995; Chalker *et al.*, 2001). *H. pylori* utilizes a variety of virulence factors including urease production, vacuolating cytotoxin A (VacA), cytotoxin-associated gene (CagA), and the *cag* pathogenicity island (*cag* PAI) to induce gastric inflammation, disruption of the gastric mucosal barrier, and alter gastric homeostasis (Dunn *et al.*, 1997; Selbach *et al.*, 2002). These virulence factors also include motility utilizing four to six sheathed flagella allowing *H. pylori* to penetrate the gastric mucosa and colonize the gastric epithelium during infection (Karianne *et al.*, 2005). *H. pylori* encode several chemoreceptors which sense environmental conditions to direct the movement of the flagella. Motility and chemotaxis are believed to be

necessary for *H. pylori* during the initial stages of infection. *H. pylori* cells produce large quantities of urease which hydrolyzes urea to produce ammonia and carbon dioxide. Ammonia raises the pH around each cell allowing *H. pylori* to survive in the acidic environment of the stomach (Fan *et al.*, 2000).

Integrated into the *H. pylori* genome is the 40-kb *cag* pathogenicity island (*cag* PAI), containing 27 to 31 genes (Couturier *et al.*, 2006). The first identified protein from the *cag* PAI is encoded by a gene called the cytotoxin-associated gene (*cag*A) (Oliveira *et al.*, 2003). Further studies of CagA showed that *H. pylori* translocates the CagA protein into cultured human gastric epithelial cells (AGS) utilizing the type IV secretion system encoded by other genes of the *cag* PAI (Higashi *et al.*, 2002). Once inside, CagA protein is tyrosine phosphorylated and involved with several signal transduction pathways leading to changes in structure, function, and morphology of the AGS cells (Bourzac *et al.*, 2007).

Another important virulence factor thought to be involved in generating gastric epithelial damage is the vacuolating cytotoxin A (VacA). VacA toxin is initially translated as a 140 kDa protein, and later modified into an 87 kDa secreted toxin (Nguyen *et al.*, 2001). This toxin reaches the plasma cell membrane and forms channels that are subsequently endocytosed and sent to the late endosomes (Gauthier *et al.*, 2005). VacA exerts its action in the cytoplasm of AGS cells, inducing the formation of large cytoplasmic vacuoles (Terebiznik *et al.*, 2006). These vacuoles grow in size to fill the entire cytosol, eventually leading to the death of cells by apoptosis (Cover *et al.*, 2003).

Even though CagA is frequently coexpressed with VacA, *vac*A gene expression is independent of expression of the *cag*A gene, which has led to dividing *H. pylori* isolates into type I or type II strains (Censini *et al.*, 1996). Type I strains, associated with a more aggressive course of infection, contain the *cag* PAI, CagA and functional VacA (Nilsson *et al.*, 2003). Type II strains lack CagA, carry a nontoxic form of VacA, and are regarded as less virulent (Nilsson *et al.*, 2003). Strains of *H. pylori* containing the *cag* PAI mark an important event in the evolution of this bacterium by allowing the organism to become more virulent compared to other organisms in this genus (Censini *et al.*, 1996).

H. pylori strains from around the world are very diverse genetically (Kersulyte *et al.*, 1998; Blaser *et al.*, 2001; Israel, *et al.*, 2001; Kersulyte *et al.*, 2002; Kersulyte *et al.*, 2004; Kalia *et al.*, 2004). Sources for this diversity include point mutations, substitutions and deletions, chromosomal inversions, recombination, and plasmids (Kleanthous *et al.*, 1991; Tomb *et al.*, 1997; Alm *et al.*, 1999; Kersulyte *et al.*, 1998; Blaser *et al.*, 2001). Another possible source for the diversity of this organism could be the result of adaptation by *H. pylori* within a host during years of infection and horizontal gene transfer (Suerbaum *et al.*, 1998; Kersulyte *et al.*, 1999; Blaser *et al.*, 2004). Insertion sequences (IS) have been implicated in contributing to the diversity of *H. pylori* (Kersulyte *et al.*, 1998; Kersulyte *et al.*, 2000; Kersulyte *et al.*, 2002; Kersulyte *et al.*, 2004; Kalia *et al.*, 2004). These IS elements are transposable DNA segments that can randomly insert into genomic or plasmid DNA. Several IS elements are found in *H. pylori* strains and plasmids in single or multiple copies. So far there have been five IS elements identified in *H. pylori*. These include: IS*605* (Kersulyte *et al.*, 1998), IS*606*

(Kersulyte *et al.*, 1998), IS*607* (Kersulyte *et al.*, 2000), IS*Hp608* (Kersulyte *et al.*, 2002), and IS*Hp609* (Kersulyte *et al.*, 2004).

Four complete *H. pylori* genome sequences have been reported: J99 (Alm *et al.*, 1999); 26695 (Tomb *et al.*, 1997); HPAG1 (Oh *et al*, 2006); and G27 (Baltrus *et al.*, 2009). Although *H. pylori* was believed to exhibit a large degree of genomic and allelic diversity, each of these genomes appear to be similar in size, gene composition, predicted proteomes (set of proteins encoded by the genomes) and appear to have a similar GC % between 38-39% (Tomb *et al.*, 1997; Alm *et al.*, 1999; Oh *et al*, 2006; Baltrus *et al.*, 2009). Additional unpublished genomes of *H. pylori* include strains P12 (NC_011498); Shi470 (NC_010698); B128 (NZ_ABSY00000000); HPKX_438_AG0C1 (NZ_ABJO00000000); and HPKX_438_CA4C1 (NZ_ABJP00000000) also appear to have a GC % between 38-39%. Further analysis indicated J99, 26695, Shi470, B128, HPKX_438_AG0C1, and HPKX_438_CA4C1 do not carry a plasmid. However, strains HPAG1, G27, and P12 each carry a plasmid, pHPAG1, pHPG27, pHPP12 (unpublished, NC_011499) respectively (Oh *et al*, 2006; Baltrus *et al.*, 2009).

Approximately eighty-five percent of *H. pylori* strains carry a plasmid with sizes ranging from 1.8 to 40 kilobase pairs (kbp) (Minnis *et al.*, 1995). Plasmids are double-stranded circular DNA molecules that replicate independently of the chromosomal DNA. Several *H. pylori* plasmids have been sequenced and characterized: pAL202 (Rickets, 2004); pAL226 (Khatun, 2005); pHPM8 (Quiñones, 2000); pHPM179 (Qasem, 1995; Taylor, 1997); pHPM180 (Minnis *et al.*, 1995); pHPM185 (Liu, 1998; Gao, 1999); pHPM186 (Burnham, 1998); pHel1 (Heuermann *et al.*, 1995); pHel4 (Hofreuter *et al.*,

2002); pHel5 (Hofreuter *et al.*, 2002); pHP69 (unpublished, NC_010884); pHP666 (unpublished, NC_010932); pHPAG1 (Oh *et al.*, 2006); pHPG27 (Baltrus *et al.*, 2009); pHPP12 (unpublished, NC_011499); pHPS1 (De Ungria *et al.*, 1999); pHP51 (Song *et al.*, 2003); pHPO100 (unpublished, AFO56496); pKU701 (Hosaka *et al.*, 2002); pHP489 (Song *et al.*, 2003); and pHPK255 (Kleanthous *et al.*, 1991). The specific role(s) of plasmids in *H. pylori* has not been identified. Accession numbers and references of the plasmids cited in this study are discussed later.

Plasmids from *H. pylori* vary from strain to strain. These variations include size, gene content, and restriction digestion profiles (Rickets, 2004; Khatun, 2005). The only gene that seems to be conserved among all of these plasmids is the replication (*rep*) gene that produces the replication (Rep) protein that is necessary for plasmid replication. The majority of plasmids from *H. pylori* replicate by the theta mechanism. Theta replicons include: pAL202; pAL226; pHPM8; pHPM179; pHPM180; pHPM185; pHPM186; pMCU1 through pMCU7; pHel1; pHel4; pHel5; pHP69; pHP666; pHAG1; pHPG27; pHPP12; pHPS1; pHP51; pHPO100; and pKU701. A small portion of *H. pylori* plasmids replicate by the rolling-circle mechanism, including replicons pHP489 and pHPK255.

Based on their replication proteins, *H. pylori* plasmids that replicate by the theta mechanism can be divided into two groups. The first group of plasmids replicate via the RepA protein. These plasmids include: pAL202; pAL226; pHPM8; pHPM179; pHPM180; pHPM185; pHPM186; pMCU1 through pMCU7; pHel1; pHel4; pHel5; pHP69; pHP666; pHAG1; pHPG27; pHPP12; and pHPS1. Plasmid pHPM179 can be divided into pHPM179a and pHPM179b as a result of pHPM179b missing a 1.5 kbp

fragment (Qasem, 1995; Taylor, 1997). However, pHPM179a and pHPM179b have the same RepA protein. De Ungria *et al.* (1998) utilized primers to obtain the middle regions of the *repA* gene for pMCU1 through pMCU7. As a result, these plasmids have not been completely sequenced, and the RepA proteins of these plasmids are missing the N-terminus as well as the C-terminus, but are assumed to replicate by the theta mechanism.

The size for the RepA proteins of pMCU1 through pMCU7 varies from 23.3 kDa to 24.5 kDa. However, plasmids: pAL202, pAL226, pHPM8, pHPM179, pHPM180, pHPM185, pHPM186, pHel1, pHel4, pHel5, pHP69, pHP666, pHAG1, pHPG27, pHPP12, and pHPS1 have a complete RepA protein with sizes that range from 54.5 kDa to 64.4 kDa. Regardless of the size of these RepA proteins they all show a strong conserved middle region while the amino acid sequences vary at the N-terminus and C-terminus (Rickets, 2004). A unique plasmid in this group is pHPAG1 that has two *repA* genes that each encodes a RepA protein (Oh *et al.*, 2006). These two proteins are not the same sequence and vary in size from 59.8 kDA to 64.4 kDa.

Another common feature of the theta replicons with RepA is a repeat sequence that is found upstream of the *repA* gene called the iterons where RepA binds to initiate plasmid replication. These iterons (R1) of pAL202, pAL226, pHPM8, pHPM179, pHPM180, pHPM185, pHPM186, pHeL1, pHel4, pHel5, pHP69, pHP666, pHPP12, and pHPS1 consist of four 22 base pair (bp) tandem direct repeat sequences. The only exception of this is found on pHPG27 which has four 23 bp tandem repeat sequences that are similar to other R1 sequences. Plasmid pHPAG1 has a complete set of the R1 region for one of the *repA* genes that encodes RepA identified as ORF (P003). However for the

second *repA* gene that encodes RepA identified as ORF (P001), the R1 region consists of only three 22 bp tandem repeat sequences.

The second group of plasmids that replicate by the theta mechanism include plasmids pHPS1, pHP51, pHPO100, and pKU701 that replicate via the RepB protein. One plasmid in this group is pHPO100 which has a replication protein identified as RepA. However, it appears this protein shares more identity with the RepB proteins than with the RepA proteins (this study). Besides pHPO100, the other plasmids have a *repB* gene which encodes the RepB protein involved with plasmid replication. The *repB* gene for *H. pylori* plasmids was first described for pHPS1 which has both *repA* and *repB* genes (De Ungria *et al.*, 1999). The sizes for RepB of pHPS1, pHP51, pKU701, and RepA of pHPO100 vary from 35.6 kDa to 51.3 kDa.

A common feature of the RepB theta replicons is an iteron sequence that is found upstream of the *repB* gene. The iterons for RepB are referred to as R3 (De Ungria *et al.*, 1999 and Hosaka *et al.*, 2002). Like R1, the R3 iterons are believed to be the binding site of RepB protein for initiation of plasmid replication. Song *et al.* (2003) identified the iteron region of pHP51 as R1; however, the R1 of pHP51 appears to be more similar to R3 iterons than the R1 iterons found with plasmids that replicate via RepA, with the exception being pHPO100 (this study). The R3 iterons of pKU701 and pHPS1 consist of a 3.5 time 33 bp direct tandem repeat. While plasmid pHP51 R1 iteron region consists of a 4 time 27 bp repeat with 5 bp spacer sequence between each repeat. Lastly, the R3 iterons of pHPO100 consists of a 3.5 time 21 bp repeat with 11-12 bp spacer sequences between each repeat.

All the *H. pylori* plasmids that utilize theta mechanism have areas of interest that can be found on some but not all of these plasmids. Plasmids pHPM8, pHPM180, pHPM185, pAL202, pAL226, pHPO100, pHel4, pHPS1, pHP51, pMCU1, pMCU2, pMCU4, pMCU5, pHP666, pHP69, pHPAG1, pHPG27, and pHPP12 contain a conserved protein termed ORF2. The *orf2* gene is positioned downstream of the *rep* gene of pHPM8, pHPM180, pHPM185, pAL202, pAL226, pHPO100, pHel4, pHPS1, pHP51, pHP666, pHP69, pHPAG1, pHPG27, and pHPP12. The *orf2* sequences of pMCU1, pMCU2, pMCU4, and pMCU5 were obtained using primers (De Ungria *et al.* 1998) like the *repA* gene of pMCU1 through pMCU7.

Each of these ORF2 proteins with the exception of pHPS1 contains a conserved domain HPFXXGNG found with proteins that are members of the cAMP-induced filamentation (*fic*) gene (Utsumi *et al.*, 1982; Komano *et al.*, 1991). De Ungria *et al.* (1999) stated that ORF2 had 95% and 88% identity to the ORF2 of pHPM180 and pHPO100 respectively; however they also mention that a 99-bp sequence found in pHPM180 and pHPO100 is missing from pHPS1. The missing region of pHPS1 included the HPFXXGNG conserved domain that is found with the other ORF2 proteins of *H. pylori* plasmids (this study). Entering ORF2 of pHPS1 into BlastP (Altschul *et al.*, 1997 and Altschul *et al.*, 2005) through NCBI revealed ORF2 of pHPS1 did have a conserved domain (COG3177) which includes members of the Fic family of proteins (Marchler-Bauer *et al.*, 2004 and Marchler-Bauer *et al.*, 2009).

Both the Fic family of proteins and cAMP are involved in cell division via the synthesis of folate in *Escherichia coli* (Utsumi *et al.*, 1982; Komano *et al.*, 1991). Some

suggestions have been made for the possible role of ORF2 in *H. pylori* plasmids. De Ungria *et al.* (1999) suggested that ORF2 might be involved with recombination while Quiñones, (2000) suggested that ORF2 might be associated with replication of the plasmids or control of copy number of plasmids in *H. pylori.*

Also included in some but not all of theta replicons of *H. pylori* are repeat regions termed R2, first identified in pHPM180 (Minnis *et al.*, 1995). Minnis *et al.* (1995) defined R2 as a two time 232 bp direct repeat separated by 36 bp. Examples of plasmids that have been identified as having partial or complete R2 repeats include pAL202, pAL226, pHel4, pHel5, pHPM180, pHP51, pHPS1, pHPP12 and pHPAG1. For plasmids pHPP12 and pHPAG1 the potential R2 repeats were identified in this study. The exact role of the R2 has been suggested to be involved with plasmid recombination events (Minnis *et al.*, 1995; Hofreuter *et al.*, 2002).

Besides R2 there was an additional repeat region first observed in pHel4 and pHel5 and was termed R4 (Hofreuter *et al.*, 2002). R4 repeats are defined as 36 bp located at the 3' end of the *repA* gene with another copy separated by 1.7 kb from the first R4. Not all the plasmids contain two copies of R4 like pHel4; however there are plasmids that only have one R4 at the 3' end of the *repA* gene. Examples of plasmids that have been identified with complete or partial R4 repeats include pHel1, pHel5, pAL202, pAL226, pHPM180, pHPM186. Like the R2 repeats the exact role of the R4 repeats is suggested to have involvement with plasmid recombination events (Hofreuter *et al.*, 2002).

As stated previously, many *H. pylori* strains contain one or more copies or partial copies of insertion sequences (IS) IS*605*, IS*606*, IS*607*, IS*Hp608*, and IS*Hp609*. These IS sequences are not only found on the genomes of *H. pylori* but also on some of the *H. pylori* plasmids that replicate by the theta mechanism as well. *H. pylori* plasmids that have partial or whole insertion sequences, and chromosomal DNA include pHPM8, pHPM179, pHPM180, pHPM185, pHP186, pAL202, pAL226, pHel1, pHel4, pHel5, pKU701, pHPS1, pKU701, and pHP51. As a result, insertion sequences could be the result of recombination events between plasmid and genome DNA of *H. pylori* or the transfer of DNA between strains (Alm *et al.*, 1999).

Besides theta replication another type of replication known as rolling-circle (RC) replication does exists in *H. pylori* plasmids. Plasmid pHPK255 was the first identified *H. pylori* plasmid that replicates by the RC mechanism that is common for plasmids in Gram positive bacteria (Kleanthous *et al.*, 1991). Unlike the plasmids that replicate by the theta mechanism, this plasmid also does not contain ORF2, R1, R2, R3, R4, and does not contain any insertion sequences. Even the replication protein is not the same as the plasmids that replicate by the theta mechanism. Instead, pHPK255 has a single open reading frame (ORF) that encodes a Rep protein that is 216 amino acids (aa) or 25.5 kDa. This Rep protein is called a "replication-initiation protein homolog", not RepA or RepB.

Instead of iteron sequences, like R1 and R3, pHPK255 contains inverted repeats that have the ability to form hairpin-like structures that contain a plus (+) origin of replication known as the double-stranded origin (DSO) that contains a nick site involved with leading strand synthesis. Located downstream of this hairpin-like structure are two

7 bp direct repeats which are possibly involved with binding the Rep protein. When the hairpin-like structure is formed the nick site protrudes from the stem, allowing the Rep protein to nick the (+) strand to initiate plasmid replication. Another origin of replication called the minus origin of replication (*ori* (-)) known as a single-stranded origin (SSO) is also found in pHPK255 involved with lagging strand synthesis. To form this origin, imperfect inverted palindromes form another hairpin-like structure. Near the top of this structure is a consensus signal sequence. Upstream of the signal sequence are two 9 bp repeated sequences which may allow attachment of host cofactors for the conversion of single stranded DNA (ssDNA) to double stranded DNA (dsDNA), thereby completing one round of replication.

Lastly, plasmid pHP489 is another *H. pylori* plasmid that has one ORF, a Dna-A binding site, one direct repeat, and three inverted repeats (Song *et al.*, 2003). Like pHPK255 this plasmid does not contain ORF2, R1, R2, R3, R4, or insertion sequences. Plasmid pHP489 does contain two 57 bp repeats identified as iterons (Song *et al.*, 2003). However, it appears that these iterons are not the same as R1 and R3 found with theta replicons from *H. pylori.*

The ORF in pHP489 encodes the replication protein that is not RepA or RepB. However, it does share 57% identity to the Rep protein of pHPK255. Song *et al.* (2003) named this protein RepH where the H stands for *Helicobacter* and consists of 239 aa or 28.0 kDa. This protein contained two of the three conserved motifs present in the Rep proteins of plasmids that replicate by the rolling circle mechanism. Besides RepH, no other evidence was available to indicate that pHP489 replicated by the rolling circle

mechanism. Of particular interest is the observation that mutational studies of the RepH protein revealed that RepH is not required for replication in *H. pylori* strain 489 (Song *et al.*, 2003). The exact role of RepH is not clear, although Song *et al.* (2003) suggested that other genomic proteins may be involved with the replication of this plasmid.

Our lab characterizes *H. pylori* plasmids in order to determine their potential role in *H. pylori*. The initial 15 strains of *H. pylori* came from patients at the Dallas VA Medical Center. DNA restriction analysis of these plasmids indicated that 12 strains carried a single plasmid. Sequence analysis of these plasmids showed the presence of a *repA* gene that could regulate theta type replication. Since the characterized strains were obtained from patients in the same geographical area (Dallas/North Texas), another set of strains from a different geographical location were obtained from Drs D. Berg and A. J. Parkinson. These strains were isolated from Alaskan Indians, presumably from persons with no contact with the Dallas patients. North Texas plasmids that have been sequenced and characterized include: pHPM8, pHPM179, pHPM180, pHPM185, and pHPM186. Alaskan plasmids that have been sequenced and characterized include pAL202, pAL226, and pAL236. Strain AL236, was initially screened by Dr. I. Rickets using the EZ-Tn5™ <R6K*γori*/KAN-2> insertion kit (Rickets, 2004). Her results indicated a possibility that AL236 contained three separate plasmids. These plasmids containing the <R6K*γori*/KAN-2> transposon, were identified as pAL236-2; pAL236-5; and pAL236-11 (Rickets, 2004).

OBJECTIVES OF THIS STUDY:

1. Determine the DNA sequence of pAL236-2, pAL236-5, and pAL236-11 isolated from *H. pylori* strain AL236 and include analysis of open reading frames (ORFs), ribosome binding sites (RBS), and promoter sequences;
2. Compare the sequences of pAL236-2, pAL236-5, and pAL236-11 with other *H. pylori* plasmids to determine gene products that may indicate geographical differences among *H. pylori* isolates.

CHAPTER II

MATERIALS AND METHODS

MATERIALS

Bacterial Strains and Plasmids

H. pylori strain AL236 was isolated from an Alaskan Indian. Strain AL236 along with other Alaskan strains of *H. pylori* were provided by Douglas E. Berg, Washington University School of Medicine, St. Louis, MO, and Alan J. Parkinson, Centers for Disease Control and Prevention, Anchorage, AK (Rickets, 2004). AL236 was stored by Rickets at -80°C in freezer medium consisting of trypticase soy broth (Difco), containing 25% glycerol (Sigma) and 10% horse serum (Sigma). Rickets cultured the AL236 cells at 37°C, under microaerophilic conditions, 5-10% CO_2 on brain heart infusion medium (Difco), supplemented with 10% horse serum.

The EZ-Tn5™ <R6K*γori*/KAN-2> transposon (EPICENTRE®) was randomly inserted into plasmid DNA from AL236 by Rickets. Her results revealed the presence of three Kan^R recombinant plasmids, pAL236-2, pAL236-5, and pAL236-11 (Rickets, 2004). Each recombinant plasmid had been transformed into *Escherichia coli* pir^+ (low copy number) cells. The genotype of the pir^+ cells is F^- *mcr*A, Δ(*mrr-hsd*RMA-*mcr*BC), ϕ80*lacZ*ΔM15, Δ*lac*X74, *rec*A1, *end*A1, ara*ara*D139, Δ(*ara, leu*)7697, *gal*U, *gal*K, λ-, rpsL(Strr), nupG, pir^+. The *pir* gene product (π protein) is needed for replication of vectors containing R6K*γori* by controlling the copy number of plasmids. Transformants

containing the recombinant plasmids pAL236-2, pAL236-5, and pAL236-11 used in this study were obtained from Rickets. After transformation each of the *E. coli* cells containing the individual recombinant plasmids were stored by Rickets at -80°C with Protect™ Bacterial Preservers (Key Scientific Products).

Additional *H. pylori* plasmid DNA sequences were obtained from our lab data and from NCBI database (http://www.ncbi.nlm.nih.gov/). These plasmids are shown in Table 1.

Media

Luria-Bertani (LB) medium (Bio 101 systems) was prepared according to the manufacturer's recommendations. Transformants were grown on LB medium supplemented with 1.5% agar (Difco) and kanamycin sulfate (Amresco) at a final concentration of 50 μg/ml. *E. coli* strains containing the recombinant plasmids pAL236-2, pAL236-5, and pAL236-11 were stored at -80°C with CryoCare™ Bacterial Preservers (Key Scientific).

Buffers and Reagents

The buffers and reagents were: 5X and 1X Tris-Borate-EDTA (TBE) (Sigma); 70% and 95% ethanol (Fisher); 100% isopropanol (Fisher); urea (Amresco); agarose (SeaKem® GTG); ethidium bromide at 10 mg/ml (Sigma); Long Ranger® gel solution (Cambrex); ammonium persulfate (Amresco), kanamycin sulfate (Amresco); 10% acetic acid (Fisher), silane stock [0.5 ml γ-methacryloxypropyltri-methoxysilane (Sigma), in 100 ml of 100% ethanol], TEMED (N,N,N'N'-tetramethylethylenediamine) (Amresco);

Table 1. Accession Numbers and/or References of *H. pylori* Plasmids and Strains Cited in this Study. *Chromosomal DNA

Plasmid or Chromosome	Accession Number	Reference
pHPM8	AF275307	Quiñones, 2000.
pHPM179		Qasem, 1995. Taylor, 1997.
pHPM180	U12689	Minnis *et al.*, 1995.
pHPM185		Liu, 1998. Gao, 1999.
pHPM186	AF077006	Burnham, 1998.
pAL202	NC_005917	Rickets, 2004.
pAL226	DQ239897	Khatun, 2005.
pHP0100	AF056496	
pHei1	Z49272	Heuermann *et al.*, 1995.
pHel4	AF469112	Hofreuter *et al.*, 2002.
pHel5	AF469113	Hofreuter *et al.*, 2002
pHPS1	AF019894	De Ungria *et al.*, 1999.
pHP51	NC_004767	Song *et al.*, 2003.
pHPG27	NC_011334	Baltrus *et al.*, 2009.
pHPK255	AAB21515	Kleanthous *et al.*, 1991.
pHPP12	NC_011499	
pMCU1	AF019895 & AF055275	De Ungria *et al.*, 1998
pMCU2	AF019896 & AF055274	De Ungria *et al.*, 1998
pMCU3	AF019897	De Ungria *et al.*, 1998
pMCU4	AF019898 & AF055276	De Ungria *et al.*, 1998
pMCU5	AF019899 & AF055277	De Ungria *et al.*, 1998
pMCU6	AF019900	De Ungria *et al.*, 1998
pMCU7	AF019901	De Ungria *et al.*, 1998
pHPAG1	NC_008087	Oh *et al.*, 2006
pHP489	NC_001843	Song *et al.*, 2003.
pHP666	NC_010932	
pHP69	NC_010884	
pKU701	BAC06188	Hosaka *et al.*, 2002.
HPAG1*	NC_008086	Oh *et al.*, 2006
26695*	NC_000915	Tomb *et al.*, 1997
J99*	NC_000921	Alm *et al.*, 1999
G27*	NC_011333	Baltrus *et al.*, 2009
P12*	NC_011498	
Shi470*	NC_010698	
B128*	NZ_ABSY00000000	
HPKX_438_CA4C1*	NZ_ABJP00000000	
HPKX_438_AG0C1*	NZ_ABJO00000000	

BenchTop 1 kbp DNA ladder (Promega); tracking dye (5 ml 5X TBE, 5 ml glycerol, 0.01 gm bromophenol blue).

Enzymes

Restriction endonucleases were obtained from Promega Inc. The restriction endonucleases and other enzymes were used with buffers provided by, and as recommended by, the manufacturer.

Commercial Kits

The kits used were: QIAGEN Plasmid Midi Kit (QIAGEN Inc.); EZ-Tn5™ <R6K*γori*/KAN-2> transposon kit (Epicentre); Gene Amp XL PCR Kit (Applied Biosystems); Elu-Quik® DNA Purification Kit (Whatman); SequiTherm EXCEL™ II DNA Sequencing Kit-LC (EPICENTRE®).

Primers

Labeled primers used in initial plasmid DNA sequencing and PCR product sequencing were obtained from LI-COR Inc. (Table 2). Unlabeled EZ-Tn5™ <R6K*γori*/KAN-2> transposon primers used for producing PCR products were provided with the EZ-Tn5™ <R6K*γori*/KAN-2> transposon kit (EPICENTRE®). The computer program, Vector NTI® Advance™ 10 and 11(InforMax®), was used to design M13 tailing primers used in the production of PCR products that were utilized for obtaining the complete DNA sequence of pAL236-2, pAL236-5, and pAL236-11. These unlabeled M13 tailing primers for producing PCR products were obtained from BioSynthesis Inc., Lewisville, TX. Sequences of all unlabeled primers are shown in Table 3.

Table 2. Primers used for DNA Sequencing. Primers used for the initial DNA sequencing of pAL236-2, pAL236-5, and pAL236-11 were obtained from LI-COR Inc. and labeled with a 5'-IRDye® 800. Primers used for PCR product sequencing were obtained from LI-COR Inc. The M13 Forward primer was labeled with a 5'-IRDye® 800. The M13 Reverse primer was labeled with a 5'-IRDye® 700.

Initial DNA Sequencing	Primer Sequence (Labeled)
KAN-2 FP-1 Forward	5'-ACCTACAACAAAGCTCTCTCATCAACC-3'
R6KAN-2 RP-1 Reverse	5'-CTACCCTGTGGAACACCTACATCT-3'
PCR Sequencing	
M13 Forward	5'-CACGACGTTGTAAAAGGA-3'
M13 Reverse	5'-GGATAACAATTTCACAGG-3'

Table 3. Unlabeled M13 PCR Primers that were used to Obtain M13 PCR Products. Each primer is underlined as either forward (F) or Reverse (R) to indicate which strand the primer annealed to on pAL236-2, pAL236-5, and pAL236-11. Underlined sequences represent M13 forward and reverse primer sites used for obtaining the complete DNA sequence of pAL236-2, pAL235-5, and pAL236-11. Approximate PCR product sizes are indicated in base pairs (bp). Primers were obtained from either Biosynthesis, Inc., or EPICENTRE®.

PCR Product	Primers (Unlabeled)	~Size (bp)
pAL236-2 #1	F-5'-CACGACGTTGTAAAAGGACGCAAATCTCACTCGTTCAGC-3' R-5'-CTACCCTGTGGAACACCTACATCT-3'	1000
pAL236-2 #2	R-5'-CACGACGTTGTAAAAGGACGCTGAACGAGTGAGATTTGC-3' F-5'-ACCTACAACAAAGCTCTCATCAACC-3'	1200
pAL236-5 #1	F-5'-CACGACGTTGTAAAAGGACCTCTAGGTGAGGGGTCGTAGTAG-3' R-5'-CTACCCTGTGGAACACCTACATCT-3'	1000
pAL236-5 #2	R-5'-CACGACGTTGTAAAAGGACCTACTACGACCCCTCACCTAGAG-3' F-5'-ACCTACAACAAAGCTCTCATCAACC-3'	1000
pAL236-11 #1	F-5'-CACGACGTTGTAAAAGGACCAAACTCATAGGGCTTGCGC-3' R-5'-GGATAACAATTTCACAGGCGTTATTTGAGTTTCACAGC-3'	1500
pAL236-11 #2	R-5'-GGATAACAATTTCACAGGGCGCAAGCCCTATGAGTTTG-3' F-5'-ACCTACAACAAAGCTCTCATCAACC-3'	1400
pAL236-11 #3	F-5'-CACGACGTTGTAAAAGGACGCTGTGAAACTCAAATAACG-3' R-5'-CTACCCTGTGGAACACCTACATCT-3'	1500
pAL236-11 #4	F-5'-CACGACGTTGTAAAAGGACGGATCTCTTGGATATAGACCG-3' R-5'-CTACCCTGTGGAACACCTACATCT-3'	1000
pAL236-11 #5	F-5'-CACGACGTTGTAAAAGGACGATAGCTCTTTGTTAATTCCAG-3' R-5'-GGATAACAATTTCACAGGCAATGACCAAACACACAGCC-3'	800
pAL236-11 #6	F-5'-CACGACGTTGTAAAAGGACGCGTTATTAAGGTGAATTGTAGC-3' R-5'-GGATAACAATTTCACAGGGTTGATCAGTAGTGAAGTCGC-3'	1000
pAL236-11 #7	R-5'-GGATAACAATTTCACAGGGCTCCCTTATTTAGCCG-3' F-5'-CACGACGTTGTAAAAGGACGCGTTATTAAGGTGAATTGTAGC-3'	1500
pAL236-11 #8	R-5'-GGATAACAATTTCACAGGGGCTTATGATGATGCTCC-3' F-5'-CACGACGTTGTAAAAGGACGCGTTATTAAGGTGAATTGTAGC-3'	1100

Lab Equipment

The equipment used in this study were: Spectrafuge Model 16M and 24D (Labnet); Beckman J2-HS centrifuge with a JS 7.5 rotor and the Beckman L8-70 ultracentrifuge with a type 30 rotor; Falcon® 50 ml polypropylene conical tubes; Shimadzu UV160U spectrophotometer; GeneAmp® PCR System Model 2400 and 2700 (Applied Biosystems); EASY-CAST™ Electrophoresis System (Owl Scientific, Inc.); Minne-the-Gel-Cicle™ Submarine Agarose Gel Unit (Hoefer); Gel XL Ultra™ Complete Compact Electrophoresis System (Labnet); RC 10.10. vacuum centrifuge (Jouan, Inc.); 302 nm UV transluminator (VILBER LOURMAT); LI-COR Automated DNA Sequencer model 4300. Agarose gels were visualized using Kodak Gel Logic 100 imaging system with an orange filter (Tiffen 15 orange).

Computer Programs

Individual plasmid DNA sequences were assembled utilizing the computer program AlignX® and ContigExpress® features on the Vector NTI® Advance™ 10 and 11(InforMax®). Vector NTI® Advance™ 10 and 11 was also utilized to design primers, analyze the plasmid DNA of pAL236-2, pAL236-5, and pAL236-11 for open reading frames, non-coding regions, ribosome binding sites, promoter sequences, repeat sequences, and GC content. Basic Local Alignment Search Tool (BLAST) algorithms BLASTN (Altschul *et al.*, 1990), BLASTP (Altschul *et al.*, 1997; Altshcul *et al.*, 2005), and BLASTX (Altschul *et al.*, 1997) by the National Center for Biotechnology Information database (NCBI) (http://www.ncbi.nlm.nih.gov/) were used to find identities

to other DNA and protein sequences in GenBank. Conserved domains of proteins were detected with the NCBI Conserved Domain Search (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (Marchler-Bauer *et al.*, 2004; Marchler-Bauer *et al.*, 2009). Open reading frames were further analyzed utilizing programs available through the internet through ExPASy Proteomics tools (http://us.expasy.org/): SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui/) (Hirokawa *et al.*, 1998; Mitaku *et al.*, 1999; Mitaku *et al.*, 2002); ProtParam (http://us.expasy.org/tools/protparam.html) (Gasteiger *et al.*, 2005); and TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0) (Tusnandy *et al.*, 1998; Tusnady *et al.*, 2001). Global multiple sequence alignments (MSA) of replication (Rep) protein amino acids from *H. pylori* plasmids was performed by the web server ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) (Larkin *et al.*, 2007). Phylogenetic trees were constructed as cladograms based on the average distance using percent identity by Jalview (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009) based on the MSA performed by ClustalW2.

METHODS

Bacterial Strains and Plasmids

H. pylori strain AL236 was isolated from an Alaskan Indian. Rickets initially cultured AL236 at 37°C under microaerophilic conditions, 5-10% CO_2, on brain heart infusion (BHI) medium (Difco), supplemented with 10% horse serum (Sigma). Rickets isolated the total plasmid DNA from AL236 and performed the initial EZ-Tn5™<R6K *γori*/KAN-2> (EPICENTRE®) transposon reaction, which yielded the three recombinant plasmids, pAL236-2, pAL236-5 and pAL236-11.

Transposon Reaction and Transformation

The EZ-Tn5™ <R6Kγ*ori*/KAN-2> Insertion Kit (EPICENTRE®) utilized by Rickets was used to insert the <R6Kγ*ori*/KAN-2> transposon randomly into the total plasmid DNA isolated from *H. pylori* strain AL236. The transposon insertion reaction was prepared as recommended by the manufacturer and incubated for 2 hours at 37°C. A stop solution was added and the tube placed in a 70°C water bath for 10 minutes and stored at -20°C until used.

After the transposon reaction was completed by Rickets the EZ-Tn5™ <R6Kγ*ori*/ KAN-2> transposon was randomly inserted into the total plasmid DNA of AL236 (Fig. 1). Taking advantage of the EZ-Tn5™ <R6Kγ*ori*/KAN-2> transposon containing the R6Kγ*ori*, origin of replication, allowed Rickets to transform *E. coli pir*$^+$ cells that would replicate the recombinant plasmids. Transformation of *E. coli pir*$^+$ cells with the recombinant plasmids was carried out by electroporation. After transformation Rickets screened and selected all transformants on either NZYM or LB agar containing kanamycin (50 µg/ml) by utilizing the kanamycin resistance gene present on the EZ-Tn5™ <R6Kγ*ori*/KAN-2> transposon. Analysis of the transformants by Rickets revealed the presence of three recombinant plasmids from AL236. Each recombinant plasmid was named according to the isolate number of *E. coli pir*$^+$ cells that contained one of the recombinant plasmids. As a result, each recombinant plasmid was named pAL236-2, pAL236-5, and pAL236-11. Rickets stored each transformant containing the recombinant plasmids in CryoCare™ Bacterial Preservers (Key Scientific) at -80°C. These plasmids were provided by Rickets and used for this study.

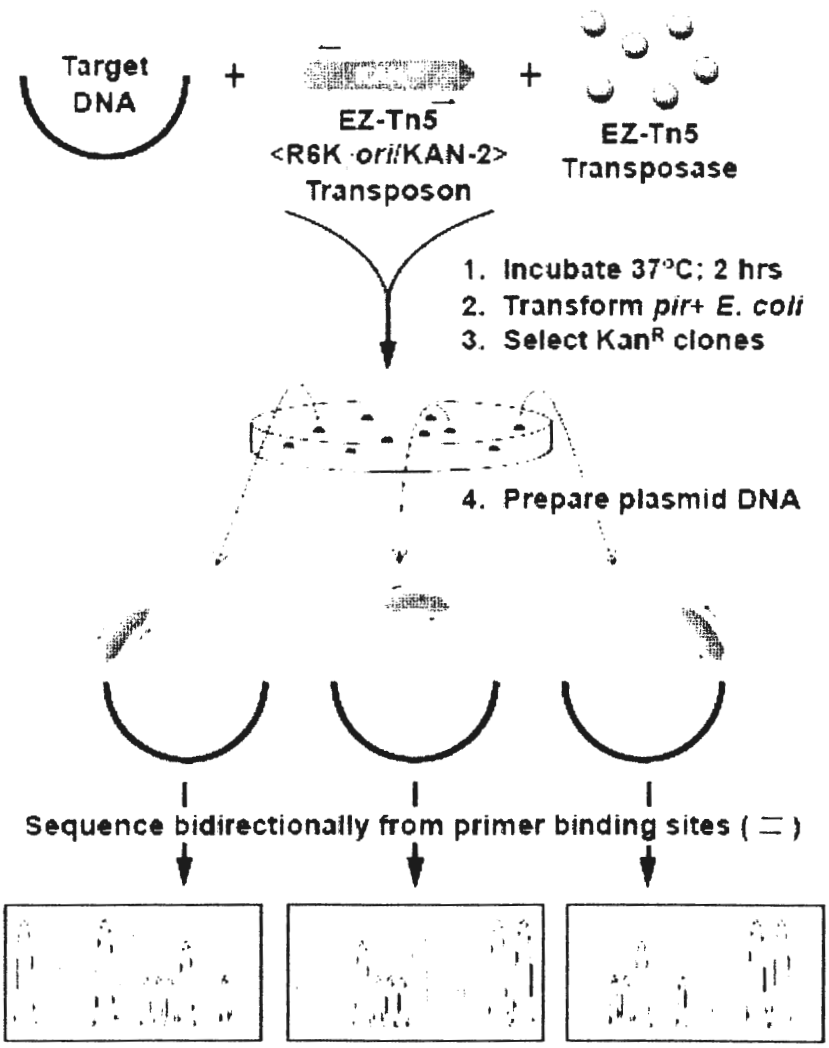

Figure 1. EZ-Tn5™ <R6Kγ*ori*/KAN-2> Transposon Insertion Protocol performed by I. Rickets to obtain recombinant pAL236 Plasmids. Image obtained from www.epicentre.com.

Plasmid Isolation and Purification

Large quantities of recombinant plasmid DNA were isolated and purified utilizing the QIAGEN Plasmid Midi Kit (QIAGEN Inc.). Each *E. coli pir*$^+$ transformant containing either pAL236-2, pAL236-5, and pAL236-11 were pulled from the CryoCare™ Bacterial Preservers and placed on separate LB agar plates containing kanamycin (50 μg/ml) and grown in an incubator at 37°C. The next day an individual colony for each *E. coli pir*$^+$ transformant containing the recombinant plasmid was inoculated into three separate 100 ml LB broth and placed in the 37°C incubator shaker overnight.

Each 100 ml LB broth culture was divided into two Falcon® 50 ml polypropylene conical tubes for a total of six tubes. Each 50 ml tube was placed in the Beckman J2-HS centrifuge with a JS 7.5 rotor pre-chilled to 4°C. The tubes were spun at 6000 rpm in the JS7.5 rotor for 15 minutes at 4°C. The supernatant was discarded and all traces of the supernatant were removed by inverting each tube and tapping them on paper towels.

After centrifugation the first pellets from each of the six tubes were resuspended in 4 ml of P1 resuspension buffer. The suspended solution from the first tube was transferred to a second 50 ml tube to resuspend the second pellet. The resuspended solutions were placed into a 30 type rotor tube for a total of three tubes each containing the transformant *E. coli pir*$^+$ cells that contained either pAL236-2, pAL236-5 and pAL236-11 recombinant plasmids. Each tube contained 4 ml of P2 lysis buffer and was inverted gently at 30 second intervals at room temperature for 5 minutes. Pre-chilled P3 neutralization buffer (4 ml) was added to each tube, the tube was placed on ice and

inverted at 2-3 minute intervals for 15 minutes. The tubes were placed in the Beckman L8-70 ultracentrifuge with a type 30 rotor, and spun at 25K rpm at 4°C for 15 minutes.

During centrifugation three QIAGEN tips were equilibrated by adding 4 ml of QBT buffer that was allowed to flow through via gravity. After centrifugation was completed, the supernatant from each tube was added to the QIAGEN tips and allowed to enter the resin and flow through via gravity. The tips were washed twice with 10 ml of QC buffer. The plasmid DNA was eluted from the QIAGEN tip with the addition of 5 ml of QF buffer. Each solution containing the plasmid DNA was split into 6 microcentrifuge tubes (833 µl/each) for a total of 18 tubes. Then 583 µl of isopropanol was added to each tube and vortexed to precipitate the recombinant plasmid DNA. The tubes were placed at 20°C overnight.

The next day each tube was placed into the Spectrafuge 16M or 24D (Labnet) model microcentrifuge and spun at 14K rpm for 30 minutes. The supernatant was removed and placed into fresh microcentrifuge tubes. Next, 333 µl of 70% ethanol at room temperature was added to each microcentrifuge tube. The tubes were placed back into the microcentrifuge and spun at 14K rpm for 10 minutes. The supernatant was discarded and each pellet was allowed to air dry for 5-10 minutes. Once the pellets were dried, 100 ml of warm UPH_2O was added to the first microcentrifuge tube of each set of tubes. This suspension was transferred though all six microcentrifuge tubes resuspending all pellets so that the last tube would contain all recombinant plasmid DNA. Overall, three separate tubes each contained one of the recombinant plasmids; pAL236-2, pAL236-5, and pAL236-11. The plasmid DNA concentration was determined using the

Shimadzu UV160U spectrophotometer (OD_{260} = 50 µg/ml). Plasmid DNA was stored at 4°C until used.

Restriction Endonuclease Analysis

The EZ-Tn5™ <R6K*γori*/KAN-2> transposon contains a single *Bam*HI restriction site at base pair (bp) 1939. Each restriction reaction contained restriction enzyme *Bam*HI and the buffer recommended by the manufacturer. To estimate the size of pAL236-2, pAL236-5, and pAL236-11 one unit of restriction endonuclease *Bam*HI was added to 250 ng of each recombinant plasmid DNA, 2 µl 10X reaction buffer, and UPH_2O to a final volume of 25 µl. Each restriction digest was incubated overnight in a 37°C in a water bath. Analysis of the recombinant plasmids was performed by agarose gel electrophoresis as described below.

Agarose Gel Electrophoresis

Analysis of recombinant plasmids and M13 PCR products (described below) were analyzed on 1% agarose gels. Agarose (SeaKem® GTG) was weighed in grams and added to either 40 ml or 60 ml of 1X TBE (Sigma) buffer and total weight was recorded. The solution was heated in a microwave to dissolve the agarose. After the agarose was dissolved, the solution was placed back on the scale and the original volume was obtained by the addition of UPH_2O. The liquid agarose solution was placed into a gel casting tray and allowed to solidify. The solidified gel was placed in a gel electrophoresis unit and 1X TBE buffer was added. The electrophoresis unit was connected to a power supply unit and allowed to run at 20 V/cm until the tracking dye traveled three-fourths the length of the gel. The gels were stained with ethidium bromide

(1 μg/ml) for 20-30 minutes at room temperature. Bands were visualized on the 302 nm UV transluminator (VILBER LOURMAT) and pictures were taken with Kodak Gel Logic 100 imaging system with an orange filter (Tiffen 15 orange).

Initial DNA Sequencing

Each recombinant plasmid pAL236-2, pAL236-5, and pAL236-11 had a single randomly inserted EZ-Tn5™ <R6K*γori*/KAN-2> transposon. Specifically labeled EZ-Tn5™ <R6K*γori*/KAN-2> transposon primers KAN-2 FP-1 Forward and R6KAN-2 RP-1 Reverse (LI-COR, Inc.), shown in Table 2, were used in bi-directional sequencing to determine the sequence near the transposon insertion site for each of the three recombinant plasmids (see Fig. 1). Each sequencing reaction utilized the SequiTherm EXCEL™ II DNA Sequencing Kit-LC protocol (EPICENTRE®) and LI-COR model 4300 Automated DNA Sequencer.

Each PCR sequencing reaction was performed with the SequiTherm EXCEL™ II DNA Sequencing Kit-LC. Each recombinant plasmid was placed in a separate sequencing reaction master mix in a 1.5 ml microcentrifuge tube. The master mix contained 8.10 μl 3.5X SequiTherm EXCEL™ II Sequencing Buffer, 2.25 μl labeled sequencing primer (1 pmol/μl), 7.65 μl of DNA (~250 ng) plus UPH_2O, and 1.12 μl of SequiTherm EXCEL™ II DNA Polymerase for a total volume of 19.12 μl. After the master mix was made, 2 μl of SequiTherm EXCEL™ II-LC Termination Mix A (532.5 pmol/μl), SequiTherm EXCEL™ II-LC Termination Mix T (720 pmol/μl), SequiTherm EXCEL™ II-LC Termination Mix G (570 pmol/μl), and SequiTherm EXCEL™ II-LC Termination Mix C (450 pmol/μl), were each placed into separate 0.2 ml PCR reaction

tubes. Each SequiTherm EXCEL™ II-LC Termination Mix contained all four deoxynucleotides; dATP, dCTP, dTTP, and dGTP along with the specific individual dideoxynucleotide; ddATP, ddCTP, ddTTP, and ddGTP. For each 2 µl of SequiTherm EXCEL™ II-LC Termination Mix, 4 µl of the master mix was added to each tube for a final volume of 6 µl. The PCR reaction tubes were placed in the GeneAmp® PCR System 2400 or 2700 models and allowed to run for 25 cycles. PCR amplification started with an initial 1 min at 94°C; 25 cycles of denaturing at 94°C for 1 min and 30 sec, annealing at 55°C for 1 min and 30 sec, elongation at 72°C for 5 min; and held ∞ at 4°C. Next, 3.0 µl of stop solution was added to each tube for a total volume of 9 µl. The tubes were heated in the PCR machine at 95°C for 3 min. After 3 min the PCR machine kept the tubes at 4°C until the samples were ready to load into the sequencing gel.

In preparation for sequencing each front plate and back plate was washed with 100% isopropanol and UPH_2O three times on each side. After the plates were cleaned 330 µl of silane stock and 10 µl of 10% acetic acid were mixed in a microcentrifuge tube and placed about 1 inch between the ears on the front sequencing plate to help prevent the sequencing gel from slipping down between the plates when placed vertically in the DNA sequencer. Each plate was allowed to dry for 30 minutes before the plates were sandwiched together and the gel solution was poured.

Each gel was a 66 cm denaturing polyacrylamide gel that contained 4 % polyacrylamide and 8 M urea. The gel solution was prepared by placing a plastic bottle with a nozzle lid on a scale that was zeroed. Next, 25.2 g of urea (Amresco), 4.8 ml Long Ranger® 50% polyacrylamide (Cambrex), 14.4 ml 5X TBE, and UPH_2O was added into

the plastic bottle for a total weight of 67.5 g. The nozzle lid was screwed on top of the plastic bottle containing the gel solution and was kept at room temperature overnight. After the urea was dissolved, 400 µl of fresh 10% ammonium persulfate and 40 µl of TEMED (Amresco) were added to the gel solution. The solution was immediately poured in between the assembled sequencing plates and allowed to polymerize for 3 hr.

Once polymerized, the gel was placed vertically inside of the DNA sequencer and the rest of the gel apparatus was assembled. For the bottom buffer tray 125 ml of 1X TBE buffer was added. Next, 600 ml of 0.75X TBE buffer was added to the upper buffer tray. The DNA sequencer laser was focused and all pre-run conditions were completed. After the pre-run, 2.0 µl of each sample was loaded and the gel was run overnight. To obtain the sequencing data the computer program e-Seq v3.0, which is part of the LI-COR sequencer, was utilized. Since the initial reactions did not provide complete DNA sequence of each plasmid, a new methodology using M13 primers was used to obtain the complete DNA sequence of pAL236-2, pAL236-5, and pAL236-11.

M13 Protocol

This section provides an overview of how the M13 protocol was utilized. Details for how each part of the M13 protocol was completed will be described below. To obtain the remaining DNA sequences of pAL236-2, pAL236-5, and pAL236-11, primers were designed based on the initial DNA sequences obtained from each recombinant plasmid. Each primer contained both known pAL236 nucleotide (nt) sequences (17-23 nt) and the standard M13 phage forward (19 nt) or M13 phage reverse primer (18 nt). Figure 2 is a schematic of one forward or reverse primer.

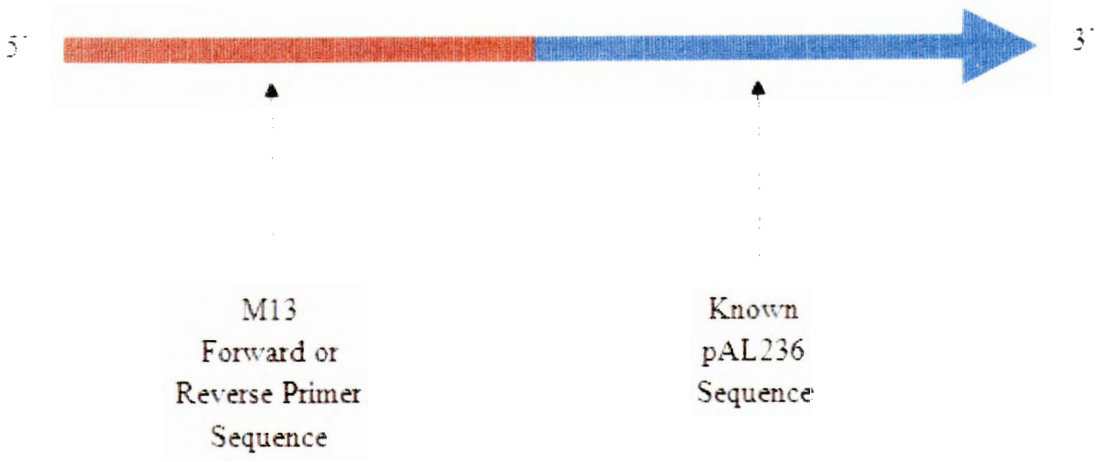

Figure 2. Schematic of designed PCR Primers. Solid red line represents known M13 forward or reverse primer sequences and solid blue line indicates known pAL236 sequence. 5' and 3' ends are indicated.

A PCR reaction utilizing the Gene Amp XL PCR Kit (Applied Biosystems) was performed to produce double-stranded products using two primers. The first primer was an unlabeled forward M13 primer that was specific for one recombinant plasmid ("known sequence" in Fig. 2). The second primer was unlabeled reverse primer specific for the end of the EZ-Tn5™ <R6Kγ*ori*/KAN-2> transposon (see Fig. 1). Additional PCR reactions were performed with both unlabeled forward or reverse M13 primers to obtain the remaining plasmid sequences. The template DNA was one of the recombinant plasmids. The M13 PCR products of ~ 800 to 1500 base pairs (bp) was amplified with the GeneAmp® PCR System 2400 or 2700 models through 25 cycles of PCR with the final double-stranded product containing a replicated M13 primer site for sequencing (Fig. 3). The same M13 protocol was used to obtain double-stranded PCR products for both strands of all three recombinant plasmids. Each PCR product was purified with the Elu-Quik® DNA Purification Kit (Whatman) protocol and sequenced utilizing the SequiTherm EXCEL™ II DNA Sequencing Kit-LC protocol to obtain the remaining plasmid DNA sequence of pAL236-2, pAL236-5, and pAL236-11 (Fig. 4).

Primer Design

The unlabeled M13 forward and reverse trailing primers (described above) were designed using Vector NTI® Advance™ 10 and 11(InforMax®). At least three to five forward or reverse initial DNA sequences obtained from each of the recombinant plasmids were entered into the AlignX® feature on the Vector NTI® Advance™ 10 and 11 computer program. This provided a consensus sequence that was used to design the

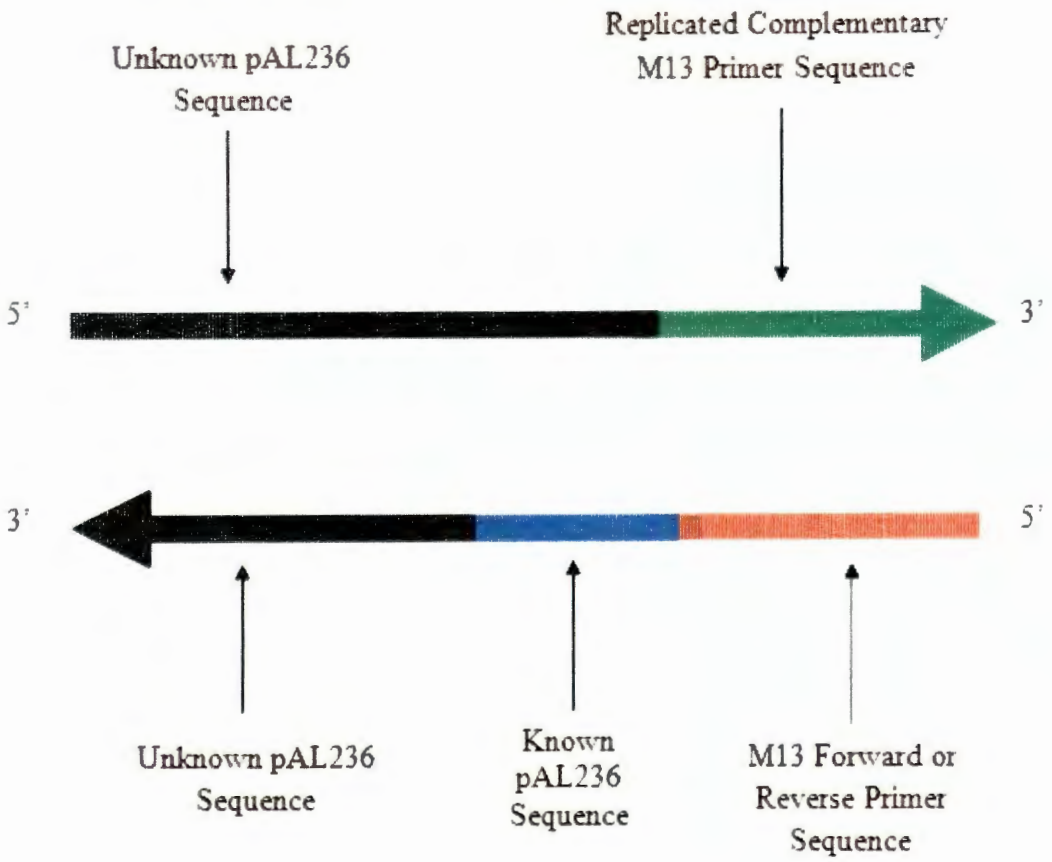

Figure 3. PCR Product containing a replicated M13 Primer Site for sequencing. Red line represents M13 forward or reverse primer sequence. Blue line represents known pAL236 sequence. Green arrow represents replicated complementary M13 primer sequence. Black line and arrow represents unknown pAL236 sequence. 5' and 3' ends are indicated.

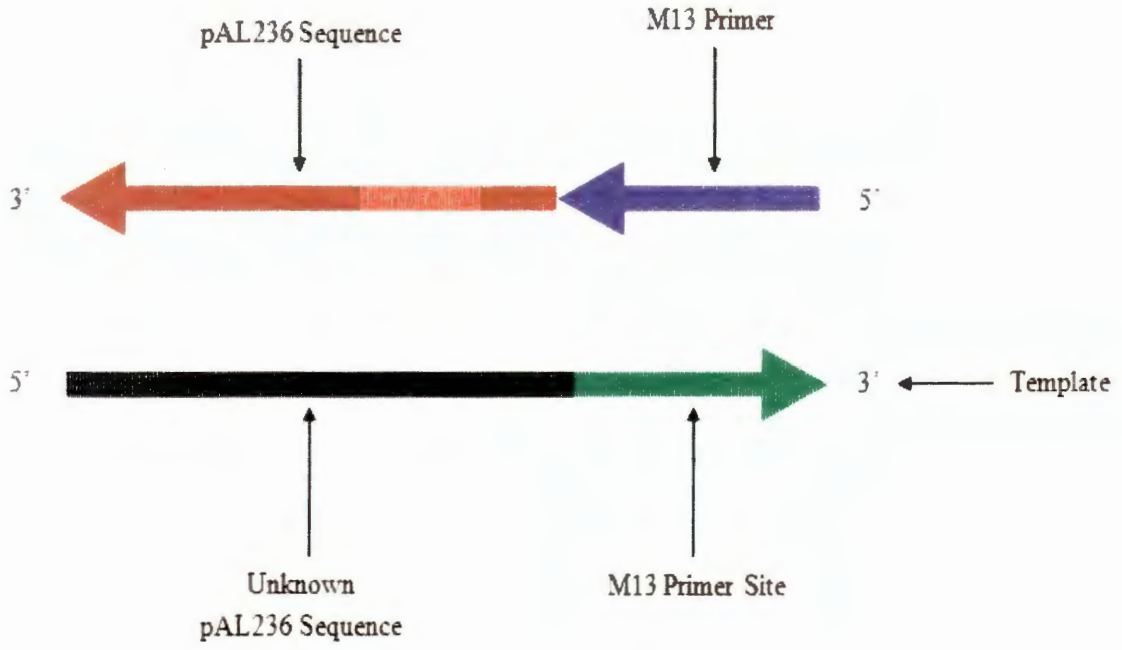

Figure 4. Sequencing of M13 PCR Products. Bottom line represents the PCR product template strand from Fig. 2. Purple arrow represents M13 primer used for sequencing the PCR product. Red arrow represents the derived sequence of pAL236. 5' and 3' ends are indicated.

primers containing known forward and reverse sequences from each recombinant plasmid. The M13 phage forward or reverse primer sequences were added to the end of the known sequence from the recombinant plasmids to complete the design of the primer. The designed primers were obtained from BioSynthesis Inc., Lewisville, TX and were shown in Table 3.

Production of M13 PCR Products

Each of the M13 PCR products described above was produced using the Gene Amp XL PCR Kit procedure. A PCR reaction was mixed on ice in separate 0.2 ml PCR tubes for pAL236-2, pAL236-5, and pAL236-11. Each tube contained 15 µl of 3.3 XL buffer; 4 µl of dNTPs (10,000 pmol/µl); 2.4 µl of $Mg(OAc)_2$; 24.2 µl of UPH_2O; 1.4 µl of DNA template (250 ng); 1 µl (25 pmol/µl) of unlabeled M13 tailing primers designed specifically for each recombinant plasmid; 1 µl (50 pmol/µl) of unlabeled primers obtained from the EZ-Tn5™ <R6K*γori*/KAN-2> Insertion Kit specific for the EZ-Tn5™ <R6K*γori*/KAN-2> transposon or an unlabeled M13 tailing primer (25 pmol/µl); 1.0 µl of rTth DNA polymerase for a final volume of 50 µl. Each PCR reaction was placed into the GeneAmp® PCR System 2400 or 2700 models and were amplified through 25 cycles of PCR. PCR amplification started with an initial 1 min at 94°C; 25 cycles of denaturing at 94°C for 1 min and 30 sec, annealing at 55°C for 1 min and 30 sec, elongation at 72°C for 5 min; and held ∞ at 4°C. Each PCR product was analyzed by electrophoresis on a 1% agarose gel to confirm the production and size of all M13 PCR products.

Purification of M13 PCR Products

All M13 PCR products described above were purified using the Elu-Quik® DNA Purification Kit. Each 50 µl reaction was placed into separate 1.5 ml microcentrifuge tubes. For each tube, 50 µl of UPH_2O was added to bring the overall volume to 100 µl. Binding buffer (280 µl) was added to each tube containing the 100 ul of DNA. Then 20 µl of glass concentrate was added to each tube and the contents were mixed by tapping each tube. The tubes were allowed to stand at room temperature for 10 minutes, while inverting the tubes at 1 minute intervals.

The tubes were placed in the Spectrafuge Model 16M or 24D microcentrifuge at 8800 rpm for 30 seconds. The supernatant was discarded without disturbing the pellet. Wash buffer (500 µl) was added and each pellet was teased from the microcentrifuge tube wall by pipette until the pellet was suspended and inverted once. The tubes were placed back into the microcentrifuge, spun again at 8800 rpm for 30 seconds and the supernatant was discarded. An additional 500 µl of wash buffer was added to each tube and each pellet was teased again from the microcentrifuge tube until the pellet was suspended and inverted once. The tubes were again placed into the microcentrifuge to spin at 8800 rpm for 30 seconds and the supernatant was discarded.

Salt reduction buffer (500 µl) was added to each tube, the pellet was teased from the microcentrifuge tube wall until suspended and inverted once. The tubes were placed back into the microcentrifuge and were allowed to spin at 8800 rpm for 2 minutes. The supernatant was discarded. Once the supernatant was discarded the tubes were placed

back into the microcentrifuge to spin at 8800 rpm for an additional 30 seconds and the remaining supernatant was discarded.

The tubes were placed into the RC 10.10. vacuum centrifuge (Jouan, Inc.) and each pellet was allowed to dry by vacuum for 30 minutes at room temperature, after which 50 µl of UPH_2O was added to each tube and the pellet was resuspended by flicking each tube. The resuspended pellet was placed in a 50°C water bath for 5 minutes and each tube was tapped at 1 minute intervals. Each tube was placed in the microcentrifuge to spin at 8800 rpm for 30 seconds. The supernatant from each tube was placed into a fresh 1.5 ml microcentrifuge tube and placed back into the microcentrifuge to spin for an additional 30 seconds at 8800 rpm. The supernatant was withdrawn and placed into a fresh 1.5 ml microcentrifuge tube. Each purified M13 PCR product was analyzed by electrophoresis on a 1% agarose gel. The tubes containing the PCR products were placed at 4°C and stored until further use.

DNA Sequencing of PCR Products

Labeled M13 forward and reverse primers (LI-COR, Inc.) (see Table 2) specific for each M13 PCR product (described above) were used to determine the DNA sequence utilizing the SequiTherm EXCEL™ II DNA Sequencing Kit-LC protocol described above. DNA sequences obtained from recombinant plasmids and PCR products were used to determine the sequence of pAL236-2, pAL236-5, and pAL236-11.

Plasmid DNA Analysis

Individual DNA sequences obtained from initial DNA sequencing and PCR product sequencing of each recombinant plasmid were submitted to the BLASTN

(Altschul *et al.*, 1990) program of NCBI not only to confirm the sequences were *H. pylori*, but to also confirm that the sequences were not transposon sequences. Individual DNA sequences obtained for each plasmid were assembled using AlignX® feature on the Vector NTI® Advance ™ 10 and 11 to obtain consensus sequences. The consensus sequences for each plasmid were submitted to the ContigExpress® feature on the Vector NTI® Advance ™ 10 and 11 for the construction of pAL236-2, pAL236-5, and pAL236-11 by assembling overlapping DNA fragments into contiguous sequence for each plasmid. Each assembled plasmid was analyzed with Vector NTI® Advance™ 10 and 11 to search for open reading frames (ORFs), ribosome binding sites (RBS), promoter sequences, repeat sequences, and GC content. All ORF sequences for pAL236-2, pAL236-5, and pAL236-11 were submitted to BLASTP (Altschul *et al.*, 1997; Altshcul *et al.*, 2005) program of NCBI to find identities to other known protein sequences.

Conserved domains of proteins were detected with the NCBI Conserved Domain Search (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (Marchler-Bauer *et al.*, 2004; Marchler-Bauer *et al.*, 2009). ORFs were further analyzed with several programs for protein analysis available on the internet through the ExPASy Proteomics tools (http://us.expasy.org). Non coding regions of pAL236-2, pAL236-5, and pAL236-11 were analyzed with BLASTN and BLASTX (Altschul *et al.*, 1990; Altschul *et al.*, 1997) programs at NCBI to search for identities with known DNA and protein sequences. BLASTX searches the entire protein database using a translated nucleotide query.

Phylogenetic Analysis of Replication Proteins in *H. pylori* Plasmids

Replication protein sequences (Rep) were obtained from the following *H. pylori* plasmids: pAL236-2 (this study), pAL236-5 (this study), pAL236-11 (this study), pAL202 (Rickets, 2004), pAL226 (Khatun, 2005), pHPM8 (Quiñones, 2000), pHPM179 (Qasem, 1995; Taylor, 1997), pHPM180 (Minnis *et al.*, 1995), pHPM185 (Liu, 1998; Gao, 1999), pHPM186 (Burnham, 1998), pMCU1 through pMCU7 (De Ungria *et al.*, 1998), pHel1 (Heuermann *et al.*, 1995), pHel4 (Hofreuter *et al.*, 1999), pHel5 (Hofreuter *et al.*, 1999), pHP69 (unpublished, NC_010884), pHP666 (unpublished, NC_010932), pHAG1 (Oh *et al.*, 2006), pHPG27 (Baltrus *et al.*, 2009), pHPP12 (unpublished, NC_011499), pHPS1 (De Ungria *et al.*, 1999), pHP51 (Song *et al.*, 2003), pHPO100 (unpublished, AFO56496), pKU701 (Hosaka *et al.*, 2002), pHP489 (Song *et al.*, 2003), and pHPK255 (Kleanthous *et al.*, 1991) (Table 1). Each Rep protein was numbered 1 through 33 to assist with identification of the Rep protein when constructing phylogenetic trees. Note that *H. pylori plasmids* pHPS1 has both *repA* and *repB* genes and pHPAG1 has two *repA* genes. The Rep proteins were sent to the web server ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) (Larkin *et al.*, 2007) and placed into a global multiple sequence alignment (MSA) using the default ClustalW2 settings: Protein Gap Open Penalty = 10.0, Protein Gap Extension Penalty = 0.2, Protein matrix = Gonnet 250, Protein ENDGAP = -1, Protein GAPDIST = 4. Each MSA shown in the Results section was colored based on the percent identity of the amino acids of the Rep proteins with the consensus of the MSA indicated with blue color, <40% identity has no color, >40% is light blue, >60% is medium dark blue, and >80% is dark blue. Based on the

MSA of the Rep proteins, phylogenetic trees were constructed as cladograms based on average distance using percent identity utilizing Jalview (Clamp *et al.*, 2004; Waterhouse *et al.*, 2009).

CHAPTER III

RESULTS

PURIFICATION AND RESTRICTION ENZYME ANALYSIS OF RECOMBINANT PLASMIDS pAL236-2, pAL236-5, AND pAL236-11

The 2001 base pair (bp) EZ-Tn5™ <R6K*γori*/KAN-2> transposon was inserted into the total plasmid DNA from *H. pylori* strain AL236 (Rickets, 2004). Analyses performed by Rickets revealed that AL236 contained three plasmids, each containing a single insert of the transposon. Each individual recombinant plasmid was transformed into *E. coli pir*$^+$ cells (Rickets, 2004). The number assigned to each plasmid of pAL236 was based on the *E. coli pir*$^+$ cell isolate number that contained the recombinant plasmid. As a result, the three recombinant plasmids were named pAL236-2, pAL236-5, and pAL235-11.

Each recombinant plasmid; pAL236-2, pAL236-5, and pAL236-11 obtained from Rickets was isolated and purified utilizing the QIAGEN Plasmid Midi Kit. The DNA concentration for each plasmid was determined utilizing the Shimadzu UV160U spectrophotometer. After the concentration and purity of each recombinant plasmid was obtained, a restriction digest of the plasmids was performed to get an estimate of the size of each plasmid in base pairs (bp) (Fig. 5). Since the transposon contains a single restriction site for *Bam*HI located at the end of the transposon at bp 1939, which after digestion, linearized each recombinant plasmid.

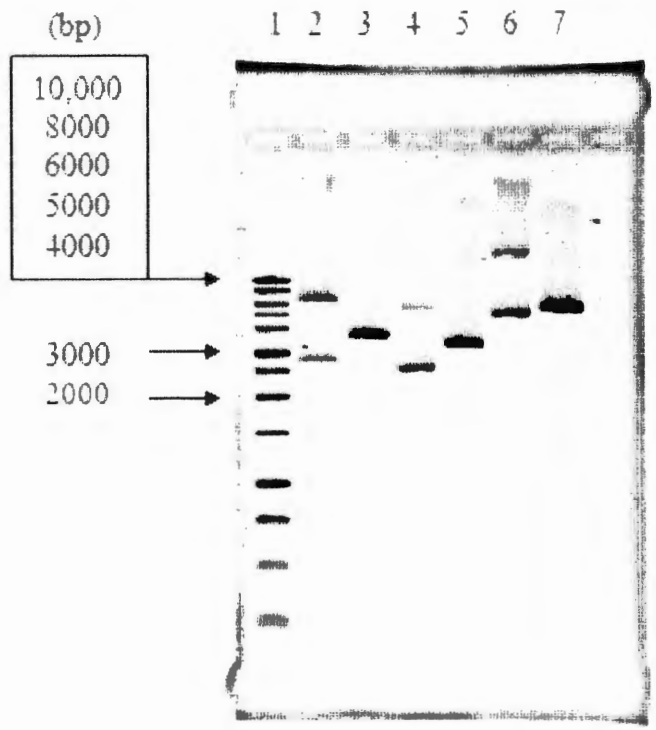

Figure 5. Restriction digests of recombinant plasmids pAL236-2, pAL236-5, and pAL236-11 on 1% agarose gel. Lane 1, BenchTop 1kb DNA Ladder (Promega); lane 2, undigested pAL236-2; lane 3, pAL236-2 digested with *Bam*HI; lane 4, undigested pAL236-5; lane 5, pAL236-5 digested with *Bam*HI; lane 6, undigested pAL236-11; lane 7, digested pAL236-11 with *Bam*HI.

Analysis of a 1% agarose gel electrophoresis confirmed that each recombinant plasmid obtained from Rickets contained a single insert of the EZ-Tn5™ <R6K*ori*/ KAN-2> transposon. The restriction digest with *Bam*HI revealed the estimated sizes for each plasmid; pAL236-2 (~3500 bp), pAL236-5 (~ 3300 bp), and pAL236-11 (~ 5500 bp). After 2001 bp of the transposon was subtracted, the estimated size for each plasmid was determined: pAL236-2 (~1500 bp), pAL236-5 (~1300 bp), and pAL236-11 (~3500 bp). The actual size of pAL236-2, pAL236-5, and pAL236-11 was determined by obtaining the complete DNA sequence of each plasmid.

PURIFICATION AND ANALYSIS OF M13 PCR PRODUCTS USED FOR OBTAINING THE COMPLETE NUCLEOTIDE SEQUENCE OF pAL236-2, pAL236-5, AND pAL236-11

Each PCR product was created using the Gene Amp XL PCR Kit (Applied Biosystems) utilizing a set of unlabeled primers specific for each plasmid (see Table 2). After 25 cycles of PCR were completed each PCR product was purified with the Elu-Quik® DNA Purification Kit (Whatman) (Fig. 6). All M13 PCR products were analyzed by electrophoresis on a 1% agarose gel (Fig. 7). The estimated size for each PCR product; pAL236-2 #1 (~1000 bp), pAL236-2 #2 (~1200 bp), pAL236-5 #1 (~1000 bp), pAL236-5 #2 (~1000 bp), pAL236-11 #1 (~1500 bp), pAL236-11 #2 (~1400 bp), pAL236-11 #3 (~1500 bp), pAL236-11 #4 (~1000 bp), pAL236-11 #5 (~800 bp), pAL236-11 #6 (~1000 bp), pAL235-11 #7 (~1500 bp), and pAL236-11 #8 (~1100 bp). Once the PCR products were purified and analyzed they were sequenced with SeqiTherm EXCEL™ II DNA Sequencing Kit-LC (EPICENTRE®) with labeled M13 forward or

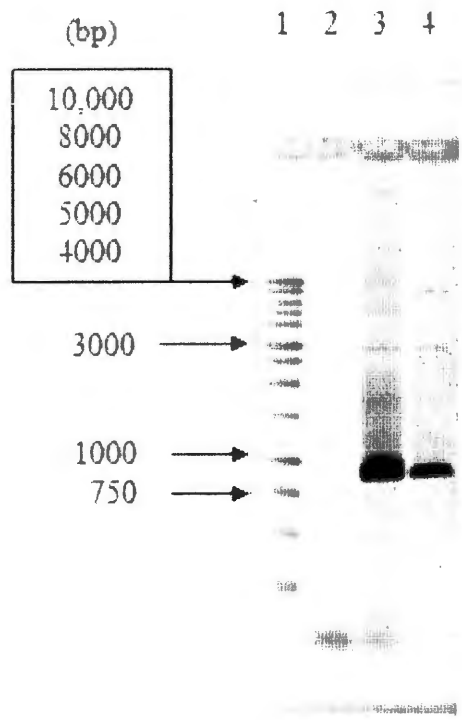

Figure 6. Analysis of a PCR Product by 1% agarose gel electrophoresis before and after purification with Elu-Quik® DNA Purification Kit. Lane 1, Benchtop 1kb DNA Ladder (Promega); lane 2, PCR reaction with no template; lane 3, PCR product before Elu-Quik®; lane 4, PCR product after Elu-Quik®. Fragment sizes in base pair (bp) are shown on left side.

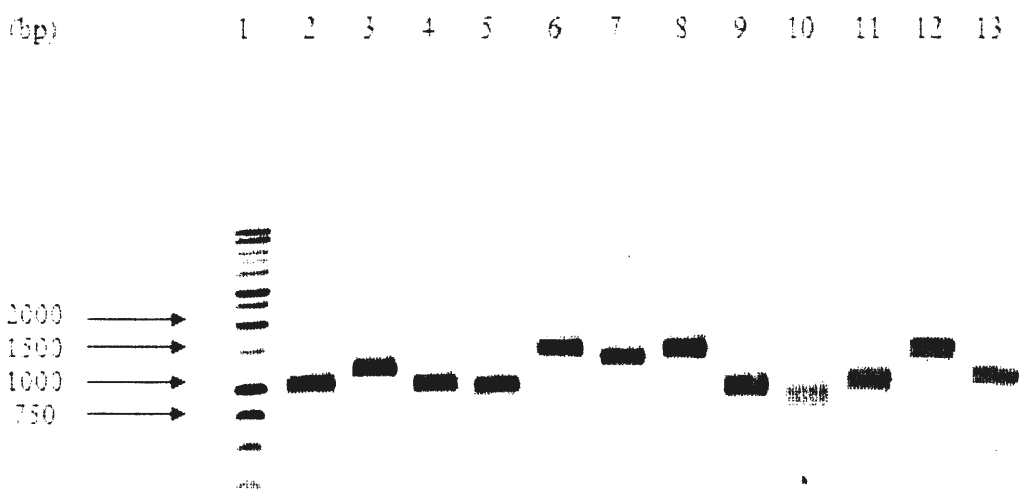

Figure 7. Analysis of M13 PCR Products on 1% agarose gel. The PCR products contained either M13 Forward, M13 Reverse, or M13 Forward and M13 Reverse primer sites used for obtaining remaining DNA sequences for pAL236-2, pAL236-5, and pAL236-11. Lane 1, BenchTop 1kb DNA Ladder; lane 2, pAL236-2 #1; lane 3, pAL236-2 #2; lane 4, pAL235-5 #1; lane 5, pAL236-5 #2; lane 6, pAL236-11 #1; lane 7, pAL236-11 #2; lane 8, pAL236-11 #3; lane 9, pAL236-11 #4; lane 10, pAL236-11 #5; lane 11, pAL236-11 #6; lane 12, pAL236-11 #7; lane 13, pAL236-11 #8.

reverse primers at least three to five times to obtain the complete sequence of pAL236-2, pAL236-5, and pAL236-11.

DNA SEQUENCING FOR pAL236-2, pAL236-5, AND pAL236-11

Initial sequencing of all three recombinant plasmids pAL236-2, pAL236-5, and pAL236-11 was done bi-directionally utilizing specifically labeled EZ-Tn5™ <R6K γ*ori*/KAN-2> transposon KAN-2 FP-1 Forward and R6KAN-2 RP-1 Reverse primers (see Table 2) to determine the sequence near the transposon insertion site for each of the three recombinant plasmids. The sequence for each strand was determined between three and five times. Approximately 1000 to 1200 bp were obtained from each individual sequence reaction. Each one of the initial DNA sequences was submitted to the Basic Local Alignment Search Tool (BLASTN) (Altschul *et al.*, 1990) program of the National Center for Biotechnology Information (NCBI) to determine if the sequences had any identity to other *H. pylori* sequences and to confirm that the sequences obtained were not EZ-Tn5™ <R6Kγ*ori*/KAN-2> transposon sequences. Multiple sequences for each recombinant plasmid were aligned in the AlignX® program from Vector NTI® Advance™ 10 and 11 to obtain consensus sequences for each recombinant plasmid. Each consensus sequence was assembled using the ContigExpress® program from Vector NTI® Advance™ 10 and 11.

The assembled sequences for pAL236-2 and pAL236-5 contained significant overlap with the initial forward and reverse DNA sequences. However, the complete sequence for these plasmids was not obtained. The assembly of the initial forward and

reverse sequences of pAL236-11 did not have any overlap. As a result, a new strategy was utilized to obtain the complete sequences for all three plasmids.

The new protocol utilized known M13 DNA sequences and was described in the method section. Each initial PCR product was purified utilizing the Elu-Quik® DNA Purification Kit (Whatman) and analyzed by electrophoresis on a 1% agarose gel (see Fig. 6 and Fig. 7). PCR products for the recombinant plasmids had varied sizes ranging from 800 bp to 1500 bp (see Fig. 7). After purification the DNA sequences of each PCR product was determined as described in methods. Each PCR product was sequenced between three to five times. Approximately 500 to 1100 bp was obtained for each individual sequencing reaction. DNA sequences were aligned, assembled, and analyzed as described above. Based on sequence, *H. pylori* plasmids pAL236-2, pAL236-5, and pAL236-11 were established as 1448, 1216, and 3148 bp respectively. Characteristics of each plasmid will be described separately.

pAL236-11 DNA AND ORF ANALYSIS

The complete sequence of pAL236-11 is shown in Figure 8. The overall GC content of pAL236-11 was 35.9%, which is lower than that of sequences of *H. pylori* chromosomal DNA of strains P12 (38.0%), B128 (38.0%), Shi470 (38.0%), 26695 (38.9%), G27 (38.9%), HPKX_438_CA4C1 (39.0%), HPKX_438_AG0C1 (39.0%), HPAG1 (39.1%), and J99 (39.2%). Interestingly, the GC content of pAL236-11 (35.9%) was higher than *H. pylori* plasmids pHPM185 (33.1%), pHP489 (33.3%), pHPM8 (33.4%), pAL236-5 (33.8%), pHPG27 (34.0%), pHP69 (34.0%), pAL202 (34.3%), pHel4 (34.3%), and pHel5 (34.4%). However, the GC content of pAL236-11 was similar to

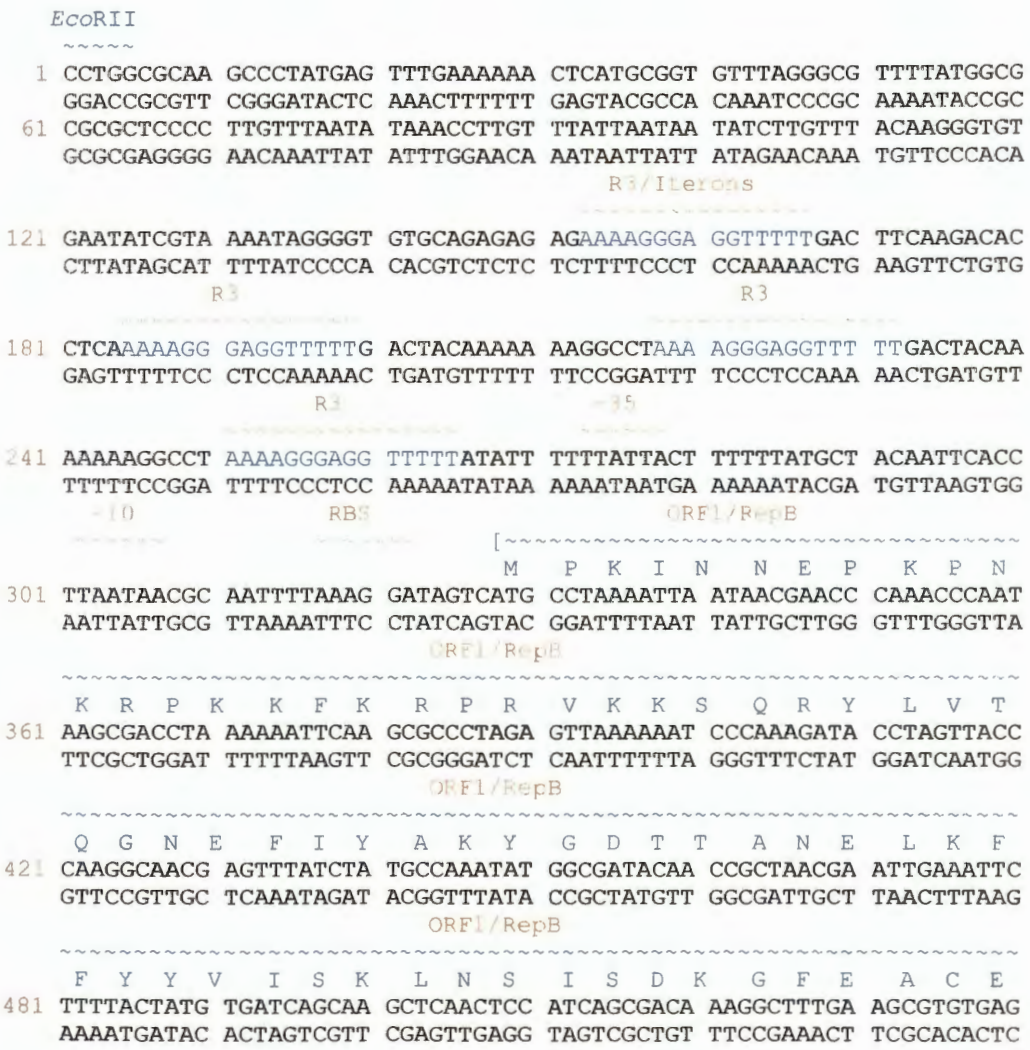

Figure 8. Nucleotide sequence of pAL236-11.
The single *Eco*RII site at nucleotide one is indicated with a blue wavy line; iterons are indicated by red wavy lines with the nucleotides highlighted in blue (153-167 nt), (185-199 nt), (218-232 nt), and (251-265 nt); -10 and -35 promoter regions (273-278 nt), (302-307 nt), (2132-2137 nt), and (2154-2159 nt) are indicated by short red wavy lines with the names of the promoter above the line in red letters; ribosome binding sites (RBS) (317-322 nt) and (2171-2176 nt) are indicated by short red wavy lines with the name of the RBS above the line in red letters; putative open reading frames (ORFs) are shown as blue wavy arrows in the direction of translation; the amino acid sequences are shown in blue letters below the ORFs.

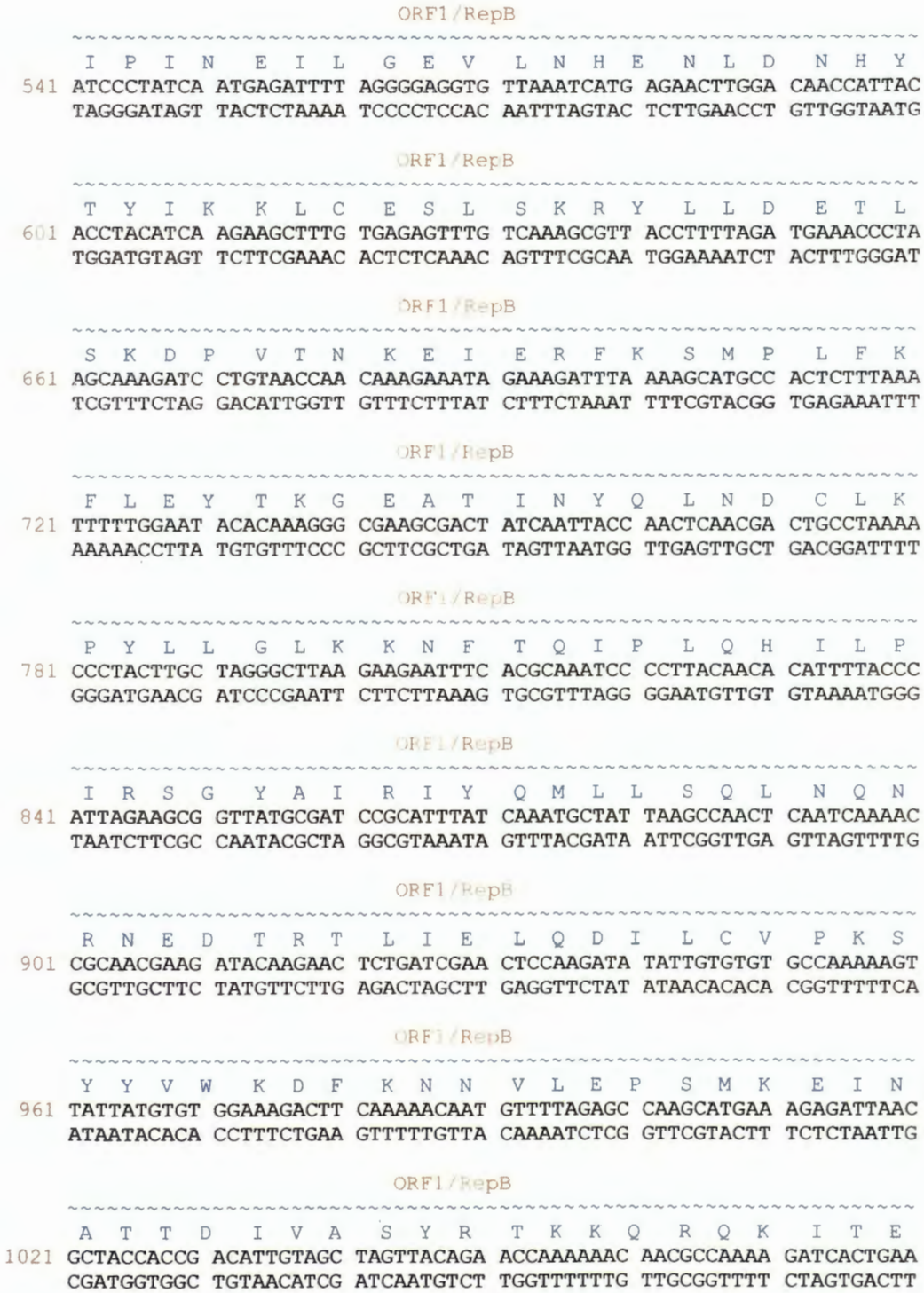

Figure 8 continued.

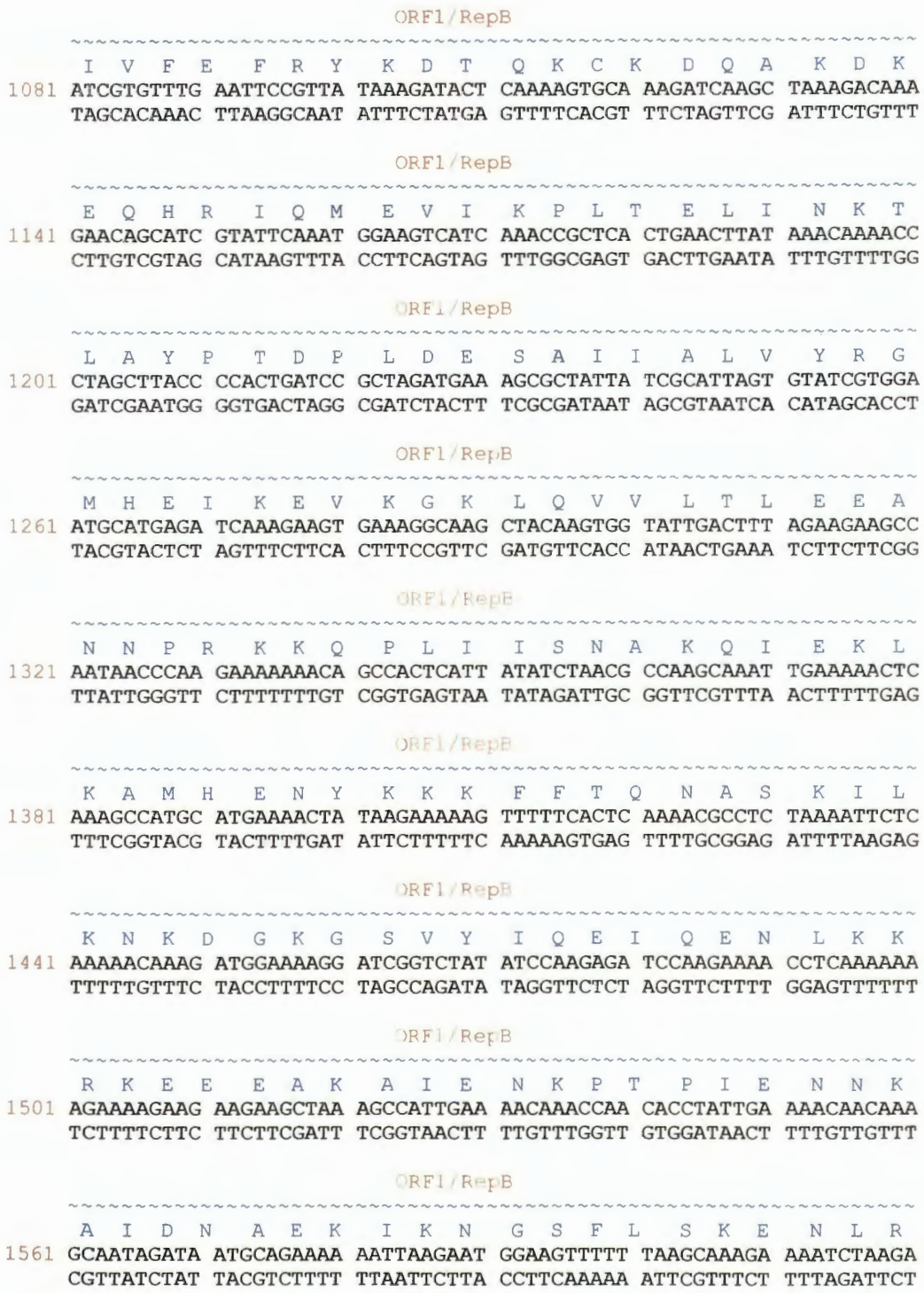

Figure 8 continued.

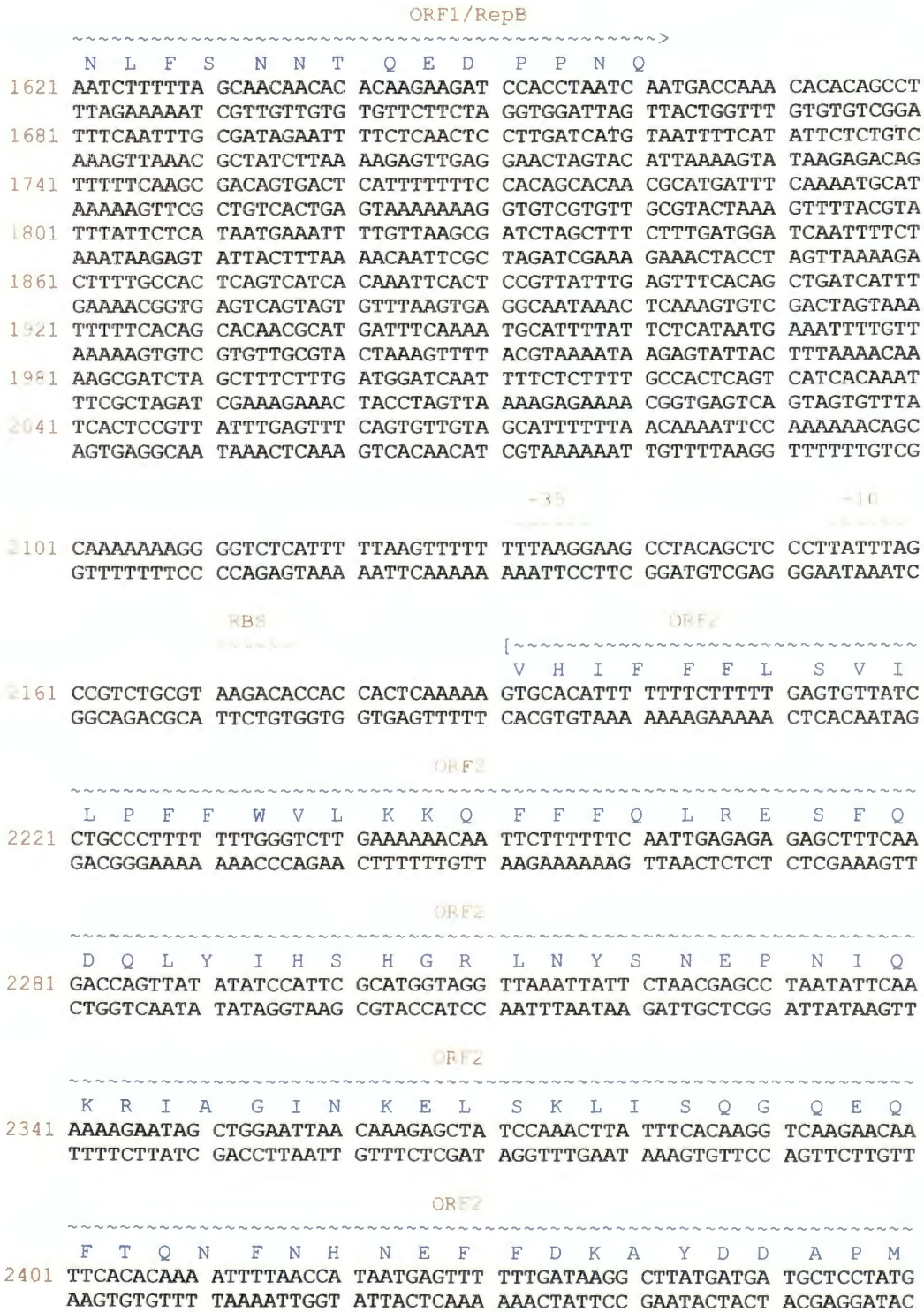

Figure 8 continued.

```
                                        ORF2
      ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
       E  M  K  L   L  A  S   V  K  N   K  E  L  G   K  L  E   K  S  K
2461   GAAATGAAAC TTCTAGCTAG TGTTAAAAAC AAAGAACTAG GCAAGTTAGA AAAATCAAAA
       CTTTACTTTG AAGATCGATC ACAATTTTTG TTTCTTGATC CGTTCAATCT TTTTAGTTTT

                      ORF2
      ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~>
       L  E  K  V   Q  I  A   I  E  V
2521   TTAGAAAAAG TTCAAATCGC TATTGAAGTG TAAAAGTTGA TCAGTAGTGA AGTCGCTCAA
       AATCTTTTTC AAGTTTAGCG ATAACTTCAC ATTTTCAACT AGTCATCACT TCAGCGAGTT
2581   TTAGTAAAAA TGAATTTTGG AATGAGCGCT TAAGTGTGAG AGGTTGAAAG AGTTGATTAC
       AATCATTTTT ACTTAAAACC TTACTCGCGA ATTCACACTC TCCAACTTTC TCAACTAATG
2641   CACTCAAAGC GTTGGAGTGT AAAAATCAAA ATTGAAAAAA GTTCAAATCG CTATTGAAGT
       GTGAGTTTCG CAACCTCACA TTTTTAGTTT TAACTTTTTT CAAGTTTAGC GATAACTTCA
2701   GTAAAAGTTG ATCAGTAGTG AAGTCGCTCA ATTAGTAAAA ATGAATTTTG GAATGAGCGC
       CATTTTCAAC TAGTCATCAC TTCAGCGAGT TAATCATTTT TACTTAAAAC CTTACTCGCG
2761   TTAAGTGTGA GAGGTTGTAA AAGTTGATCA CCACTCAAAG CGTTGGAGTG TAAAAATCAA
       AATTCACACT CTCCAACATT TTCAACTAGT GGTGAGTTTC GCAACCTCAC ATTTTTAGTT
2821   AAATTGGAAT GTGTGCTTCA CTTTAGTTAG TGCTTGGAAG TGAGTTTGAA AAAAATAAAA
       TTTAACCTTA CACACGAAGT GAAATCAATC ACGAACCTTC ACTCAAACTT TTTTTATTTT
2881   ACTAAAATCC CTTTTAGATG CTCACAGAAA GCAAATATAA CGAGTTTAGC AACTAAGAGA
       TGATTTTAGG GAAAATCTAC GAGTGTCTTT CGTTTATATT GCTCAAATCG TTGATTCTCT
2941   GCATGTAACT AGGGCTTGCT AACAACAAAA CCGCATAGCG AAGCGTAAGC GGGCATTCAA
       CGTACATTGA TCCCGAACGA TTGTTGTTTT GGCGTATCGC TTCGCATTCG CCCGTAAGTT
3001   GCGGTTTTGC ACCTATTTGT TTGCTGACAA GCAAACACCA ATAAGCGAAG CGTAAGCGAA
       CGCCAAAACG TGGATAAACA AACGACTGTT CGTTTGTGGT TATTCGCTTC GCATTCGCTT
3061   GCATGGACAA AAGCGCATCG CAGTTTGAAA TCGCAGGCGT TAGCCGGAGA TGGTTTTGCG
       CGTACCTGTT TTCGCGTAGC GTCAAACTTT AGCGTCCGCA ATCGGCCTCT ACCAAAACGC
3121   ATAGCAAAAT CAAACAAGAT AGCGCAAA
       TATCGTTTTA GTTTGTTCTA TCGCGTTT
```

Figure 8 continued.

H. pylori plasmids pHPP12 (35.0%) pHPP12 (35.0%), pHel1 (35.1%), pHP51 (35.3%), pHPM179a (35.5%), pHP666 (35.6%), pAL226 (35.8%), and pHPM186 (35.9%). Finally, the GC content of pAL236-11 was lower than *H. pylori* plasmids pHPM179b (36.2%), pHPO100 (36.2%), pHPAG1 (36.4%), pHPK255 (36.8%), pHPM180 (37.0%), pHPS1 (37.3%), pAL236-2 (37.5%) and pKU701 (38.0%). (The GC content of all plasmids included here is summarized in Table 8, which includes geographic regions.)

The open reading frames (ORFs) were determined using ATG and GTG as translation start sites and TAA, TAG, and TGA as the stop codons with a size limit of 100 amino acids. The complete sequence of pAL236-11 contained two ORFs, both on the same strand. The amino acid sequences of ORF1 and ORF2 is indicated in Figure 8 and a circular diagram of the plasmid is shown in Figure 9. Possible -10 and -35 promoter sequences and ribosome binding sites (RBS) for each ORF were determined using the consensus sequences of *E. coli*. Locations for these regions along with ORF1 and ORF2 identities are shown in Table 4.

Nucleotides 328-1662, *orf1*, showed identity to the *orf1/repB* nucleotide sequence of several *H. pylori* plasmids: pKU701 (90%); pHP51 (85%); pHPS1 (79%); and the *repA* gene of pHPO100 (85%). The amino acid sequence of ORF1 of pAL236-11 showed strong identity to the replication protein, RepB, of *H. pylori* plasmids pHP51 (79%), pHPS1 (69%), pKU701 (81%), and to the RepA protein of pHPO100 (79%). The amino acid alignment is shown in Figure 10. Two important findings were obtained from these results. First, this alignment shows that ORF1 of pAL236-11 is RepB.

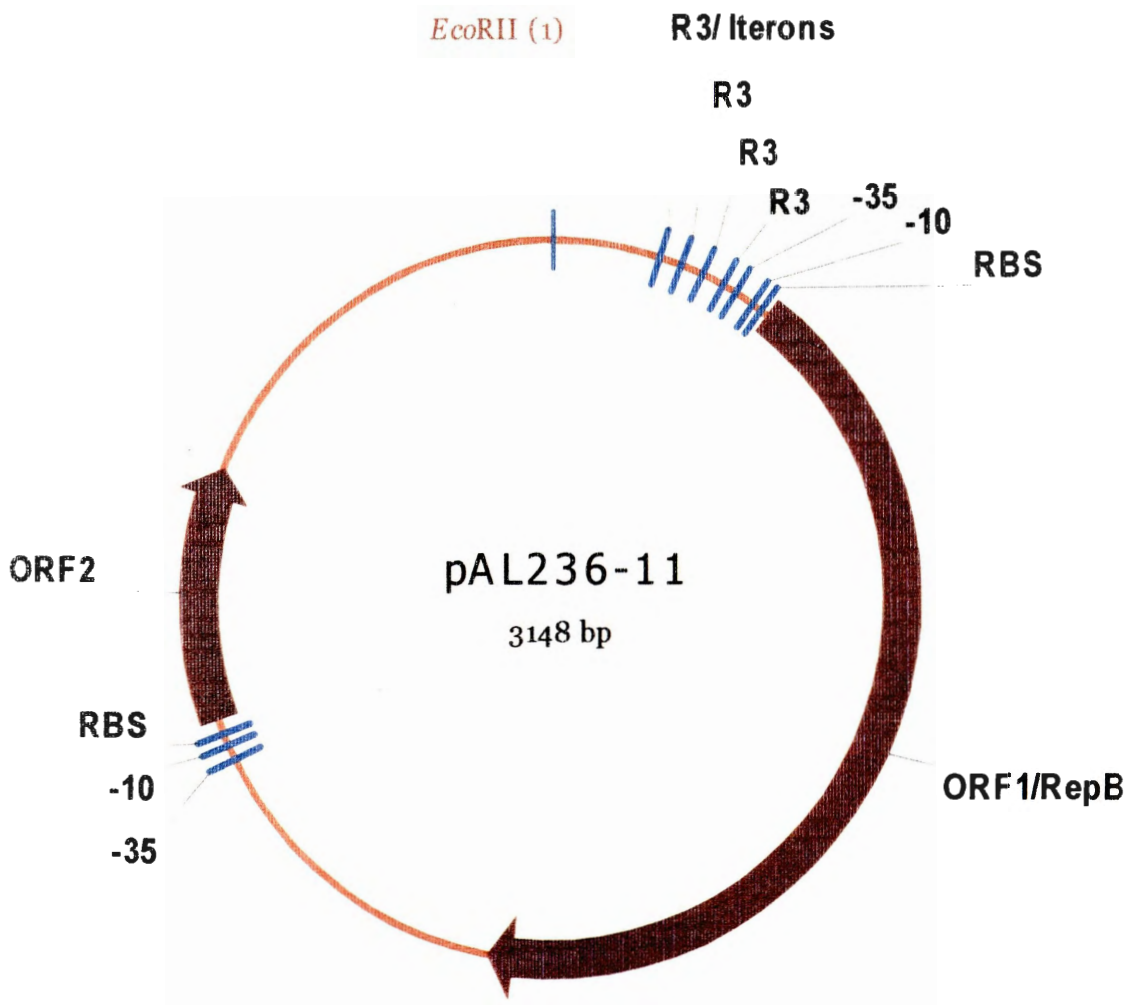

Figure 9. A diagram of pAL236-11 showing Iterons, Promoter Regions, Ribosome Binding Sites (RBS), and Open Reading Frames (ORFs). Blue vertical lines represent iterons (R3), promoters (-10 and -35), and RBS. Arrowheads represent putative ORFs and the direction of translation. The single restriction endonucleases site, *Eco*RII is assigned as nucleotide number 1.

Table 4. Analysis of ORF1 and ORF2 from pAL236-11. a) pAL236-11 ORF analysis by Vector NTI Advance™ 11. b) Promoter and ribosome binding site (RBS) locations of pAL236-11. c) Nucleotide identities of ORF1 and ORF2 with known plasmid sequences.

a)

ORF	Sequence	Length	M.W.[1]	IP[2]
	n.t.#	a.a#		
1	328-1662	445	52152	9.40
2	2132-2137	120	14254	8.85

1-Molecular weight; 2-Isoelectric point.

b)

Promoter		RBS
-35 5' TTGACA 3'	-10 5' TATAAT 3'	5' AGGAGG 3'
TTATTA (273-278)	TAATAA (302-307)	AAAGGA (317-322)
TTAAGG (2132-2137)	TATTTA (2154-2159)	AAGACA (2171-2176)

c)

ORF	Main Identities	Percent Identity	Proposed Function
1	RepB pKU701 RepB pHP51 85 RepA pHPO100 RepB pHPS1	81 79 79 69	Plasmid Replication
2	None	No Identities	Unknown Function

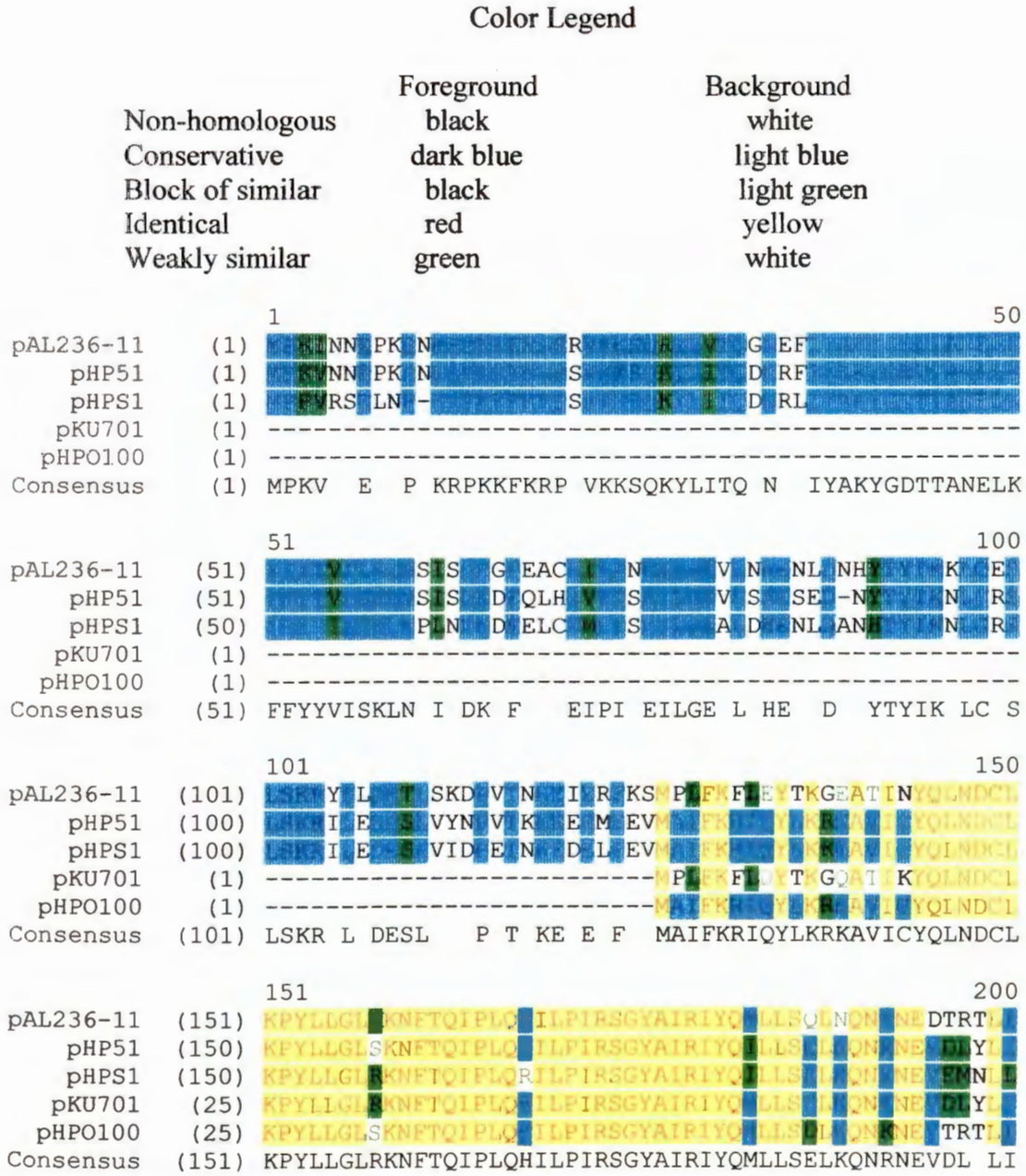

Figure 10. Amino acid alignment of pAL236-11 ORF1/RepB compared to related *H. pylori* Plasmids.

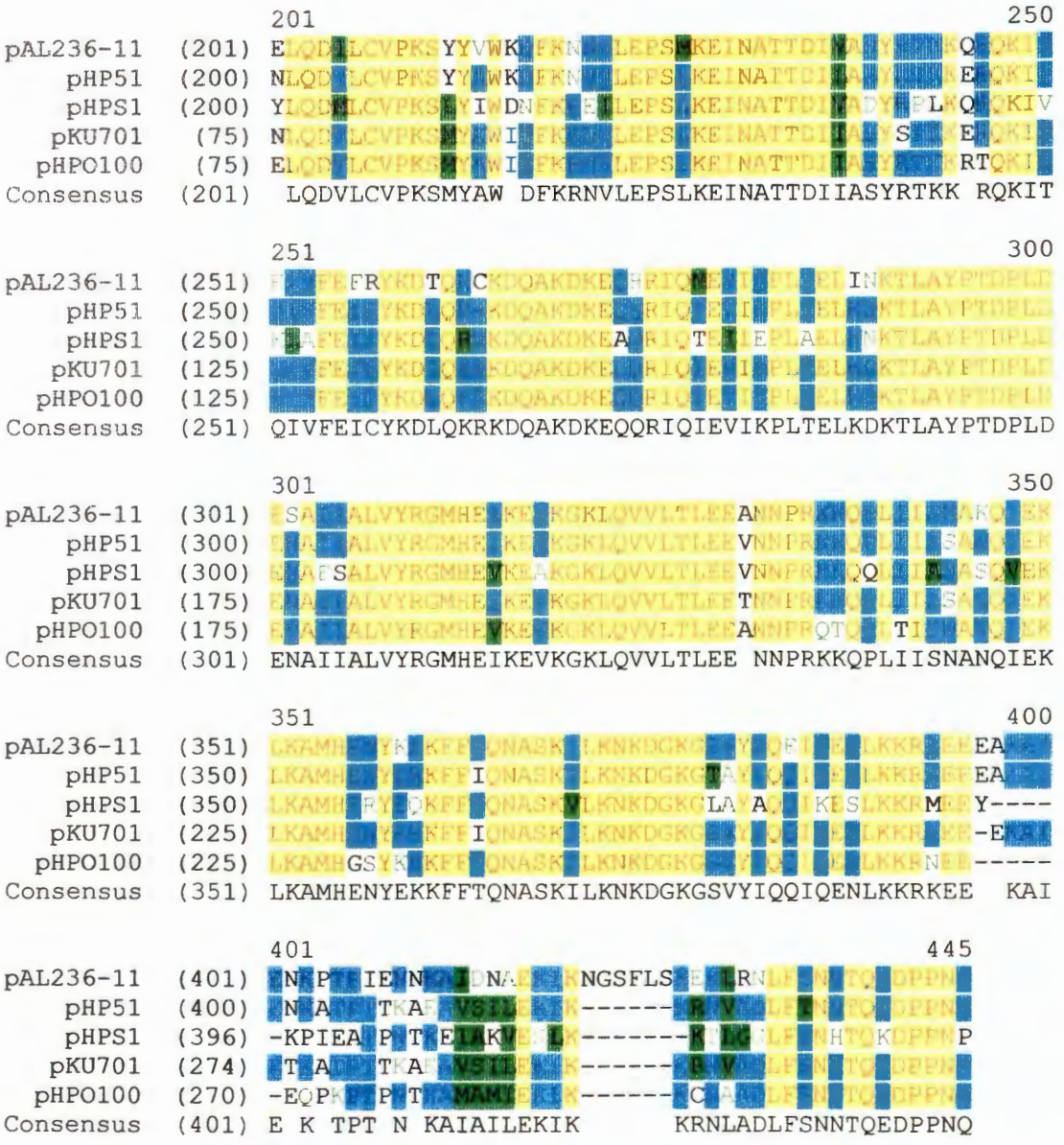

Figure 10 continued.

Second, RepA of pHPO100 belongs with the RepB replication proteins that are also found on plasmids pHP51, pHPS1, pKU701, and pAL236-11. There was additional identity to the replication proteins of plasmids from unrelated species such as *Campylobacter jejuni* (41%), *Campylobacter coli* (41%), and *Geobacillus kaustophilius* (28%).

The ORF1/RepB protein consisted of 445 amino acid residues. Submission of ORF1/RepB to BLASTP (Altschul *et al.*, 1997; Altshcul *et al.*, 2005) at NCBI, indicated that this protein contained a conserved domain that belonged to the Rep_3 superfamily. According to the Conserved Domain Database (CDD) the Rep_3 superfamily are proteins that are involved with the initiation of plasmid replication (Marchler-Bauer *et al.*, 2004; Marchler-Bauer *et al.*, 2009). These proteins included both RepA and RepB and possess nicking-closing (topoisomerase I) like activity which allows these proteins to perform a strand transfer reaction on ssDNA that contains its target.

ORF1/RepB was analyzed by Vector NTI® Advance™ 11 and indicated that the protein has a molecular weight of 52152 (52.1 kDa) and an isoelectric point of 9.40 (see Table 4a). Additional analysis of ORF1/RepB was performed by software programs available through the internet through ExPASy Proteomics tools (http://us. expasy.org). SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui) (Hirokawa *et al.*, 1998; Mitaku *et al.*, 1999; Mitaku *et al.*, 2002), a program that indicates the hydrophobicity index of proteins, indicated that ORF1/RepB is soluble. Analysis with TMHMM (http://www.cbs. dtu.dk/services/TMHMM-2.0) (Tusnady *et al.*, 1998; Tusnady *et al.*, 1999), which predicts transmembrane helices in proteins, indicated that ORF1/RepB did not contain

any transmembrane helices. ProtParam (http://us.expasy.org /tools /protparam.html) (Gasteiger *et al.*, 2005), analysis determines the instability index and provides an estimate of the stability of the protein in a test tube; proteins with an instability index above 40 might be unstable. Analysis by ProtParam indicated that ORF1/RepB was stable with an instability index (II) of 39.31.

Nucleotides 2191-2550, *orf2*, did not have significant similarity to *H. pylori* plasmids or unrelated species according to BLASTN (Altschul *et al.*, 1990) at NCBI. Next, the protein sequence of ORF2 was submitted to BLASTP at NCBI. BLASTP results showed that ORF2 did not have significant similarity to any *H. pylori* or unrelated species proteins. Results with the Conserved Domain Search indicated that this protein did not contain any conserved domains.

The ORF2 protein consisted of 120 amino acid residues. Analysis of ORF2 by Vector NTI® Advance™ 11 indicated that ORF2 protein had a molecular weight of 14254 (14.2 kDa) and an isoelectric point of 8.85 (see Table 4a). Additional analysis of ORF2 was performed using software programs available via the internet through ExPASy Proteomics tools. SOSUI and TMHMM determined that ORF2 was a soluble protein with no transmembrane helices. ProtParam analysis determined that the instability index (II) of ORF2 was 41.25 and classified the protein as unstable. The exact function of this protein is unknown. As a result, the ORF2 protein from pAL236-11 was not the ORF2 protein that is found on other theta replicons that belong to the cAMP-induced filamentation (Fic) protein family.

There were two main regions of pAL236-11 that and were determined to be non-coding sequences. The first non-coding sequence was located at nucleotide (nt) positions 1663-2190, a total of 538 nt. Submission of this sequence to BLASTN at NCBI, indicated that only the nucleotides 2082-2190 had identity with non-coding regions of pHPS1 and pHPO100 (Table 5). Results from submission of this sequence in BLASTX (Altschul *et al.*, 1997), which searches the entire protein database using a translated nucleotide query, indicated that the pAL236-11 non-coding region did not have significant identity to any *H. pylori* chromosomal, plasmid, or non-related species proteins.

The other non-coding region of pAL236-11 was located at nt positions 2551-3148 and 1-327, a total of 925 nt. Results for submission of this sequence to BLASTN indicated that nucleotides 2551-2887 had no identity to DNA sequences of *H. pylori* chromosomal, plasmid, or non-related species. The remaining non-coding region showed identity with non-coding regions of pKU701 and pHP51 (Table 5). These non-coding regions of pKU701 contained the R3 iteron region, while pHP51 contained both the R1 iteron region along with the common region (CR). The non-coding region of pAL236-11 had identity with two non-coding regions of pHPS1, pHPP12, and pHPAG1, which contained the R2 repeat regions (Table 5). The R2 repeats of pHPP12 and pHPAG1 were identified in this study. The R2 repeats for pHPP12 were located at nt positions 5758-6440 and 6487-7169 and consisted of a two time 683 bp direct repeat sequence separated by 46 bp. Plasmid pHPAG1 had R2 repeats located at nt positions 2785-3384 and 8774-

Table 5. Non-coding Region Identities of pAL236-11 with Other *H. pylori* Plasmids. Plasmid features located within the non-coding region are indicated. (R2=Repeat Regions, R3=Repeat Regions, and C1=Common Region) 1- R2 repeat regions identified in this study. 2- Intially identified as R1.

pAL236-11 Sequence (nt)	Main Plasmid Identities	Plasmid Sequence (nt)	Coding Region	Percent Identity with pAL236-11
2084-2190	pHPS1	5025-5132	Non-Coding	92%
2082-2190	pHPO100	158-265	Non-Coding	90%
2888-3148, 1-294	pKU701	1329-1891	(R3)	83%
2892-3148, 1-295	pHP51	3121-3676	(R3[2],C1)	82%
2910-3148, 1-112 2910-3148, 1-7	pHPS1	2893-3245 5432-5681	(R2) (R2)	81% 83%
2888-3148, 1-7	pHel4	5783-6048	Non-Coding	84%
2964-3148, 1-7	pHPP12	5989-6178 6718-6907	(R2)[1] (R2)[1]	91% 91%
2967-3148, 1-7	pHP666	6248-6454	Non-Coding	91%
2965-3148, 1-7	pHP69	8756-8942	(C1)	91%
2923-3148, 1-7	pAL226	9455-9694	Non-Coding	86%
2914-3148, 1-7	pHel5	1-243	Non-Coding	85%
2967-3148, 1-14	pHPAG1	3022-3214 9011-9203	(R2)[1] (R2)[1]	89% 89%
2965-3148, 1-2	pHPG27	473-600	Non-Coding	89%
2910-3148, 1-7	pHPM8	11-263	Non-Coding	83%
3014-3148, 1-14	pAL202	11781-11926	Non-Coding	91%
3043-3148, 1-8	pHPM180	2143-2253	Non-Coding	91%
3044-3148	pHel1	89-191	Non-Coding	90%
1663-2190 2551-3148, 1-327	pAL236-2	793-1448, 1-99	Non-Coding Non-Coding	35% 41%
1663-2190 2551-3148, 1-327	pAL236-5	759-1216, 1-113	Non-Coding Non-Coding	42% 32%

9370, 1-3 that consisted of a two time 600 bp repeat sequence separated by 2784 bp (2.7 kbp). Also, pAL236-11 had identities to the non-coding region of pHP69, which was similar to the CR found in pHPS1. Additionally, the non-coding region of pAL236-11 had various identities to non-coding regions of *H. pylori* plasmids including pHP666, pAL202, pAL226, pHel1, pHel4, pHel5, pHPG27, pHPM8, and pHPM180 (see Table 5). The non-coding region of pAL236-11 was submitted to BLASTX and the results indicated no significant identity to any *H. pylori* chromosomal, plasmid, or non-related species proteins.

The search for repeat and IS elements revealed no R1, R4, or *H. pylori* specific IS elements in pAL236-11. Plasmid pAL23-11 showed partial identity to R2 regions of pHPS1, pHPAG1, and pHPP12; however, no complete set of R2 repeat sequences were present. The search did reveal that pAL236-11 contained the R3 iteron region that consists of a 4 time 15 bp repeat with a spacer sequence of 17-18 bp between each repeat. The R3 region of pAL236-11 is indicated by Figure 8 and 9. Utilizing AlignX® on the Vector NTI® Advance™ 11 computer program, the R3 iterons of pAL236-11 showed identity with R3 iterons of pKU701 (72.8%), pHPS1 (63.5%), pHPO100 (67.3%), and the R1 iterons of pHP51 (60.0%). For plasmids pHPS1, pHP51, pKU701, and pAL236-11 the R3 iterons are present upstream of *repB*. The R3 iteron sequence of pHPO100 is also located upstream of *repA*. The R3 iteron region of pKU701 and pHPS1 consists of a 3.5 time 33 bp direct tandem repeat, while plasmid pHP51 R1 iteron region consists of a 4 time 27 bp repeat with 5 bp spacer sequences between each repeat. Lastly, the R3 iteron region of pHPO100 consists of a 3.5 time 21 bp repeat with 11-12 bp spacer sequences

between each repeat. A complete sequence alignment of these iterons is shown in Figure 11. Interestingly, when comparing the entire R3 iterons of plasmids pHPS1, pHPO100, pKU701, pAL236-11, and R1 of pHP51 there were a total of 51 conserved nucleotides within these sequences. This result indicates that the R1 of pHP51 belongs to the R3 iterons of pHPS1, pKU701, pHPO100, and pAL236-11. As a result of the alignment of the R3 iterons and the amino acid alignment of the replication proteins gives further evidence that the RepA found on pHPO100 is RepB while the R1 iterons of pHP51 is R3.

Color Legend

	Foreground	Background
Non-homologous	black	white
Conservative	dark blue	light blue
Block of similar	black	light green
Identical	red	yellow
Weakly similar	green	white

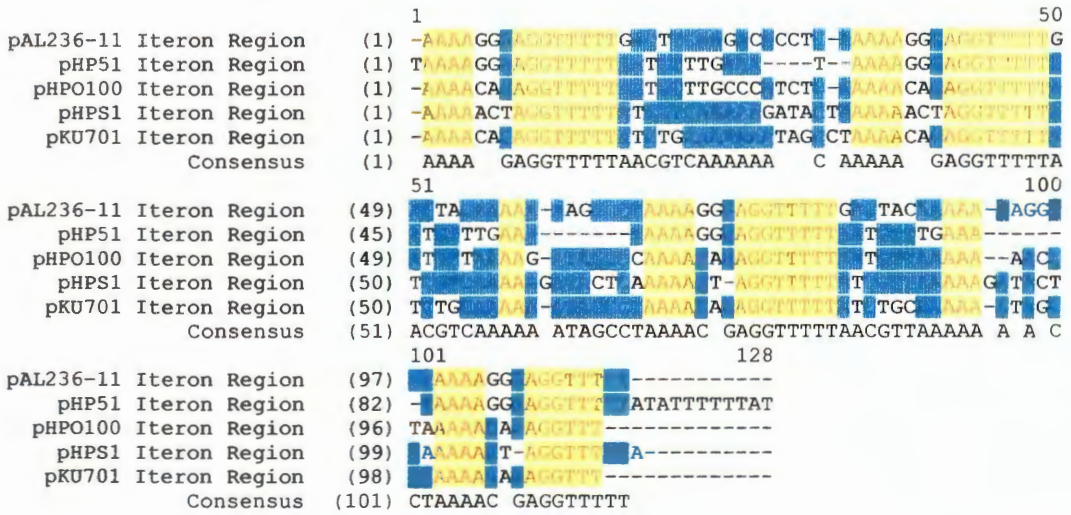

Figure 11. Nucleotide alignment of R3 Iteron Region of pAL236-11 compared to R3 Iteron Regions of pHPS1, pHPO100, pKU701, and R1 Iteron Region of pHP51.

pAL236-2 DNA AND ORF ANALYSIS

The complete sequence of pAL236-2 is shown in Figure 12. The overall GC content of pAL236-2 was 37.5 %, which is lower than that of sequences of *H. pylori* chromosomal DNA of strains, P12 (38.0%), Shi470 (38.0%), B128 (38.0%), G27 (38.9%), 26695 (38.9%), HPKX_438_CA4C1 (39.0%), HPKX_438_AG0C1 (39.0%), HPAG1 (39.1%), and J99 (39.2%). The GC content of pAL236-2 (37.5%) was similar to *H. pylori* plasmids pHPK255 (36.8%), pHPM180 (37.0%), pHPS1 (37.3%) and lower than pKU701 (38.0%). However, GC content of pAL236-2 was higher compared with other *H. pylori* plasmids pHPM185 (33.1%), pHP489 (33.3%), pHPM8 (33.4%), pAL236-5 (33.8%), pHP69 (34.0%), pHPG27 (34.0%), pAL202 (34.3%), pHel4 (34.3%), pHel5 (34.4%), pHPP12 (35.0%), pHPP12 (35.0%), pHel1 (35.1%), pHP51 (35.3%), pHPM179a (35.5%), pHP666 (35.6%), pAL226 (35.8%), pHPM186 (35.9%), pHPM179b (36.2%), pHPO100 (36.2%), and pHPAG1 (36.4%) (see Table 8).

The open reading frames (ORFs) were determined using ATG and GTG as translation start sites and TAA, TAG, and TGA as the stop codons with a size limit of 100 amino acids. The complete sequence of pAL236-2 contained one ORF (ORF1) on one strand and no ORFs on the opposite strand. The amino acid sequence of ORF1 is indicated in Figure 12 and a circular diagram of the plasmid is shown in Figure 13. Possible -10 and -35 promoter sequence and ribosome binding site (RBS) for ORF1were determined using the consensus sequences of *E. coli*. Locations for these regions and ORF1 identities are shown in Table 6. The Rep from pAL236-5, shown in Table 6, will be discussed later.

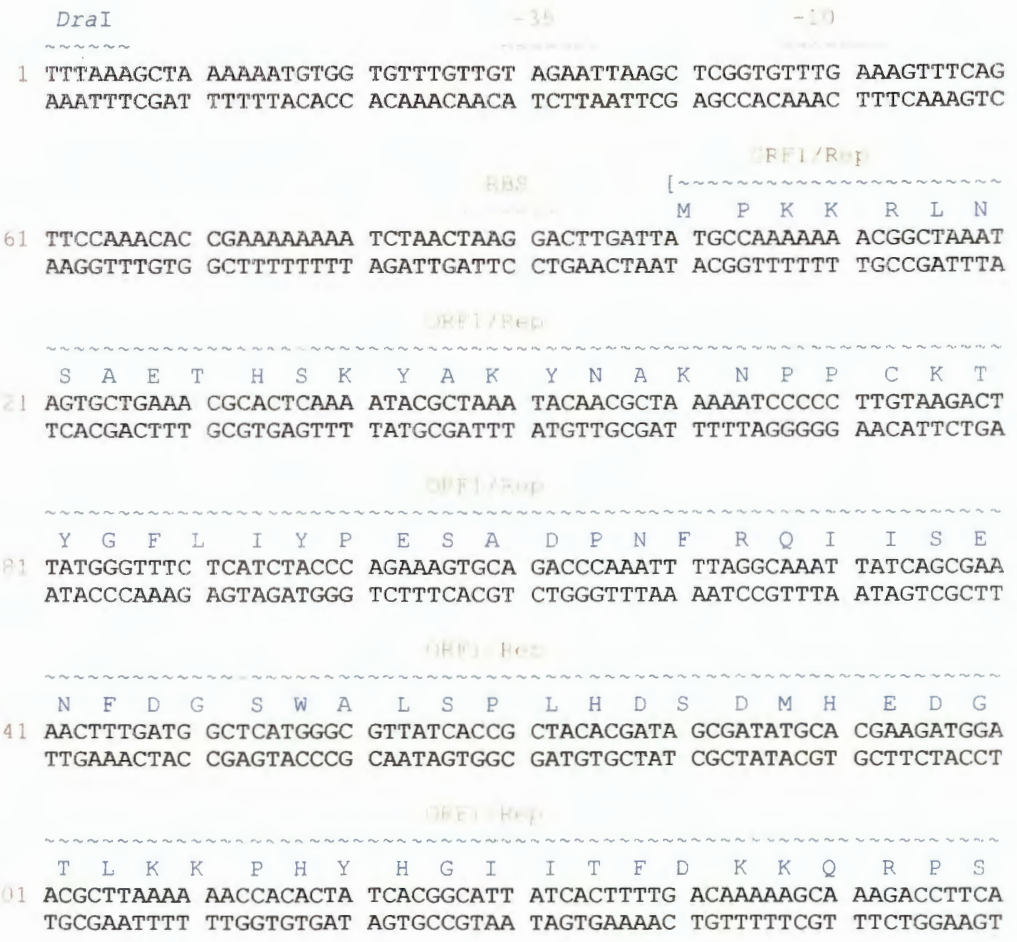

Figure 12. Nucleotide sequence of pAL236-2.
The single *Dra*I site at nucleotide one is indicated with a blue wavy line; -35 (30-35 nt) and -10 (47-52 nt) promoter regions and the ribosome binding site (RBS) (87-92 nt) are indicated by short red wavy lines; putative open reading frame (ORF) is indicated by a blue wavy arrow in the direction of translation; the amino acid sequence is shown in blue letters below the ORF; double-stranded origin (DSO) (797-818 nt) is indicated by red short wavy line with the nucleotides for the inverted repeats highlighted in red letters; the nucleotides involved with the DSO nick site (805-811 nt) are highlighted in blue letters; iterons (824-840, 847-863 nt) are indicated by red wavy lines with the nucleotides highlighted in blue; inverted repeats (IR) are indicated by red wavy lines with the nucleotides highlighted in blue. The inverted repeats are IR1 (1033-1039, 1044-1050 nt), IR2 (1057-1063, 1070-1075 nt), IR3 (1084-1089, 1095-1100 nt), IR4 (1119-1125, 1131-1137 nt), and IR5 (1202-1211, 1223-1232 nt).

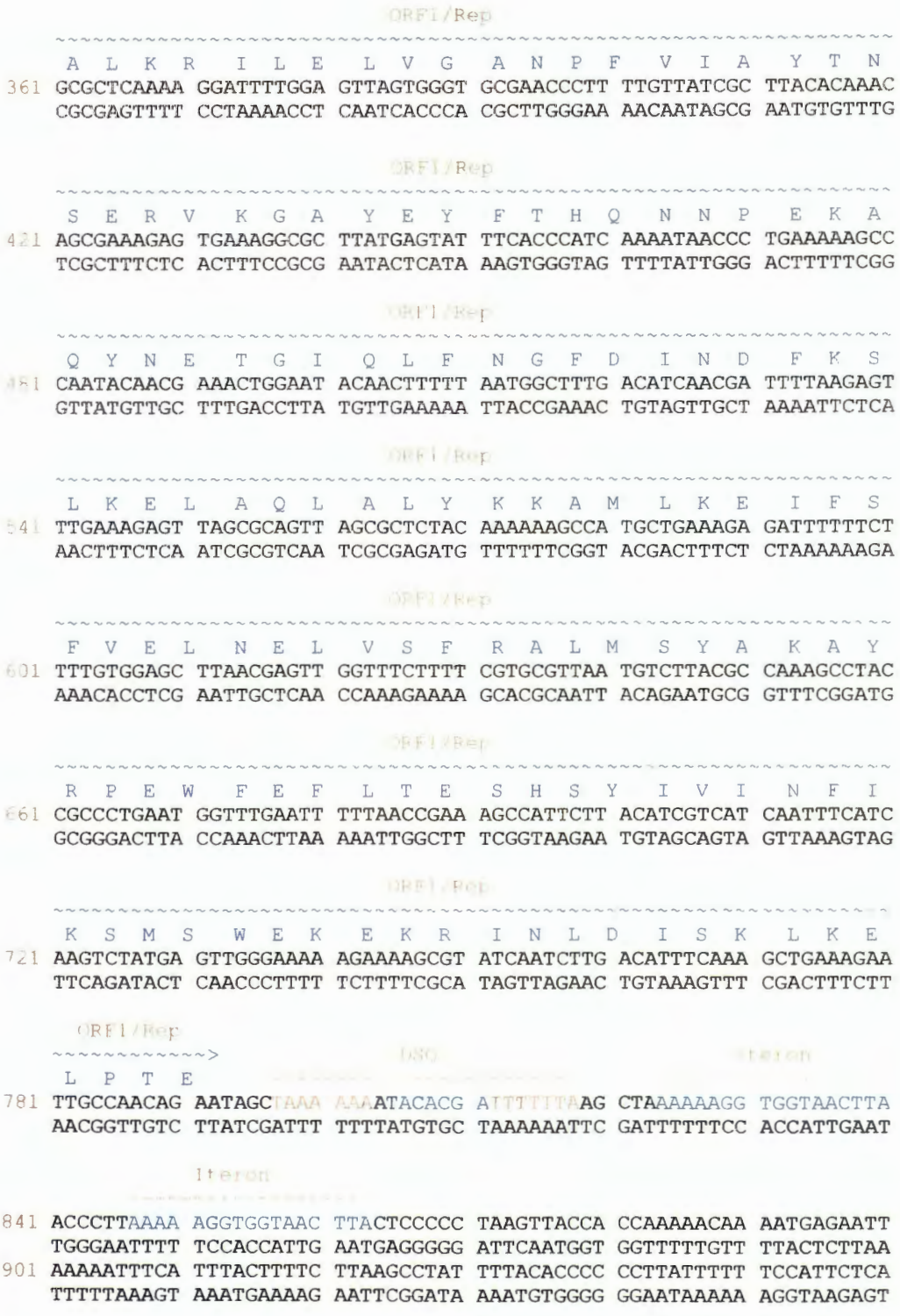

Figure 12 continued.

```
 961 TTTTCATAGG TGCTTAAAGT TTCTTTCCAA GCGTTAGCTA AAAACTTCAA TAGAATGAAA
     AAAAGTATCC ACGAATTTCA AAGAAAGGTT CGCAATCGAT TTTTGAAGTT ATCTTACTTT

                    IR1          IR1              IR2          IR2

1021 AGGGCAAGGA TACTTTGTGT AGCCACAAAG AACCTTGTTT AGGCGCTGCC CTAAAACCCT
     TCCCGTTCCT ATGAAACACA TCGGTGTTTC TTGGAACAAA TCCGCGACGG GATTTTGGGA

          IR3          IR3                        IR4          IR4

1081 TATTGGGGGA CAAGCCCCCA TGATAATGAG AATAATTTAA GGGGGATAAA CCCCCTTTTT
     ATAACCCCCT GTTCGGGGGT ACTATTACTC TTATTAAATT CCCCCTATTT GGGGGAAAAA
1141 TATTCTCATT ATTGGCGCAA AAAATGCGCC AAAATCTAAC CGGTGGTGGG TCGCGCCTTA
     ATAAGAGTAA TAACCGCGTT TTTTACGCGG TTTTAGATTG GCCACCACCC AGCGCGGAAT

          IR5          IR5

1201 CATGGGGGGG TGTCGTAGTA AAACCCCCCC ATTTTTTACC CCTAAATTTT TGCCCATTTT
     GTACCCCCCC ACAGCATCAT TTTGGGGGGG TAAAAAATGG GGATTTAAAA ACGGGTAAAA
1261 TTTTTGCTTT TTTTGGTGGG ATTGAAAGCC GAGGTTATAT GCGCTTTTTT TTGCTGAACG
     AAAAACGAAA AAAACCACCC TAACTTTCGG CTCCAATATA CGCGAAAAAA AACGACTTGC
1321 AGTGAGATTT GCTAACTGAT TAGCTTTTAA GTGTGAGTAA TTTTGCTTGA TTTTGATTTT
     TCACTCTAAA CGATTGACTA ATCGAAAATT CACACTCATT AAAACGAACT AAAACTAAAA
1381 TAATGGTGTT TTTGGTTGTT AGGTAATAAC TCTATTCGGT CTTTTGGTTT GTGTTTGTTA
     ATTACCACAA AAACCAACAA TCCATTATTG AGATAAGCCA GAAAACCAAA CACAAACAAT
1441 AAAACGAT
     TTTTGCTA
```

Figure 12 continued.

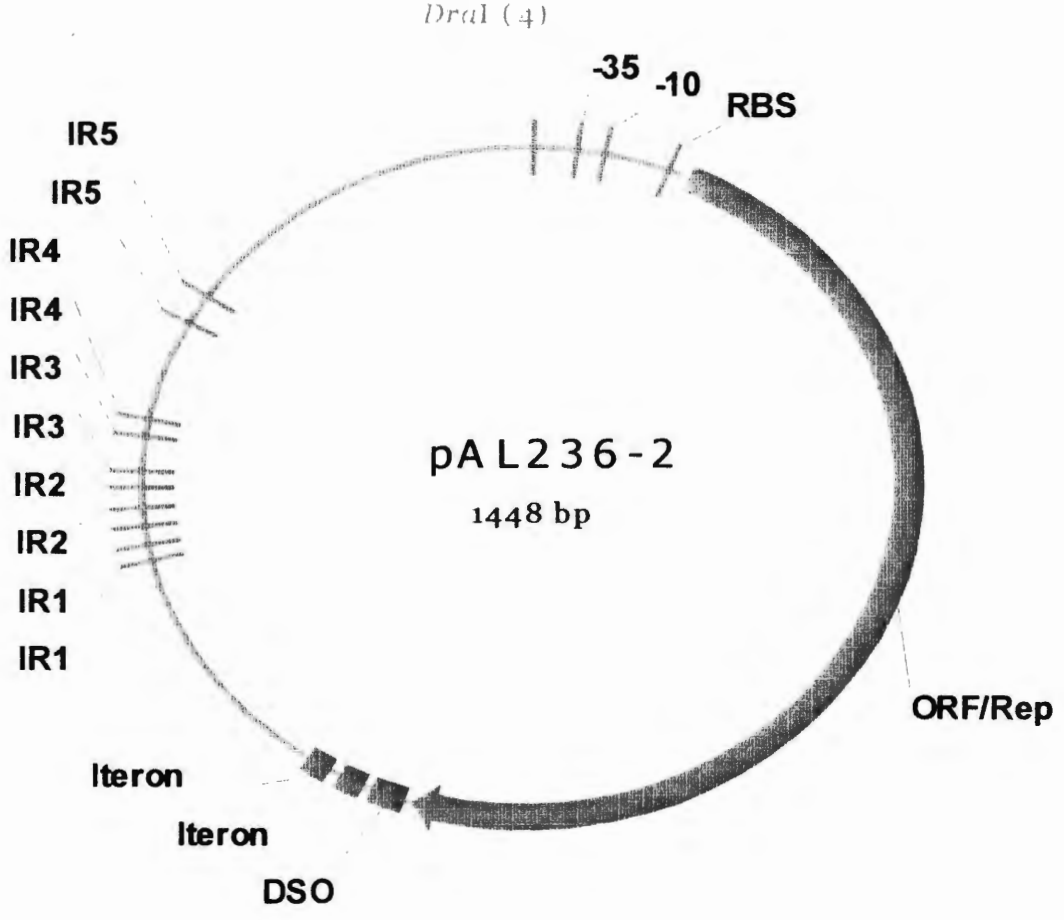

Figure 13. A diagram of pAL236-2 showing Promoter Regions, Ribosome Binding Site (RBS), Open Reading Frame (ORF), Inverted Repeats (IR), Iterons, and Double-Stranded Origin (DSO). Blue vertical lines represent promoters (-10 and -35), RBS, and IR. Maroon boxes represent DSO and iterons. Arrowhead represents putative ORF and the direction of translation. The single restriction endonucleases site, *Dra*I site is assigned as nucleotide number 1.

Table 6. Analysis of ORF1 from pAL236-2. a) ORF1 analysis by Vector NTI Advance™ 11. b) Promoter and RBS locations. c) Nucleotide identities of ORF1 with known plasmid sequences.

a)

ORF	Sequence	Length	M.W.[1]	IP[2]
	n.t.#	a.a#		
1	100-792	231	26951	8.69

1-Molecular weight; 2-Isoelectric point.

b)

Promoter		RBS
-35 5' TTGACA 3'	-10 5' TATAAT 3'	5' AGGAGG 3'
TAGAAT (30-35)	TTTGAA (47-52)	TAAGGA (87-92)

c)

ORF	Main Identities	Percent Identity	Proposed Function
1	Rep pHPK255 Rep pHP489 Rep pAL236-5	52 42 39	Plasmid Replication

Nucleotides 100-792, *orf1*, showed identity to the *orf1/rep* nucleotide sequences of *H. pylori* plasmids: pHPK255 (72%); pHP489 (71%); and pAL236-5 (55 %). The ORF1 amino acid sequence showed identity to the replication proteins, Rep, of *H. pylori* plasmids pHPK255 (52%), pHP489 (42%), and pAL236-5 (39 %). Of the three plasmids only pHPK255 showed the greatest amino acid identity with ORF1/Rep of pAL236-2 as shown in Figure 14.

Unlike other *H. pylori* plasmids which replicate by the theta mechanism, pHPK255 and pHP489 are plasmids that replicate via the rolling-circle (RC) mechanism (Kleanthous *et al.*, 1991, Song *et al.*, 2003). There were additional identities of ORF1/Rep to several Rep proteins of RC plasmids from unrelated species including the Gram-positive *Lactobacillus plantarum* (35%), *Lactococcus lactis* (31%), and *Staphylococcus aureus* (25%). Interestingly, no identities were found for of ORF1/Rep with any plasmid Rep proteins involved with theta replication from *H. pylori* or from unrelated species. As a result, ORF1/Rep identity with Rep proteins of three RC plasmids indicated that pAL236-2 should replicate as a RC plasmid.

The ORF1/Rep protein consisted of 231 amino acid residues. Submission of ORF1/Rep of pAL236-2 to BLASTP at NCBI, indicated by the Conserved Domain Search, that this protein contained a conserved domain belonging to the Rep_2 superfamily. According to the Conserved Domain Database (CDD) the Rep_2 superfamily are proteins involved with the initiation of plasmid replication (Marchler-Bauer *et al.*, 2004; Marchler-Bauer *et al.*, 2009). These proteins are topoisomerases that

Color Legend

	Foreground	Background
Non-homologous	black	white
Conservative	dark blue	light blue
Block of similar	black	light green
Identical	red	yellow
Weakly similar	green	white

Figure 14. Amino acid alignment of ORF1 compared with Rep from pHPK255.

nick the DNA at the positive strand at the plus origin of replication and at the single-strand conversion sequence. This result provides further evidence that pAL236-2 should replicate by the RC mechanism, since the proteins of pHPK255 and pHP489 also contain this conserved domain.

ORF1/Rep was analyzed by Vector NTI® Advance™ 11, which indicated a molecular weight of 26951 (26.9 kDa) and an isoelectric point of 8.69 (see Table 7a). Additional analysis on ORF1/Rep was performed by software programs available via the internet through ExPASy Proteomics tools. SOSUI and TMHMM determined that ORF1/Rep was a soluble protein with no transmembrane helices. ProtParam analysis determined that ORF1/Rep was stable with an instability index (II) of 33.03.

The non-coding regions of pAL236-2 were located at nt positions 1-99 and 793-1448, a total of 755 nt. Submission to BLASTN at NCBI indicated that the non-coding region had several short sequence identities with four non-coding regions of *H. pylori* plasmid pHPK255 as well as with two non-coding regions on plasmid pAL226, and one-non-coding region with both plasmid pHP489 and pHP666 (Table 7a). Interestingly, two of these plasmids, pHP666 and pAL226, are identified as plasmids that replicate by the theta mechanism, while pHPK255 and pHP489 are identified as RC plasmids. Results of the BLASTN search also revealed a short sequence of pAL236-2 (1018-1053 bp) with identity to a non-coding region of the genomic sequence of *H. pylori* strain P12 (Table 7b). Submission of non-coding sequence to BLASTX indicated no significant identity to any *H. pylori* chromosomal, plasmid, or non-related species proteins.

Table 7. Non-coding Region Identities of pAL236-2 with Other *H. pylori* Plasmids.

a. Plasmid features located within the non-coding region are indicated. (IR=Inverted Repeat)
b. Non-coding region identities of pAL236-2 with *H. pylori* genomic sequences.

a.

pAL236-2 Sequence (nt)	Main Plasmid Identities	Plasmid Sequence (nt)	Coding Region	Percent Identity with pAL236-2
1008-1148	pAL236-5	920-1058	(IR)	87%
1-99, 793-1448	pAL236-11	1663-2190	Non-Coding	35%
		1-327, 2551-3148	Non-Coding	41%
23-72	pHPK255	1392-1441	Non-coding	86%
930-965		519-554	Non-coding	88%
1103-1154		342-394	(IR)	86%
1275-1439		17-191	Non-coding	73%
1275-1345	pHP489	1152-1222	Non-coding	87%
1011-1053	pHP666	606-648	Non-coding	90%
1018-1050	pAL226	8561-8593	Non-coding	93%
1275-1342		9433-9500	Non-coding	80%

b.

pAL236-2 Sequence (nt)	Main Genomic Identities	Genomic Sequence (nt)	Coding Region	Percent Identity with pAL236-2
1018-1053	P12	180861-180896	Non-coding	94%

The search for repeated sequences revealed no copies of R1, R2, R3, R4 or *H. pylori* specific IS elements in pAL236-2. There were, however a set of iterons consisting of a two time 17 bp repeated sequence located at nt positions 824-840 and 847-863 (see Fig.12 and Fig.13). Plasmid pHP489 also contains a set of iterons consisting of a two time 57 bp repeated sequence, while plasmid pHPK255 has direct repeats (DR) consisting of a two time 7 bp repeat. Also found on pAL236-2 were several IR sequences: IR1; IR2; IR3; IR4; and IR5 (see Fig. 12 and Fig. 13). Plasmids pHPK255 and pHPK489 also contain IR sequences, but only IR3 of pAL236-2 was similar to an IR sequence found on pHPK255 (see Table 7a). Also IR1 and IR3 were exactly identical to IR sequences found on pAL236-5, while IR2 and IR4 were similar to IR sequences on pAL236-5 (see below). The roles of these IR repeats in pAL236-2 and pAL236-5 are not known and require further research.

An important feature indicating that pAL236-2 replicated by the RC mechanism was the presence of a possible double-stranded-origin (DSO), a feature belonging to plasmids that replicate via the RC mechanism. Several plasmids have been identified as RC plasmids and initially were grouped into five families based on the homology of their replication proteins and DSOs (Khan, 1997; Khan, 2005). As more DNA sequences of RC plasmids have been determined there are now 17 RC plasmid groups (http://www.essex.ac.uk/bs/staff/osborn/DPR_home.htm) based on homologies of the Rep proteins and DSOs (Khan, 2005).

The DSO of pAL236-2 was located at nt position 797-818 (see Fig. 12 and Fig. 13). This inverted repeated sequence contained the nucleotides that could be involved

with the DSO nick site (805-811 nt). The DSO could potentially form a hairpin-loop structure with the nick site located at the top of the hairpin-loop (Fig. 15). Downstream of the DSO were the iteron sequences that could the binding site of the replication protein to allow access to the nick site to initiate plasmid replication.

Of the 17 groups of RC plasmids described by Khan (2005), pAL236-2 could belong to Group II. Evidence for this includes the homologies of the Rep protein with other members of this group including pHPK255 and pHP489 (http://www.essex.ac.uk/bs/staff/osborn/DPR_home.htm). Also, most members within this group, including pHPK255 and pHP489, contain a conserved DSO nick sequence (TACTACG/A) where the / represents the cleavage site by the Rep protein that initiates leading-strand replication (del Solar *et al.*, 1993; Kleanthous *et al.*, 1991; del Solar *et al.*, 1998). However, the pAL236-2 potential nick sequence (TACACG/A) was not identical. This nick site was chosen as the result of the iteron sequences that were located downstream of the potential DSO nick site (Fig. 15). While the nick site is highly conserved within a family of RC plasmids, the binding regions of these plasmids can vary (del Solar *et al.*, 1993; del Solar *et al.*, 1998; Khan, 1997; Khan, 2000; Khan, 2005). Distances of these binding sequences from the nick site vary or can be in close proximity to the nick site depending on the RC plasmid family (del Solar *et al.*, 1993; del Solar *et al.*, 1998; Khan, 1997; Khan, 2000; Khan, 2005).

It should be noted that a potential nick sequence (TACTACG/A) was present between the two IR5 sites, but no iteron sequences or other direct repeats were present, unlike the iterons found for the proposed DSO nick site of pAL236-2 (see Fig. 12).

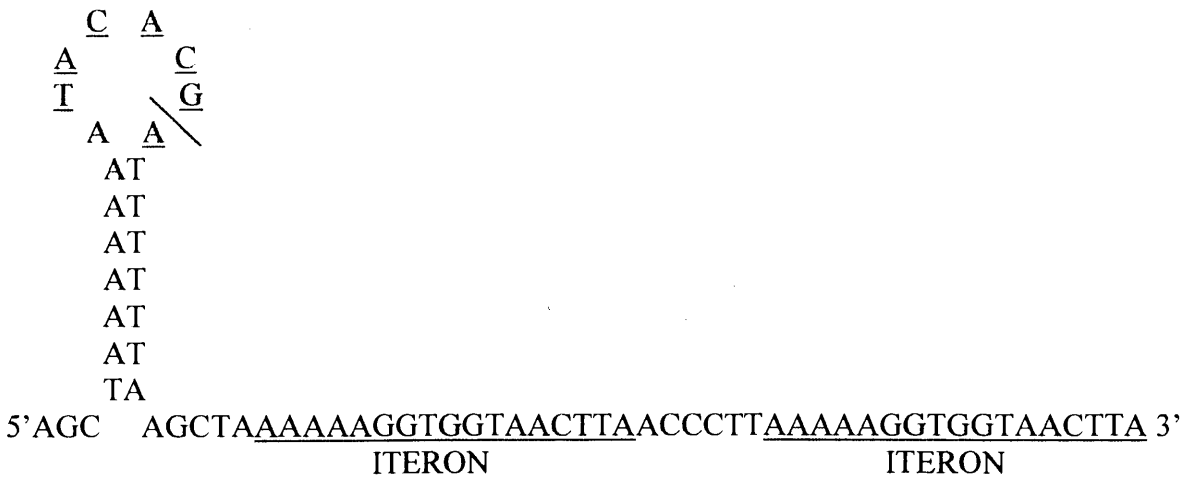

Figure 15. Graphic representation of the potential Hairpin-Loop Structure in the Double-Stranded Origin (DSO) of pAL236-2. The nick site sequence (TACACG/A) is located at the top of the hairpin-loop structure with a line representing where the replication protein nicks the plasmid DNA. Downstream of the hairpin-loop structure are two iteron repeats where the replication protein could bind allowing access to the nick site to initiate plasmid replication.

Even though there are replication proteins of RC plasmids that can bind to IR sequences located within the Rep binding region (Khan, 1997), the analysis here provide evidence that the DSO indicated in Figure 13 is the functional site.

Additional evidence that pAL236-2 is a Group II RC plasmid was found in the Rep protein itself. Located on the Rep protein were several protein motifs that are found on RC plasmids (Fig. 16 and Fig. 17). In most RC Rep proteins there are three conserved motifs (Ilyina *et al.*, 1992). However, further research of certain Group II RC plasmids, including pHPK255, revealed not only the three previously reported protein motifs but two additional motifs (del Solar *et al.*, 1993). Plasmid pAL236-2 had all five of the protein motifs as did pHPK255 and pHP489 (Fig. 17).

Motif 1 and Motif 2 functions have not been determined however it is speculated that these motifs could be necessary for recognizing DNA sequences around the nick site or in the nick site (del Solar *et al.*, 1993). Motif 3 contains a possible metal-binding domain HUH where U represents hydrophobic residues (Ilyina *et al.*, 1992). It appears that Group II RC plasmids share two conserved tyrosine residues found in motifs 4 and 5 (del Solar *et al.*, 1993). Motif 4 represents the catalytic domain that contains a catalytic tyrosine residue involved with the nucleophilic attack on the plasmid DNA for the initiation of replication (Ilyina *et al.*, 1992). Motif 5 also contains a conserved tyrosine residue that could be involved in linking the Rep protein to the plasmid DNA (del Solar *et al.*, 1993). Even though the sequences of these motifs can vary among RC plasmids, the metal-binding domain and the conserved tyrosine residue utilized to initiate plasmid

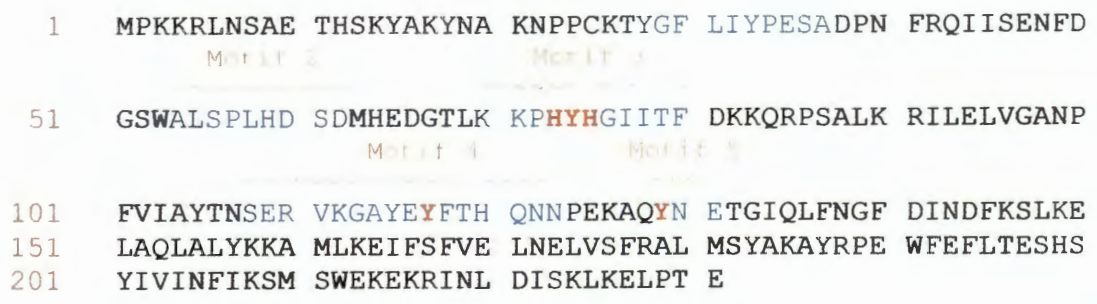

Figure 16. Replication motifs of pAL236-2. Amino acid sequence with motifs are represented by red wavy lines and amino acid residues indicated by blue letters; motif 1 (28-37 aa), motif 2 (54-62 aa), motif 3 (70-80 aa), motif 4 (108-123 aa) and motif 5 (129-131 aa). The possible metal-binding domain of motif 3 is indicated by bold red letters. A conserved tyrosine residue of motif 4 and motif 5 is indicated by a bold red letter.

Color Legend

	Foreground	Background
Non-homologous	black	white
Conservative	dark blue	light blue
Block of similar	black	light green
Identical	red	yellow
Weakly similar	green	white

Figure 17. Replication motifs of pAL236-2 compared to pHP489 and pHPK255. Protein motifs are represented by boxes. Possible metal-binding domain (HUH) is indicated by underlined bold letters. Conserved tyrosine residues (Y) is indicated by an underlined bold letter.

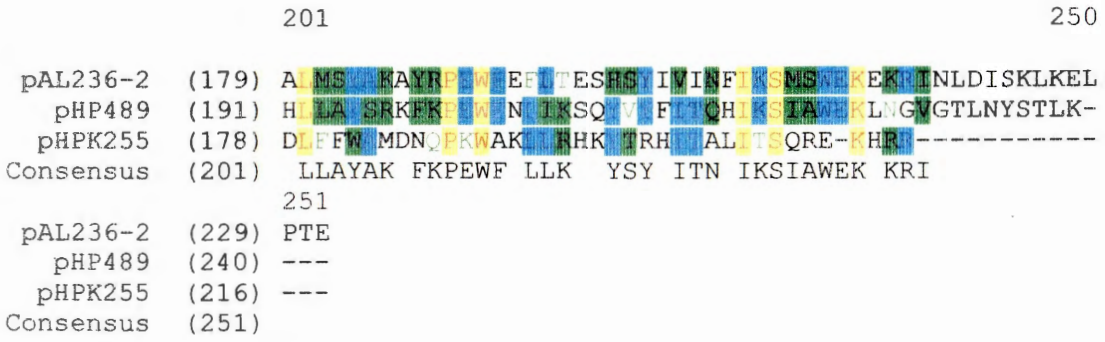

Figure 17 continued.

replication are found within members of most RC plasmids families, although one family of RC plasmids does not contain the metal-binding motif (del Solar *et al.*, 1998).

The single-stranded origin (SSO) represents the final stage of RC plasmid replication involved with lagging strand replication by converting single-stranded DNA (ssDNA) into double-stranded DNA (dsDNA) (del Solar *et al.*, 1998; Khan, 1997; Khan, 2000; Khan, 2005). A feature that is considered a hallmark for RC plasmid replication is the presence of ssDNA (Khan, 1997). A common feature found among SSOs is the ability to form secondary structures consisting of hairpin-loop structures (Khan, 1997). Based on sequence and structural similarities, there are a variety of SSOs in RC plasmids and the most extensively studied include *ssoA*, *ssoU*, *ssoT*, and *ssoW* (Boe *et al.*, 1989, Novick, 1989, Gruss *et al.*, 1989; Seegers *et al.*, 1995; Devine *et al.*, 1989, Zaman *et al.*, 1993, Meijer *et al.*, 1995; del Solar *et al.*, 1998; Khan, 1997; Khan, 2000; Khan, 2005). One feature of the SSO nucleotide sequences is that they vary among different RC plasmid families (Khan, 1997).

The DNA sequence of pAL236-2 was compared to known SSOs, including *ssoA*, *ssoU*, *ssoT*, and *ssoW* as well as the SSO found on pHPK255. No SSO sequences were identified on pAL236-2. Unlike pHPK255, plasmid pHP489 also appears to have no SSO. RC plasmids that do not contain an SSO can replicate to some extent and deletions in the SSO can lead to a reduced copy number, large quantities of ssDNA, and plasmid instability (del Solar *et al.*, 1987; Gruss *et al.*, 1987). Due to the variety of SSOs that have been reported a possibility does exist that an unknown sequence or sequences could

act as a SSO on pAL236-2 to help convert ssDNA to dsDNA. Further research is needed to determine this possibility.

pAL236-5 DNA AND ORF ANALYSIS

The complete sequence of pAL236-5 is shown in Figure 18. The overall GC content of pAL236-5 was 33.8 %, which is lower than that of *H. pylori* chromosomal DNA of strains P12 (38.0%), Shi470 (38.0%), B128 (38.0%), 26695 (38.9%), G27 (38.9%), HPKX_438_CA4C1 (39.0%), HPKX_438_AG0C1 (39.0%), HPAG1 (39.1%), and J99 (39.2%). The GC content of pAL236-5 was similar to *H. pylori* plasmids pHPM185 (33.1%), pHP489 (33.3%), pHPM8 (33.4%), pHP69 (34.0%), pHPG27 (34.0%), pAL202 (34.3%), pHel4 (34.3%) and pHel5 (34.4%). The GC content of pAL236-5 was lower compared with other *H. pylori* plasmids pHPP12 (35.0%), pHPP12 (35.0%), pHel1 (35.1%), pHP51 (35.3%), pHPM179a (35.5%), pHP666 (35.6%), pAL226 (35.8%), pHPM186 (35.9%), pHPM179b (36.2%), pHPO100 (36.2%), pHPAG1 (36.4%), pHPK255 (36.8%), pHPM180 (37.0%), pHPS1 (37.3%), pAL236-2 (37.5%), and pKU701 (38.0%) (Table 8).

The complete sequence of pAL236-5 contained one ORF. The amino acid sequence of ORF1 is indicated in Figure 18. A circular diagram of the plasmid with ORF1 is shown in Figure 19. Possible -10 and -35 promoter sequences and ribosome binding site (RBS) for ORF1 were determined using the consensus sequences of *E. coli* (Table 9). Table 9c also shows the amino acid identities between ORF1 of pAL236-5 and other *H. pylori* ORFs.

Figure 18. Nucleotide sequence of pAL236-5.
The single *Eco*RI site at nucleotide one is indicated with a blue wavy line; -35 (21-26 nt) and -10 promoter regions (44-49 nt) and the ribosome binding site (RBS) (103-108 nt) are indicated by short red wavy lines; DnaA-binding site (76-84 nt) is indicated by a red wavy line with the nucleotides highlighted in blue; putative open reading frame (ORF) is indicated by a blue wavy arrow in the direction of translation; amino acid sequence is shown in blue letters below the ORF; direct repeats (DR) (1075-1079, 1082-1086 nt) are indicated by red wavy lines with the nucleotides highlighted in blue; double-stranded origin (DSO) (1090-1147 nt) is indicated by a red wavy line with the inverted repeats IR5 (1090-1100, 1137-1147 nt) and IR6 (1109-1120, 1124-1135 nt) indicated by red wavy lines with the nucleotides highlighted in red letters; nucleotides involved with the DSO nick site (1101-1108 nt) are highlighted in blue; inverted repeats (IR) are indicated by red wavy lines with the nucleotides highlighted in blue. The inverted repeats are IR1 (946-952, 957-963 nt), IR2 (970-977, 981-988 nt), IR3 (998-1003, 1009-1014 nt), IR4 (1033-1038, 1044-1049 nt).

Figure 18 continued.

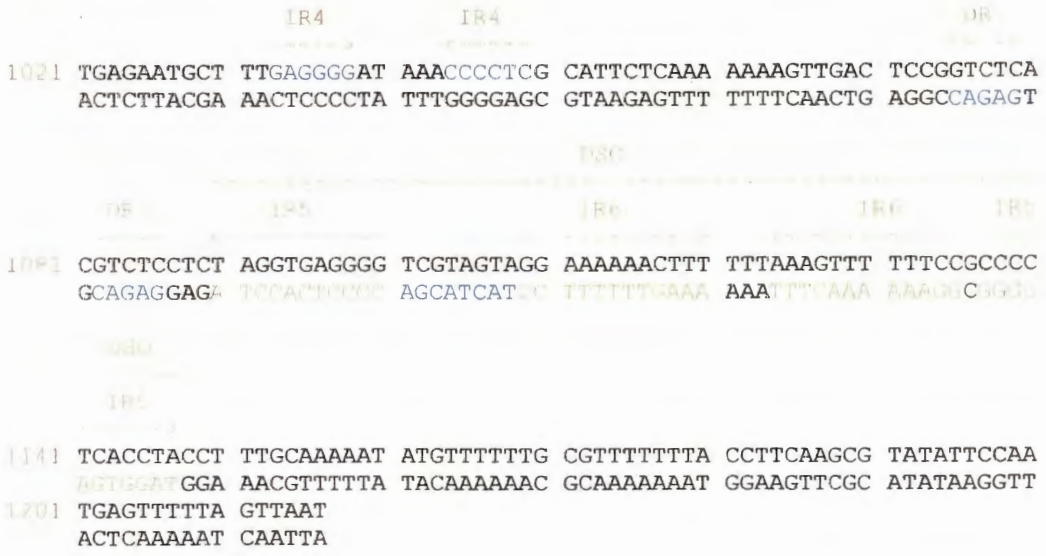

Figure 18 continued.

Table 8. Percent GC of *H. pylori* Plasmid and Chromosomal DNA.

%GC	Plasmid/Chromosomal DNA	Geographic Region
33.1	pHPM185	North Texas
33.3	pHP489	Korea
33.4	pHPM8	North Texas
33.8	pAL236-5	Alaska
34.0	pHPG27	Italy
34.0	pHP69	Korea
34.3	pAL202	Alaska
34.3	pHeL4	Germany
34.4	pHeL5	Germany
35.0	pHPP12	Germany
35.1	pHeL1	Germany
35.3	pHP51	Korea
35.5	pHPM179a	North Texas
35.6	pHP666	Italy
35.8	pAL226	Alaska
35.9	pAL236-11	Alaska
35.9	pHPM186	North Texas
36.2	pHPM179b	North Texas
36.2	pHPO100	China
36.4	pHPAG1	Sweden
36.8	pHPK255	United Kingdom
37.0	pHPM180	North Texas
37.3	pHPS1	Australia
37.5	pAL236-2	Alaska
38.0	pKU701	Japan
38.0	P12*	Germany
38.0	Shi470*	Peru
38.0	B128*	United States
38.9	26695*	United Kingdom
38.9	G27*	Italy
39.0	HPKX_438_CA4C1*	Sweden
39.0	HPKX_438_AG0C1*	Sweden
39.1	HPAG1*	Sweden
39.2	J99*	United States

* Chromosomal DNA

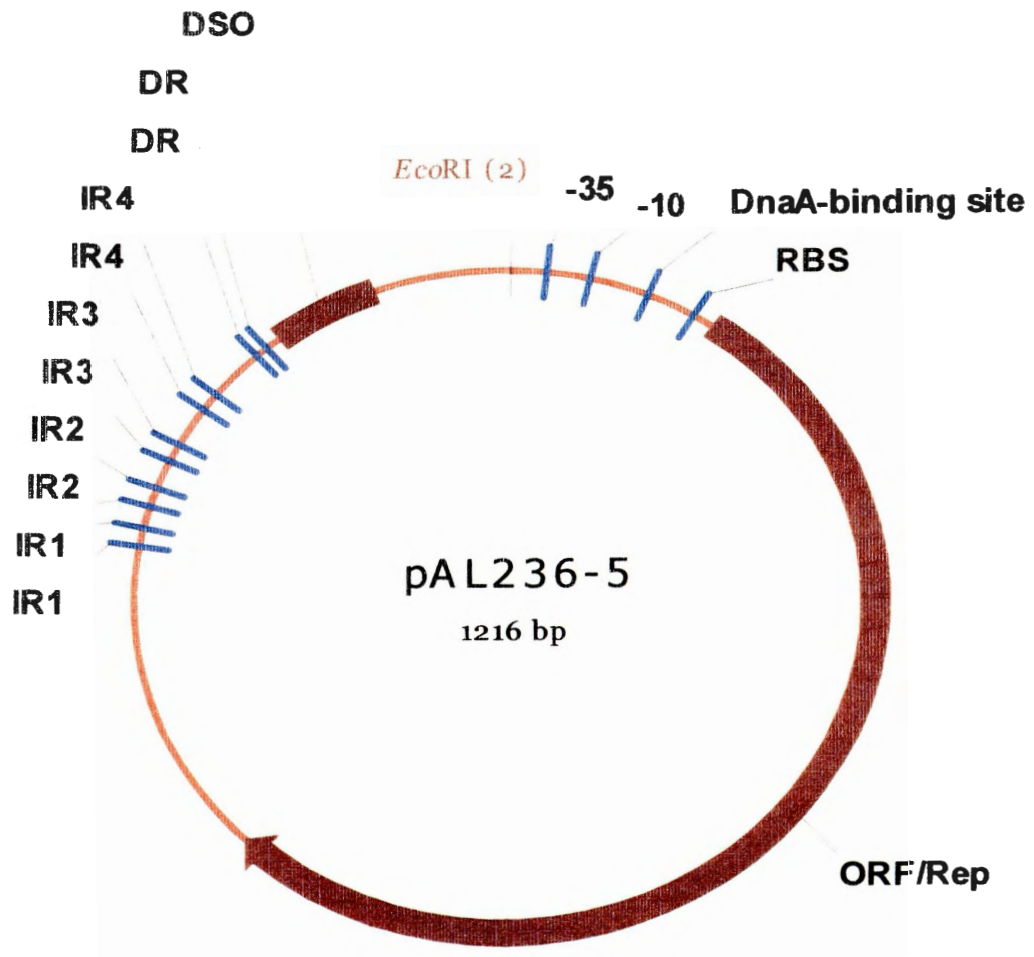

Figure 19. A diagram of pAL236-5 showing Promoter Regions, Ribosome Binding Site (RBS), Open Reading Frame (ORF), Inverted Repeats (IR), Direct Repeats (DR), DnaA-Binding Site, and Double-Stranded Origin (DSO). Blue vertical lines represent promoters (-10 and -35), RBS, IR, DR, and DnaA-binding site. Maroon box represents DSO. Arrowhead represents putative ORF and the direction of translation. The single restriction endonucleases site, *Eco*RI was assigned as nucleotide number 1.

The ORF1 amino acid sequence of pAL236-5 showed identity to the replication protein, Rep, of *H. pylori* plasmids pHPK255, pHP489, and pAL236-2 (see Table 9c). The highest alignment, with pHP489, is shown in Figure 20. As for nucleotide identity, the *orf1* gene, nt 114-758, showed identity to the *orf1/rep* nucleotide sequences of three *H. pylori* plasmids: pHPK255 (78%); pHP489 (85%); and pAL236-2 (55 %).

Unlike other plasmids found in *H. pylori*, which replicate by the theta mechanism, pHPK255, pHP489 and pAL236-2 replicate via the RC mechanism. There were additional identities of ORF1/Rep to several Rep proteins of RC plasmids from unrelated species including *Lactobacillus plantarum* (39%), *Lactococcus lactis* (37%), and *Staphylococcus aureus* (41%). Interestingly, ORF1/Rep of pAL236-5 appeared to have no identities with any plasmid Rep proteins involved with theta replication from *H. pylori* or from unrelated species. As a result, ORF1/Rep identity with Rep proteins of RC plasmids pHPK255, pHP489, and pAL236-2 along with Rep proteins of RC plasmids from unrelated species indicated that pAL236-5, like pAL236-2, should replicate as a RC plasmid.

The ORF1/Rep protein consisted of 215 amino acid residues. Submission of the ORF1/Rep of pAL236-5 to the Conserved Domain Search of BLASTP at NCBI indicated that this protein contained a conserved domain that belongs to the Rep_2 superfamily. This result indicated pAL236-5 should replicate by the RC mechanism as replication proteins of pHPK255, pHP489, and pAL236-2 also contained this conserved domain.

Table 9. Analysis of ORF1 from pAL236-5. a) ORF1 analysis by Vector NTI Advance™ 11. b) Promoter and RBS locations. c) Nucleotide identities of ORF1 with known plasmid sequences.

a)

ORF	Sequence	Length	M.W.[1]	IP[2]
	n.t.#	a.a#		
1	114-758	215	25494	7.26

1-Molecular weight; 2-Isoelectric point.

b)

Promoter		RBS
-35 5' TTGACA 3'	-10 5' TATAAT 3'	5' AGGAGG 3'
TTGATA (21-26)	TAGAAT (44-49)	AGGACA (103-108)

c)

ORF	Main Identities	Percent Identity	Proposed Function
1	Rep pHP489 Rep pHPK255 Rep pAL236-2	86 38 39	Plasmid Replication

Color Legend

	Foreground	Background
Non-homologous	black	white
Conservative	dark blue	light blue
Block of similar	black	light green
Identical	red	yellow
Weakly similar	green	white

Figure 20. Amino acid alignment of ORF1 with Rep from pHP489.

ORF1/Rep was analyzed using Vector NTI® Advance™ 11 which indicated that the protein had a molecular weight of 25494 (25 kDa) and a isoelectric point of 7.26 (see Table 9a). Additional analysis of ORF1/Rep was performed by software programs available through ExPASy Proteomics tools. SOSUI determined ORF1/Rep is soluble. Analysis with TMHMM determined that ORF1/Rep did not contain any transmembrane helices. ProtParam determined that ORF1/Rep was unstable with an instability index (II) of 41.41.

The region of pAL236-5 with no ORFs was determined to contain non-coding sequences. The non-coding sequences were located at nt positions 1-113 and 759-1216, a total of 571 nt. Results from submission to BLASTN at NCBI, indicated that the non-coding region of pAL236-5 had short sequence identities with two non-coding regions of other *H. pylori* plasmids: pHP489; pHPP12; pHPM180; pHPS1; pAL226; and pHPAG1. These results are shown in Table 10. The identities with pHPM180, pHPS1, pHPAG1, and pHPP12 included the R2 repeat regions found on these plasmids. Plasmid pAL236-5 also had several identities with non-coding regions of *H. pylori* plasmids: pHPG27; pHel4; pHP69; pHPO100; pHPM8; pHel1; and pHP666, also shown in Table 10. The non-coding region identity with pHP69 included the common region (C1), also shown in Table 10. Of these plasmid identities, the non-coding region of pAL236-5 only had identity with one RC plasmid, pHP489, and did not have any identity with the RC plasmid pHPK255. Interestingly, the remaining identities were with plasmids that

Table 10. Non-coding Region Identities of pAL236-5 with Other *H. pylori* Plasmids. Plasmid features located within the non-coding region are indicated. (IR= Inverted Repeat, R2=Repeat Regions, C1=Common Region). 1- R2 repeat regions identified in this study.

pAL236-5 Sequence (nt)	Main Plasmid Identities	Plasmid Sequence (nt)	Coding Region	Percent Identity with pAL236-5
920-1058	pAL236-2	1008-1148	(IR)	87%
1-113,759-1216	pAL236-11	1663-2190 1-327, 2551-3148	Non-Coding Non-Coding	42% 32%
804-873 1065-1148	pHP489	1152-1221 1048-1131	Non-Coding Non-Coding	78% 91%
795-898	pHPP12	5892-5995 6621-6724	(R2)[1] (R2)[1]	85% 85%
787-898	pHPG27	9594-9705	Non-Coding	80%
787-898 825-898	pHPM180	1764-1837 2032-2144	(R2) (R2)	85% 77%
828-898	pHel4	5770-5840	Non-Coding	87%
803-898 931-963	pAL226	9406-9501 931-963	Non-Coding Non-Coding	78% 90%
804-899	pHP69	8657-8752	(C1)	83%
804-898	pHPO100	605-700	Non-Coding	78%
795-894	pHPS1	2861-2960 5400-5499	(R2) (R2)	81% 81%
795-898	pHPAG1	2921-3024 8910-9013	(R2)[1] (R2)[1]	75% 75%
795-867	pHPM8	6-78	Non-Coding	78%
795-868	pHel1	9-82	Non-Coding	77%
931-964	pHP666	613-646	Non-Coding	91%

replicate by the theta mechanism. Results of the BLASTN search also revealed a short sequence of pAL236-5 (931-954 nt) that had identity to a non-coding region of the genomic sequence of *H. pylori* strain P12 (Table 11). The non-coding region of pAL236-5 was submitted to BLASTX, which indicated that the pAL236-5 non-coding region did not have significant identity to any *H. pylori* chromosomal, plasmid, or non-related species proteins.

The search for repeated sequences revealed that pAL236-5, like pAL236-2, did not have R1, R3, R4 repeat regions or *H. pylori* specific IS elements. Although pAL236-5 had a small identity to the R2 repeats found on pHPM180, pHPP12, pPHAG1 and pHPS1 (see Table 10), no complete set of R2 repeats was found on pAL236-5. Plasmid pAL236-5 did not contain any iteron sequences, unlike the iterons found on RC plasmids pHP489, pAL236-2, and other *H. pylori* plasmids that replicate by the theta mechanism. Instead of containing iterons, pAL236-5 had a two time 5 bp direct repeat (DR) (see Fig. 18 and Fig. 19). Direct repeats are found also on RC plasmid pHPK255, where they consist of a two time 7 bp repeat. They were, however, not identical to the DRs of pAL236-5. Also found on pAL236-5 were several inverted repeat (IR) sequences which included IR1, IR2, IR3, and IR4 (see Fig. 18 and Fig. 19). Plasmids pHPK255 and pHPK489 also contain IR sequences, but only IR3 of pAL236-5 was similar to an IR sequence found on pHPK255, similar to the IR3 of pAL236-2. Besides identity with the Rep of pHPK255 this IR3 appeared to be the only other area of identity between pAL236-5 and pHPK255. Also, IR1 and IR3 were exactly identical to IR sequences

Table 11. Non-coding Region Identities of pAL236-5 with *H. pylori* Genomic Sequences.

pAL236-5 Sequence (nt)	Main Genomic Identities	Genomic Sequence (nt)	Coding Region	Percent Identity with pAL236-5
931-954	P12	180863-180896	Non-coding	91%

found on pAL236-2, while IR2 and IR4 were similar to IR sequences on pAL236-2. The exact role these IR repeats have in pAL236-5 is not known and requires further research.

An important feature indicating that pAL236-5 replicates by the RC mechanism consisted of a possible double-stranded origin (DSO). The DSO of pAL236-5 was located at nt positions 1090-1147 (see Fig. 18 and Fig. 19). It consisted of two inverted repeat sequences IR5 (1090-1100, 1137-1147 nt) and IR6 (1109-1120, 1124-1135 nt) (see Fig. 18 and Fig.19). Within the DSO region, besides IR5 and IR6, were the nucleotides that could be involved with the DSO nick site (1101-1108 nt) (see Fig. 18 and Fig. 19). The DSO could potentially form a hairpin-loop structure with the nick site located at the side of the hairpin-loop (Fig. 21). Downstream of the DSO were the DR sequences that could be involved with attachment of the replication protein that would allow access to the nick site to initiate plasmid replication (Fig. 21).

Out of the 17 groups of RC plasmids described by Khan (2005), pAL236-5, like pAL236-2, could belong to Group II. Evidence for this included the homologies of the Rep protein with other members of this group including pHPK255 and pHP489 (http://essex.ac.uk/bs /staff/osborn/DPR_home.htm). Plasmid pAL236-5, like pAL236-2, along with other members of this group, contain a conserved DSO nick sequence (TACTACG/A) where the / represents the cleavage site by the Rep protein that initiates leading-strand replication (del Solar *et al.*, 1993; Kleanthous *et al.*, 1991; del Solar *et al.*, 1998). Plasmid pAL236-5 contained the exact same DSO nick sequence that is found on pHPK255 and pHP489. Additional evidence indicated that pAL236-5 was an RC plasmid and belonged with Group II of RC plasmids, similar to pAL236-2, was found

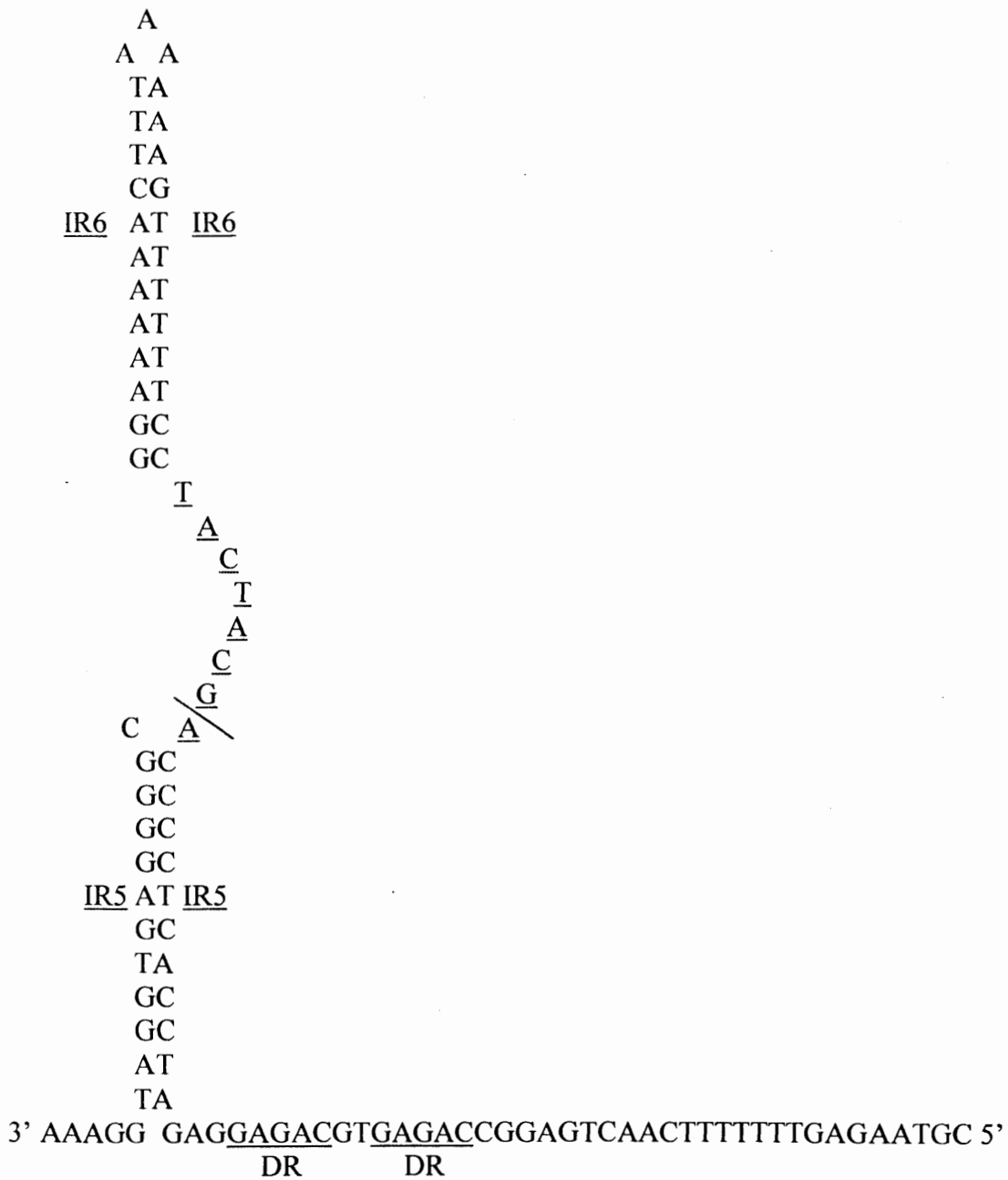

Figure 21. Graphic representation of the potential Hairpin-Loop Structure in the Double-Stranded Origin (DSO) of pAL236-5 formed by Inverted Repeats (IR) 5 and 6. The nick site sequence (TACTACG/A) is located on this side of the hairpin-loop structure with a line representing where the replication protein of pAL236-5 nicks the plasmid DNA. Downstream of the hairpin-loop structure are two direct repeats where the replication protein could bind allowing access to the nick site to initiate plasmid replication.

within the Rep protein itself. Located on the Rep protein was several protein motifs that can be found on RC plasmids (Fig. 22 and Fig. 23). Plasmid pAL236-5 appeared to have all five of the protein motifs as well as pHPK255, pHP489, and pAL236-2 (Fig. 23).

The next feature examined on pAL236-5 was the single-stranded origin (SSO). The DNA sequence of pAL236-5 was compared to identified SSOs including *ssoA*, *ssoU*, *ssoT*, and *ssoW* and the SSO found on pHPK255. As a result, pAL236-5 appeared to have no sequences that could be involved with an SSO. However, unlike pHPK255, plasmids pHP489 and pAL236-2 appeared to lack an SSO. As described previously with pAL236-2, a possibility does exist that an unknown sequence or sequences could act as an SSO on pAL236-5 to help convert ssDNA to dsDNA. However, further research is needed to confirm this possibility.

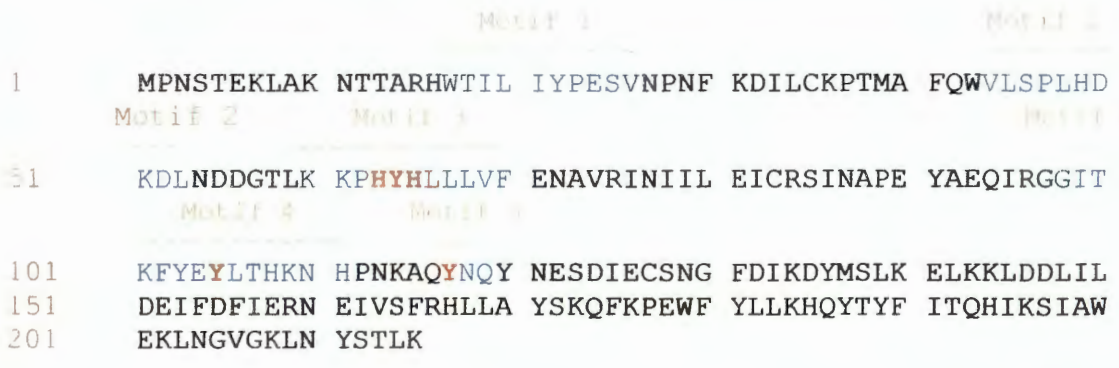

Figure 22. Replication motifs of pAL236-5. Amino acid sequence with motifs represented by red wavy lines and amino acid residues indicated by blue letters; motif 1 (17-26 aa), motif 2 (44-53 aa), motif 3 (60-70 aa), motif 4 (98-111 aa), and motif 5 (117-119).

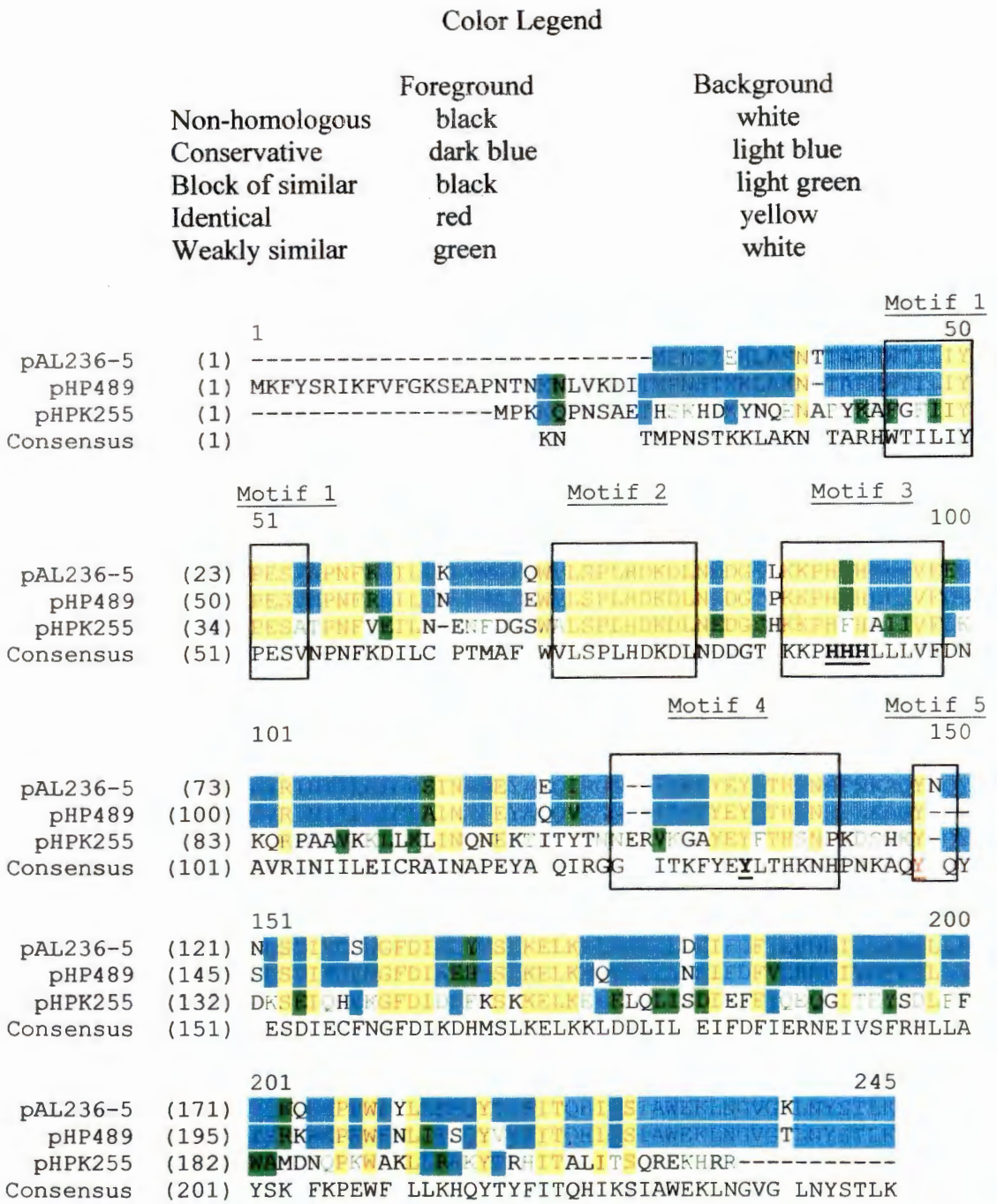

Figure 23. Replication motifs of pAL236-5 compared to pHP489 and pHPK255. Replication motifs of pAL236-5 compared to pHP489 and pHPK255. Protein motifs are represented by boxes. The possible metal-binding domain of motif 3 is indicated by bold red letters. A conserved tyrosine residue of motif 4 and motif 5 is indicated by a bold red letter.

ANALYSIS OF NON-CODING REGIONS BETWEEN pAL236-2, pAL236-5, AND pAL236-11

Plasmid pAL236-11 has two non-coding regions located at the end of ORF1 to the beginning of ORF2 at nt positions 1663-2190 (538 nt) and from the end of ORF2 to the beginning of ORF1 at nt positions 1-327 and 2551-3148 (925 nt) previously shown in Figures 8 and 9. The non-coding region of pAL236-2 is located from the end of ORF1/Rep up to the beginning of ORF1/Rep at nt positions 1-99 and 793-1448 (755 nt) previously shown in Figures 12 and 13. For plasmid pAL236-5, the non-coding region is located from the end of ORF1/Rep up to the beginning of ORF1/Rep at nt positions 1-113 and 759-1216 (571 nt) previously shown in Figures 18 and 19. Each non-coding region of pAL236-2, pAL236-5, and pAL236-11 were compared and analyzed with each other utilizing AlignX® on Vector NTI® Advance™ 11.

Results of comparing both non-coding regions of pAL236-11 against the non-coding region of pAL236-2 indicated identities of 35% and 41% respectively (see Table 5). When comparing the non-coding regions of pAL236-11 against the non-coding region of pAL236-5 results indicated identities of 42% and 32% respectively (see Table 5). Based on the alignment of these nucleotide sequences no significant area of identity was observed.

The non-coding region of pAL236-2 was compare with the non-coding region of pAL236-5. Result indicated that both regions had 48% identity with each other. However, when comparing these sequences only a small area of identity was observed. As a result, 141 nt of pAL236-2 (1008-1148 nt) was analyzed against 139 nt of

pAL236-5 (920-1058 nt). Interestingly, these two regions shared 87% identity and appeared to be the region that contained the inverted repeats (IR) found on both plasmids (see Tables 7a and 10).

PHYLOGENETIC ANALYSIS OF REPLICATION PROTEINS FROM *H. pylori* WITH GAPS INCLUDING pAL236-2, pAL236-5, AND pAL236-11

The replication protein sequences were obtained and analyzed from the following *H. pylori* plasmids: pAL236-2; pAL236-5; pAL236-11; pAL202; pAL226; pHPM8; pHPM179; pHPM180; pHPM185; pHPM186; pMCU1 through pMCU7; pHel1; pHel4; pHel5; pHP69; pHP666; pHAG1; pHPG27; pHPP12; pHPS1; pHP51; pHPO100; pKU701; pHP489; and pHPK255 (see Table 1). Each Rep protein was numbered 1 through 33 to assist with identification of the Rep proteins when constructing phylogenetic trees. The Rep proteins were submitted to the web server ClustalW2 (http://www.ebi.ac.uk/Tools/ clustalw2/index.html) (Larkin *et al.*, 2007) and placed into a global multiple sequence alignment (MSA) using the default ClustalW2 settings. Each MSA was color coded bases on percent identity of the amino acids of the Rep proteins with the consensus of the MSA and was indicated with blue colors. Following completion of the MSA, phylogenetic trees of all *H. pylori* Rep proteins were constructed as cladograms based on average distance using percent identity utilizing Jalview (Clamp *et al.*, 2004, Waterhouse *et al.*, 2009).

As shown previously, several plasmids have unique features of their respective replication proteins that are not found with other *H. pylori* plasmids. First, the *rep* gene sequences for plasmids pMCU1 through pMCU7 were obtained utilizing primers to

obtain the middle regions of the *repA* gene. As a result, the complete sequences of their RepA proteins were not reported. They are missing the N-terminus as well as the C-terminus sequences, but are assumed to replicate by the theta mechanism. Second, plasmid pHPAG1 has two different *repA* genes identified as ORF P001 and ORF P003 that each encodes RepA. Lastly, plasmid pHPS1 encodes both RepA and RepB proteins. The first MSA alignment involved all of the Rep proteins from *H. pylori* and included gaps in the sequences (Fig. 24).

Based on the MSA alignment, the plasmids with RepA which includes pAL202 (#1), pAL226 (#2), pHPM8 (#10), pHPM179 (#11), pHPM180 (#12), pHPM185 (#13), pHPM186 (#14), pMCU1 through pMCU7 (#16-22), pHel1 (#3), pHel4 (#4), pHel5 (#5), pHP69 (#6), pHP666 (#7), pHAG1 (#8 and #9), pHPG27 (#23), pHPS1 (#15) and pHPP12 (#24) share the greatest identity with the consensus sequence and show a strong conserved middle region, while the amino acid identity varied at the N-terminus and C-terminus. Plasmids encoding RepB, including pHPS1 (#27), pHP51 (#25), pKU701 (#28), pHPO100 (#26), and pAL236-11 (#29), appeared to have no significant amino acid identity to the RepA proteins or the consensus sequence. RC plasmids pAL236-2 (#32), pAL236-5 (#33), pHP489 (#30), and pHPK255 (#31) also appeared to have no significant amino acid identity to RepA, RepB, or the consensus sequence. However, RepB plasmids did have more amino acid identity with the consensus sequence than did Rep.

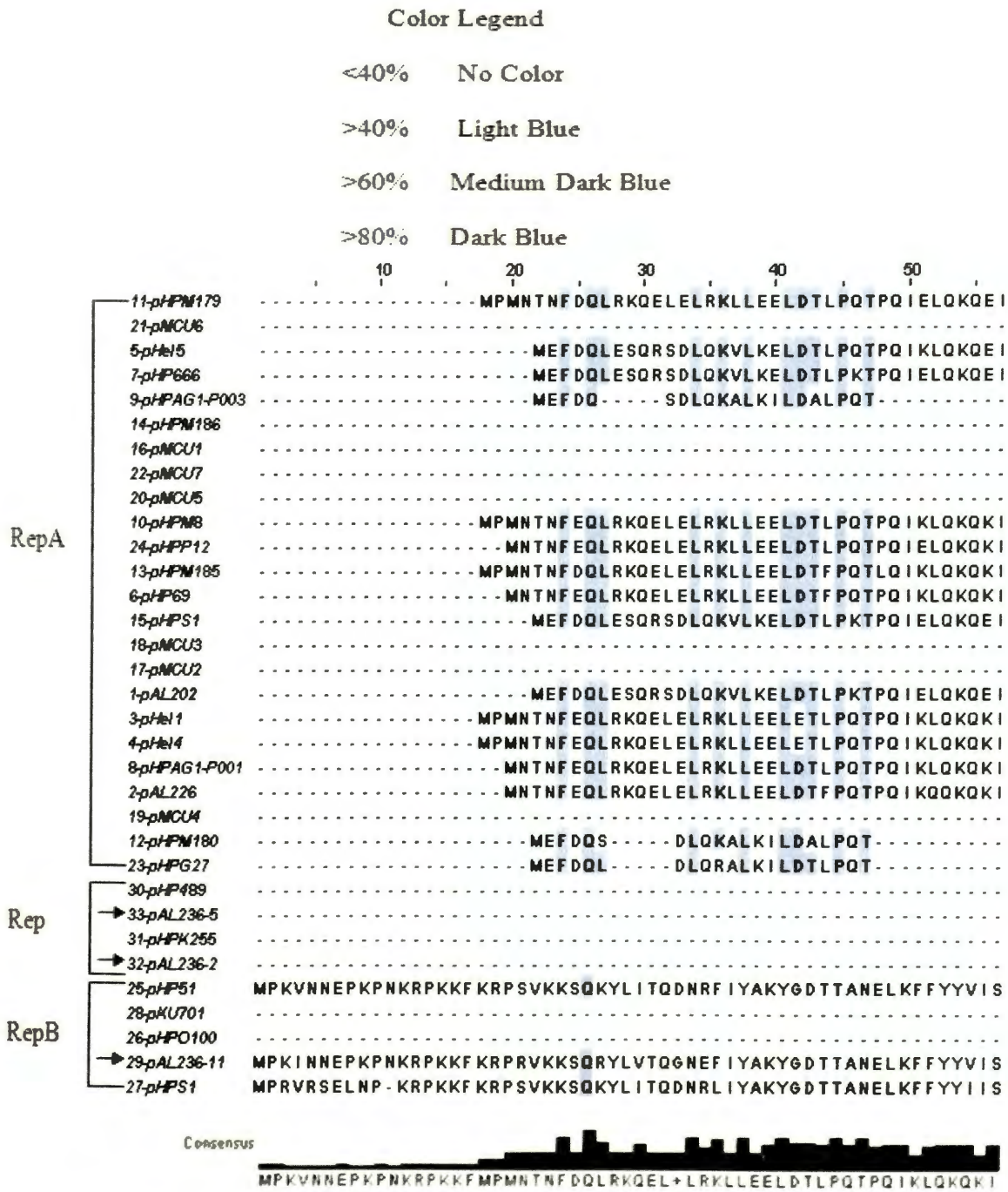

Figure 24. ClustalW2 multiple sequence alignment of all Rep Proteins from *H. pylori* Plasmids with gaps. Percent identity of amino acids with consensus are shaded. RepA, RepB, and Rep are indicated by boxes. Plasmids pAL236-2, pAL236-5, and pAL236-11 are indicated with arrows.

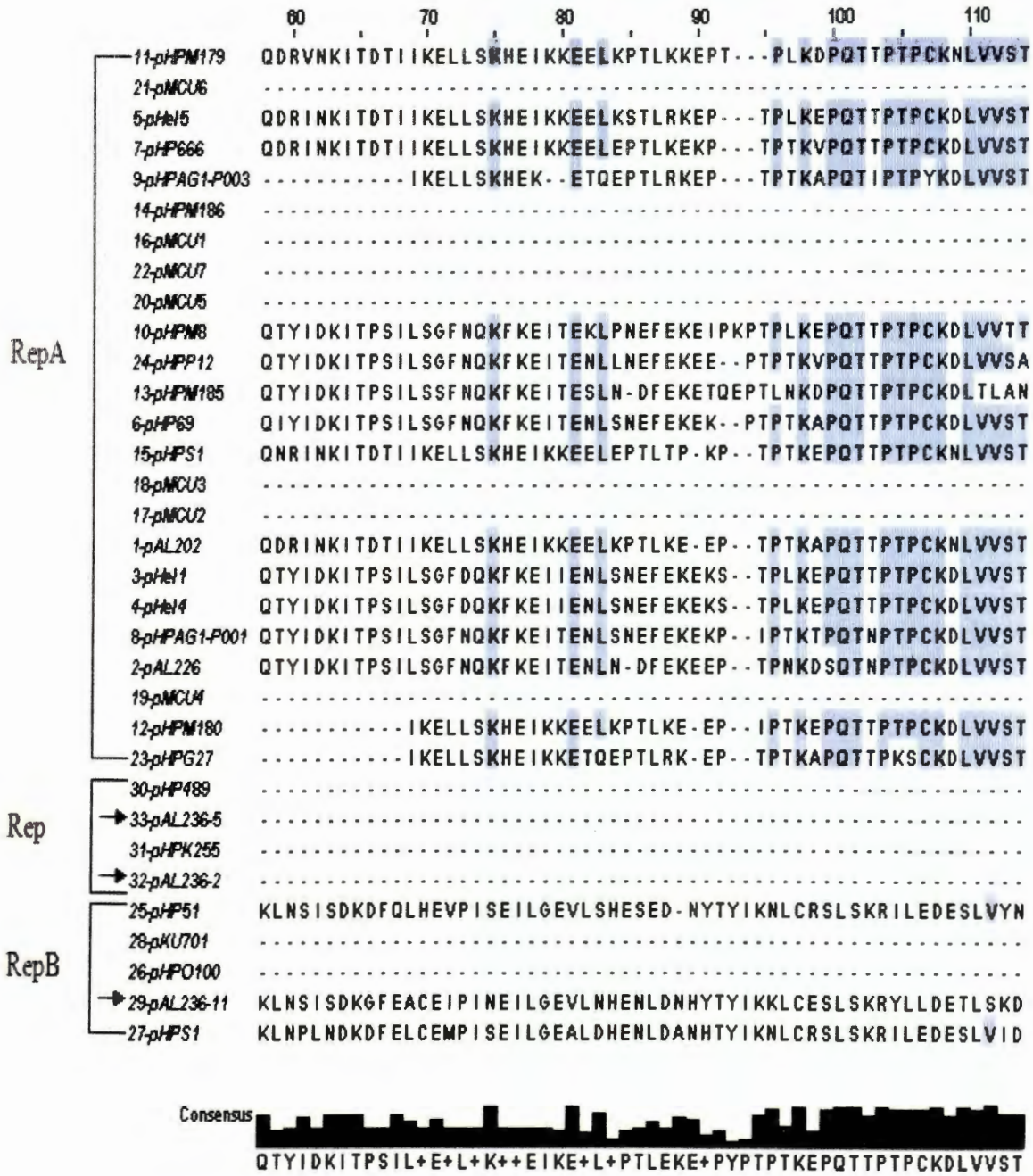

Figure 24 continued.

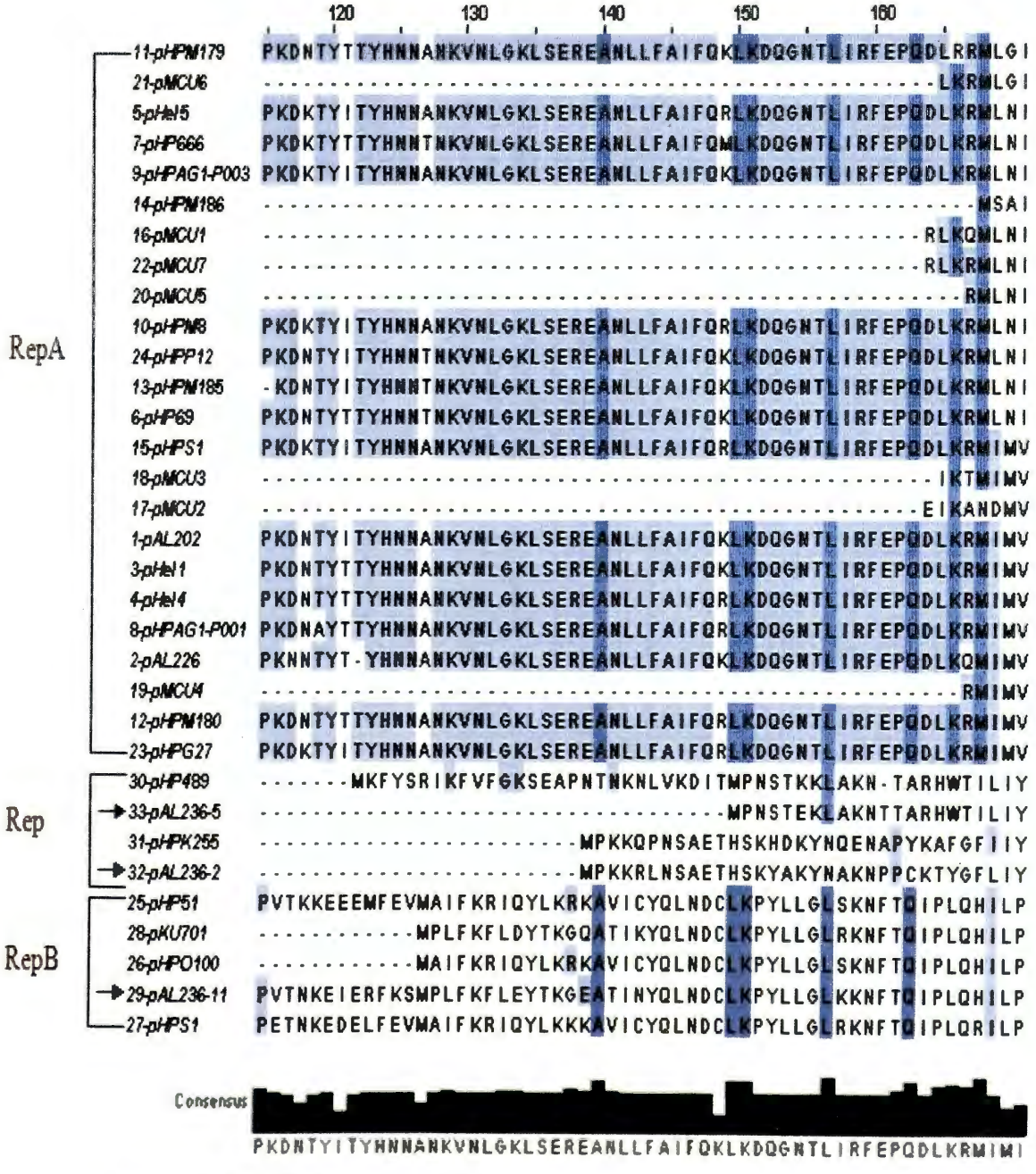

Figure 24 continued.

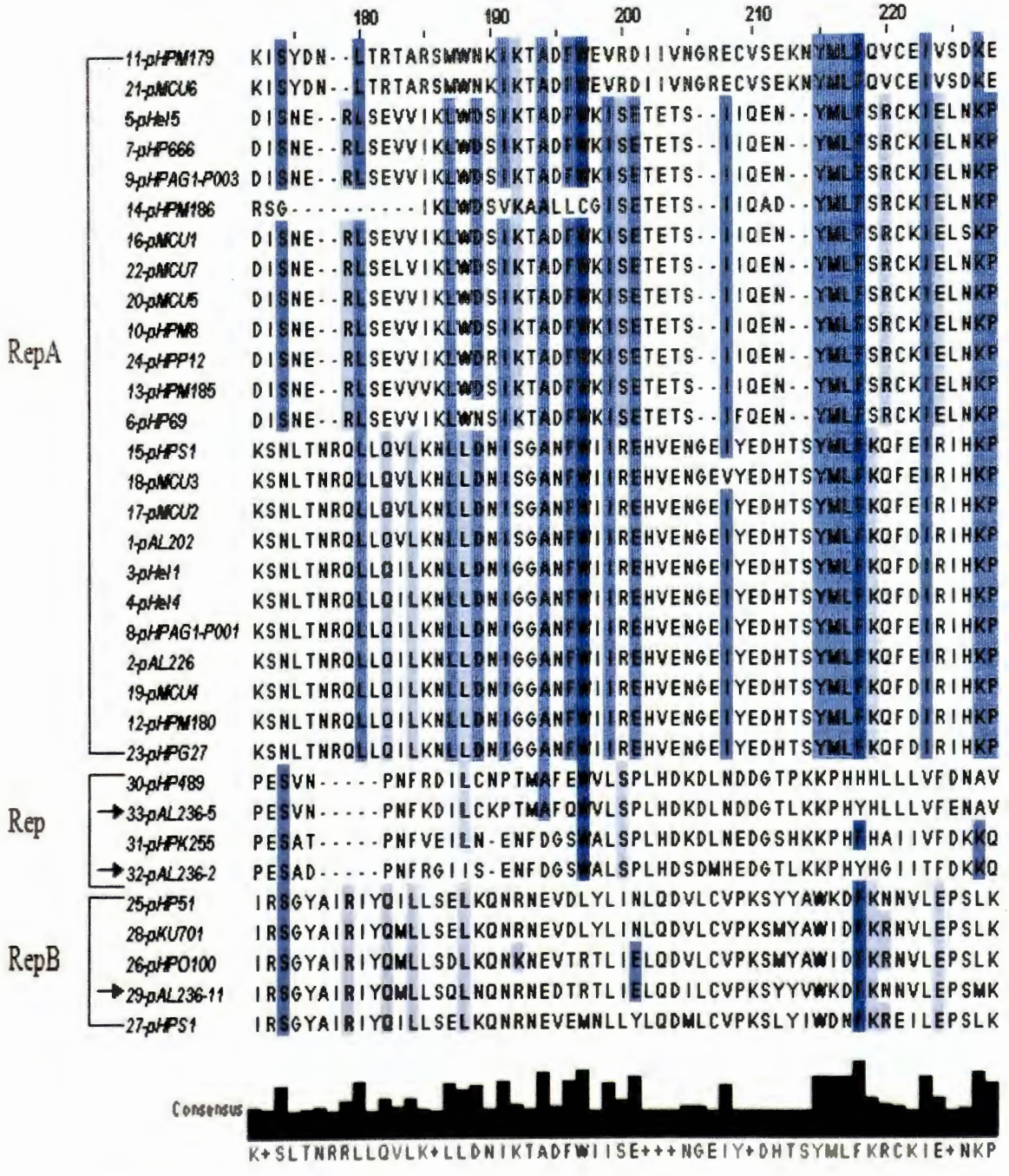

Figure 24 continued.

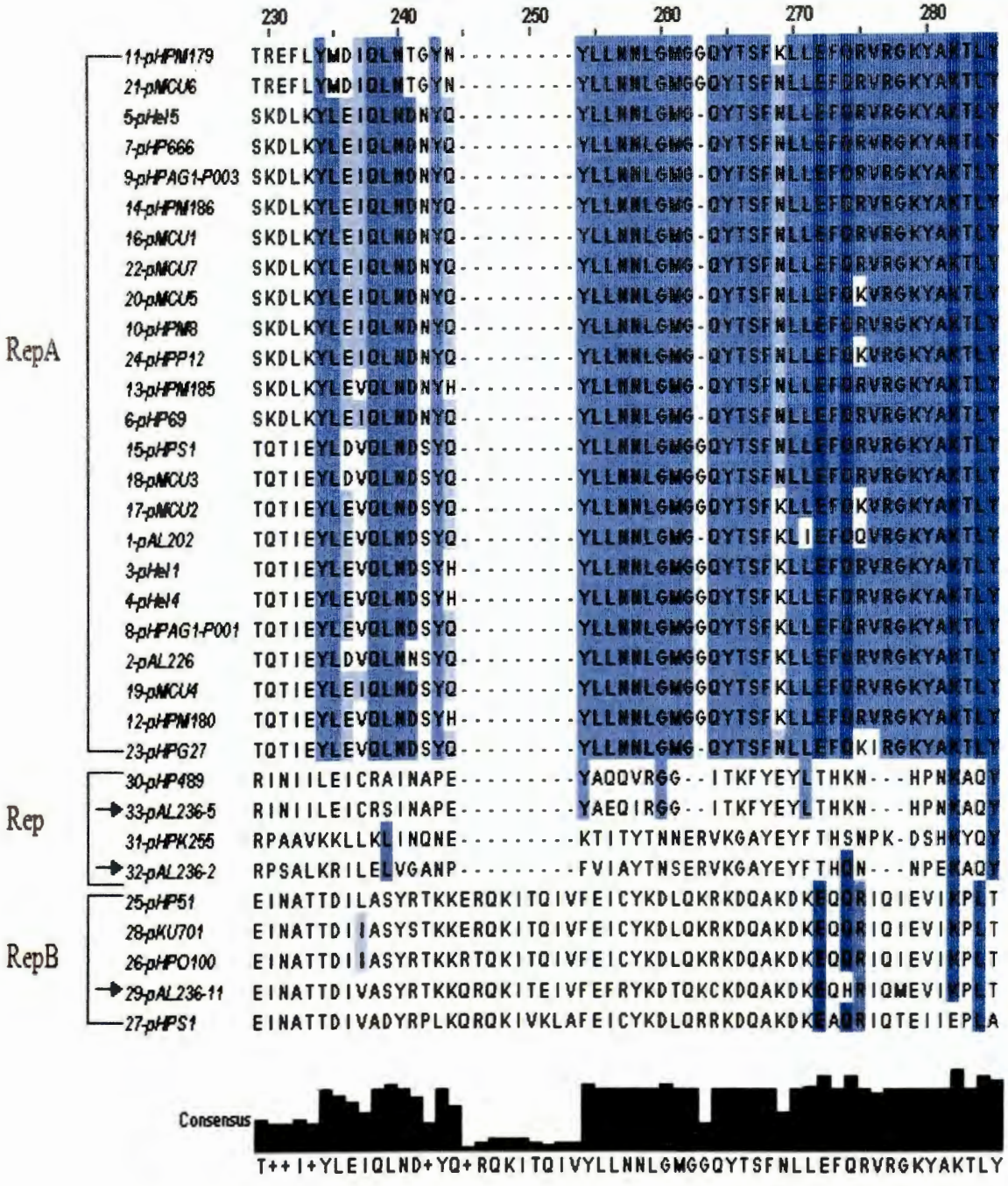

Figure 24 continued.

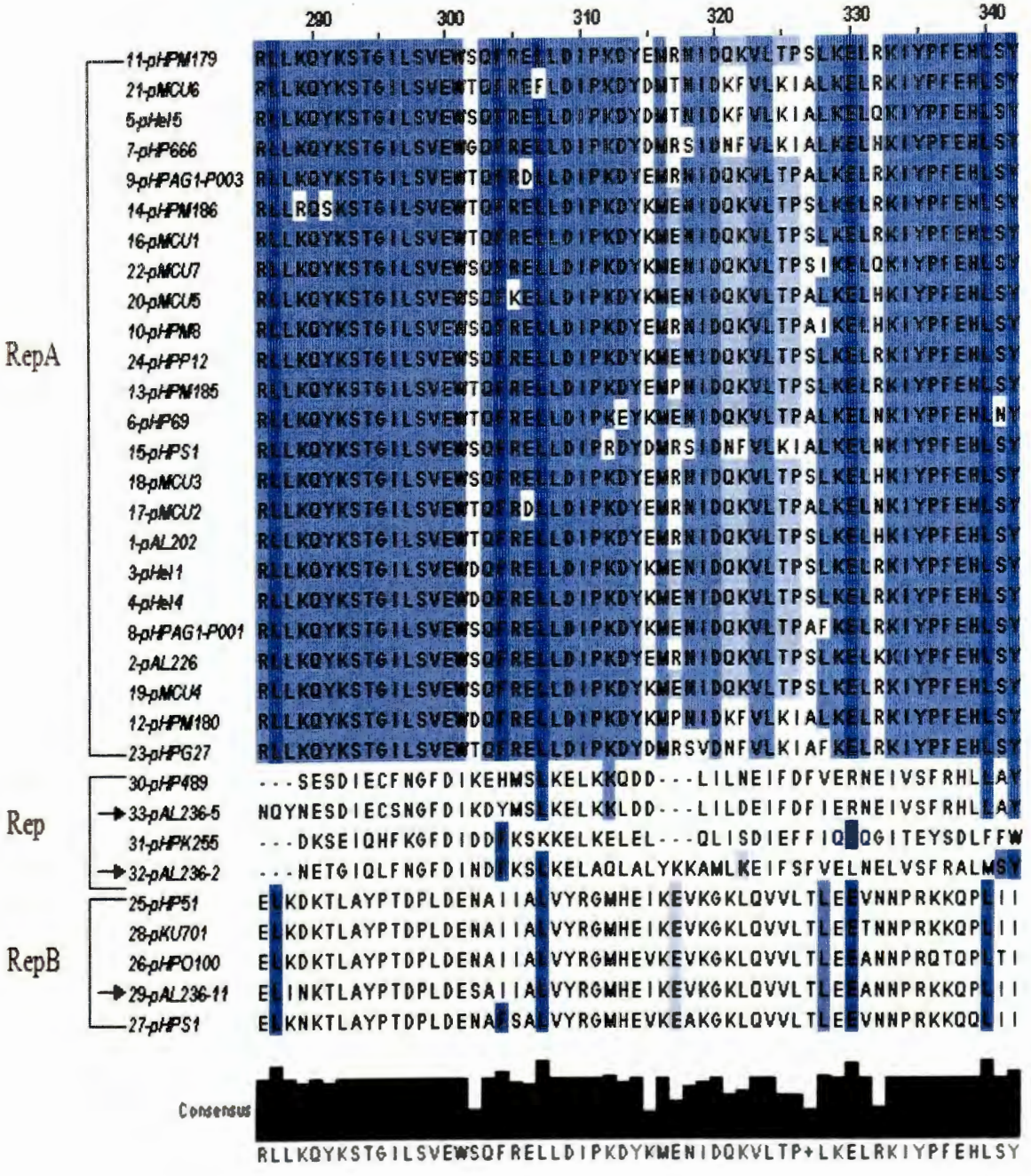

Figure 24 continued.

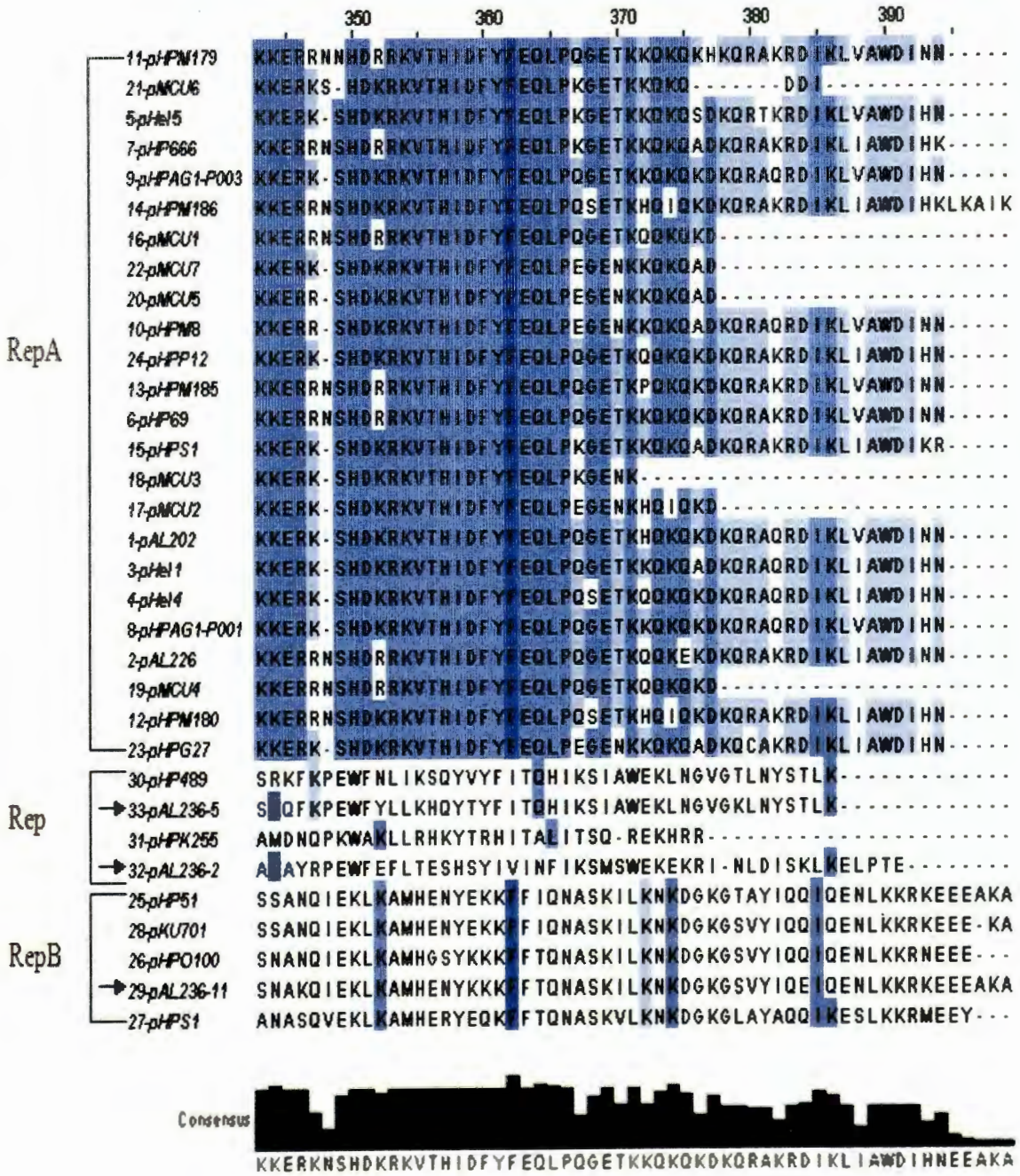

Figure 24 continued.

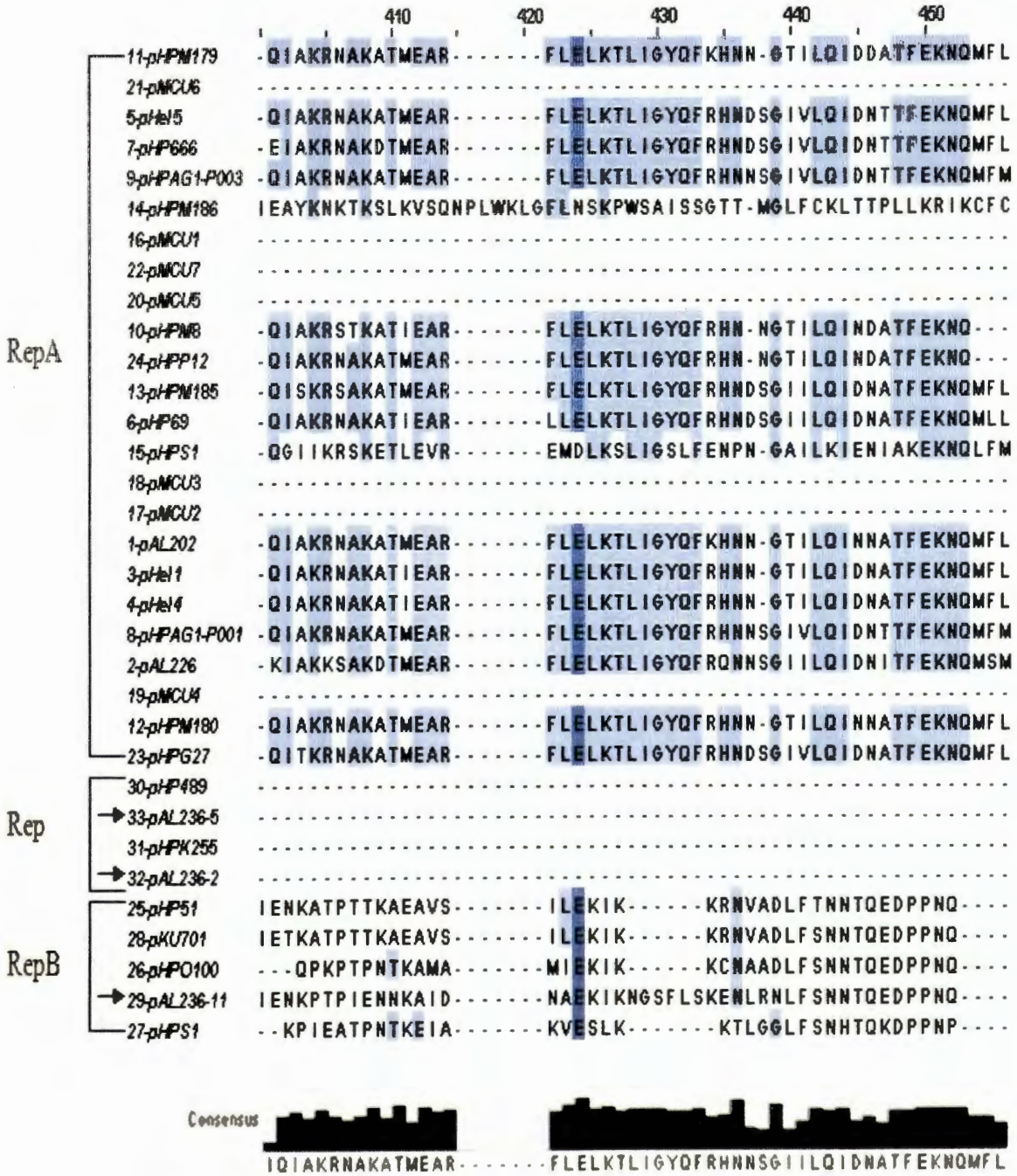

Figure 24 continued.

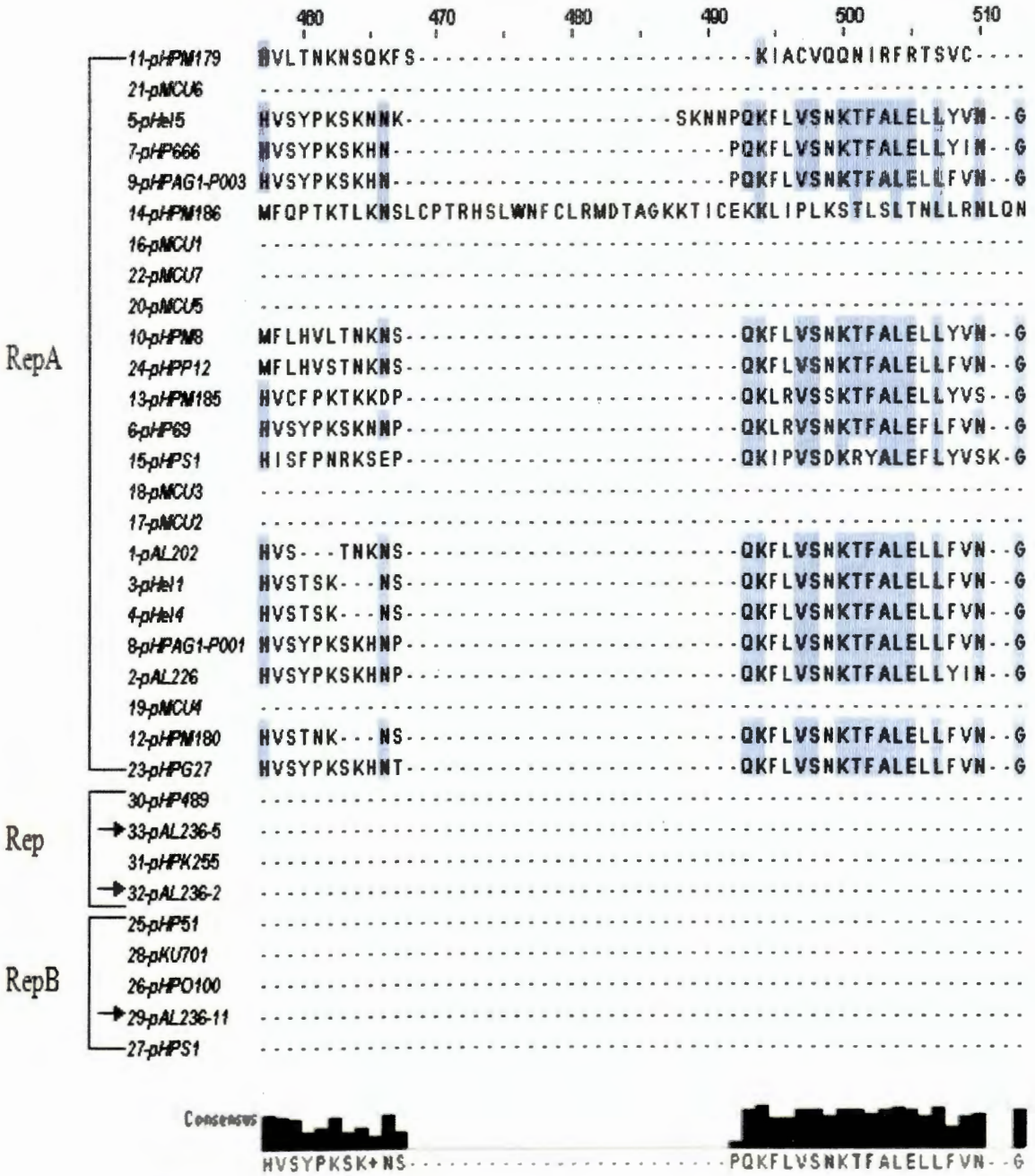

Figure 24 continued.

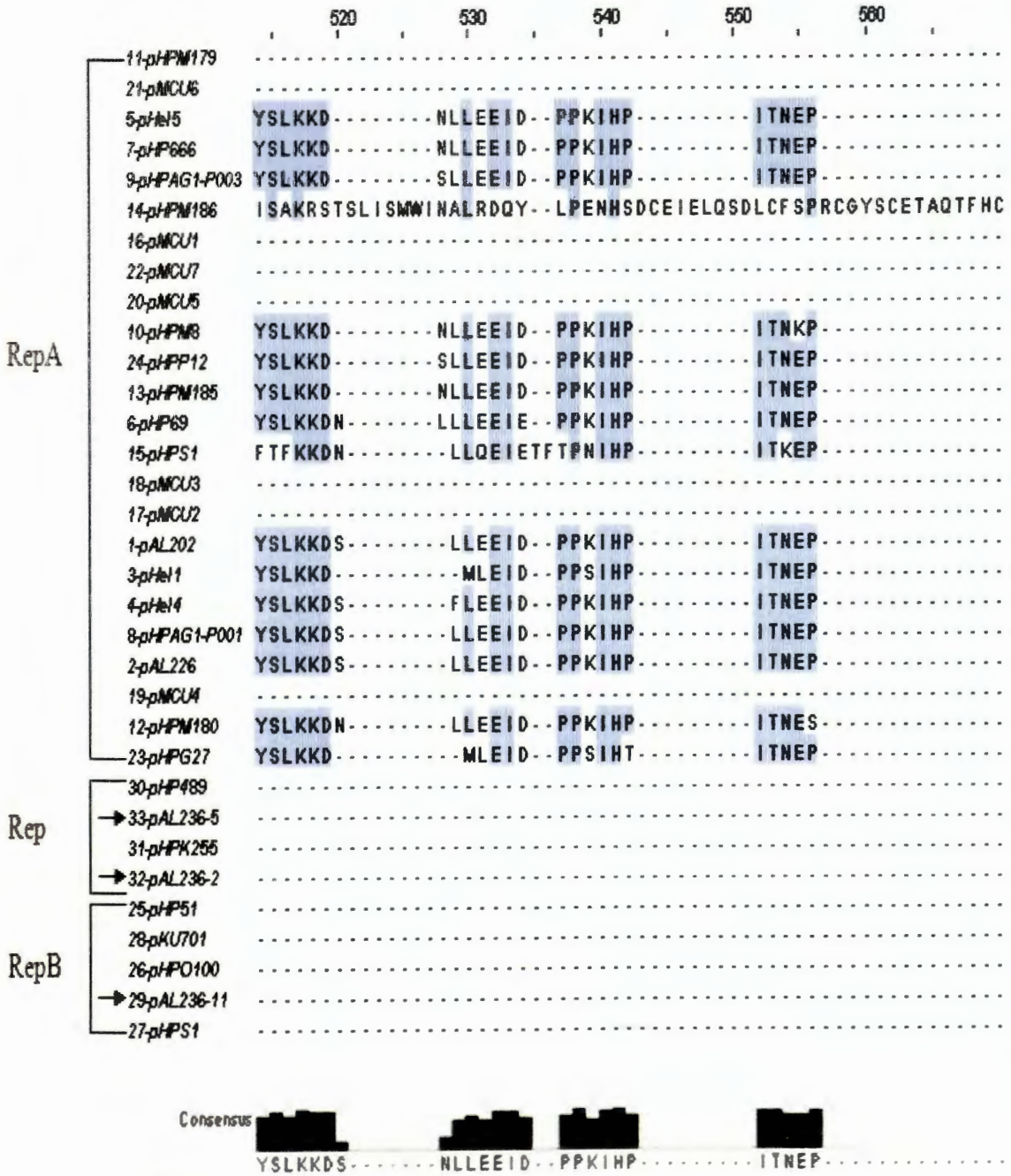

Figure 24 continued.

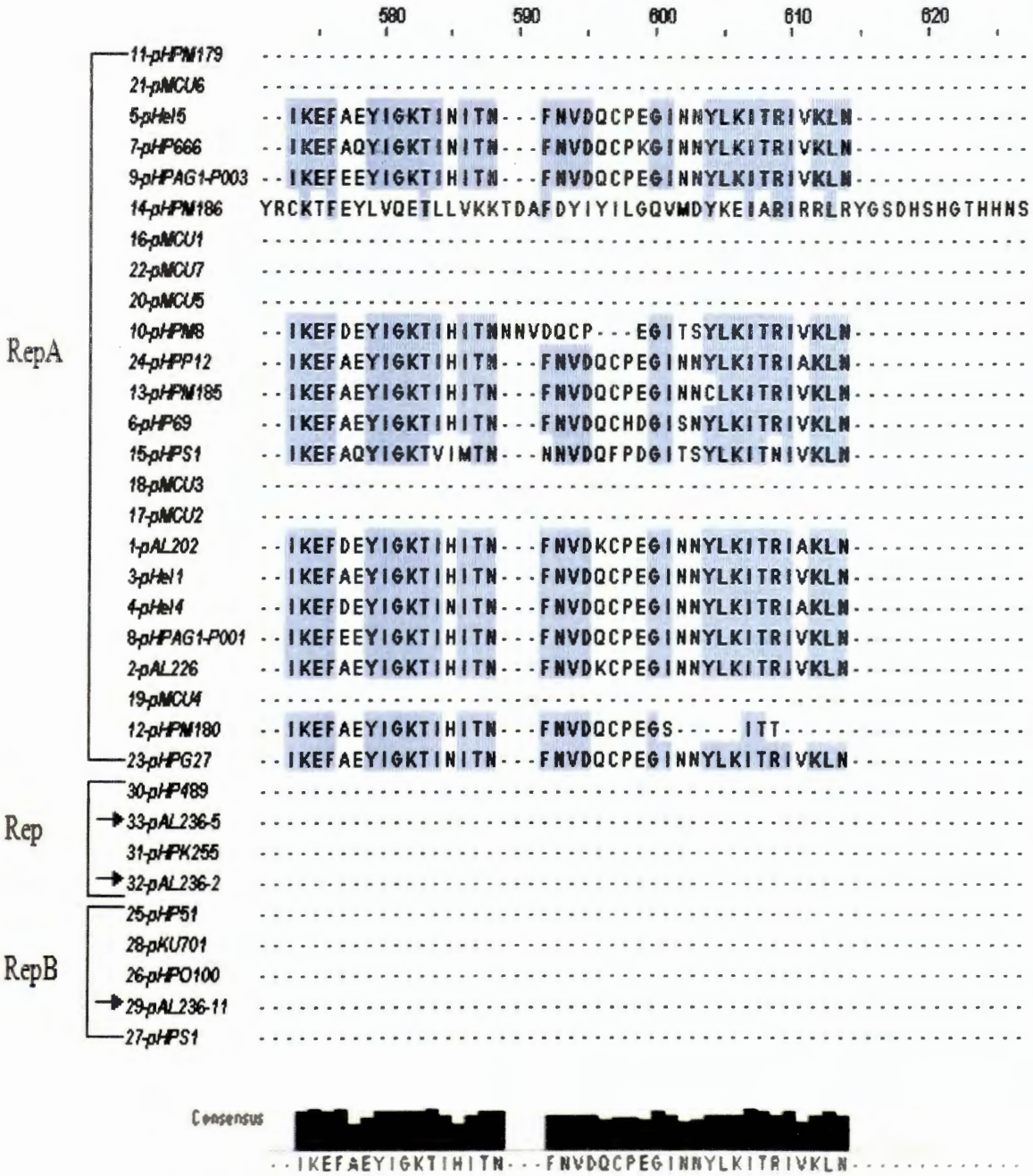

Figure 24 continued.

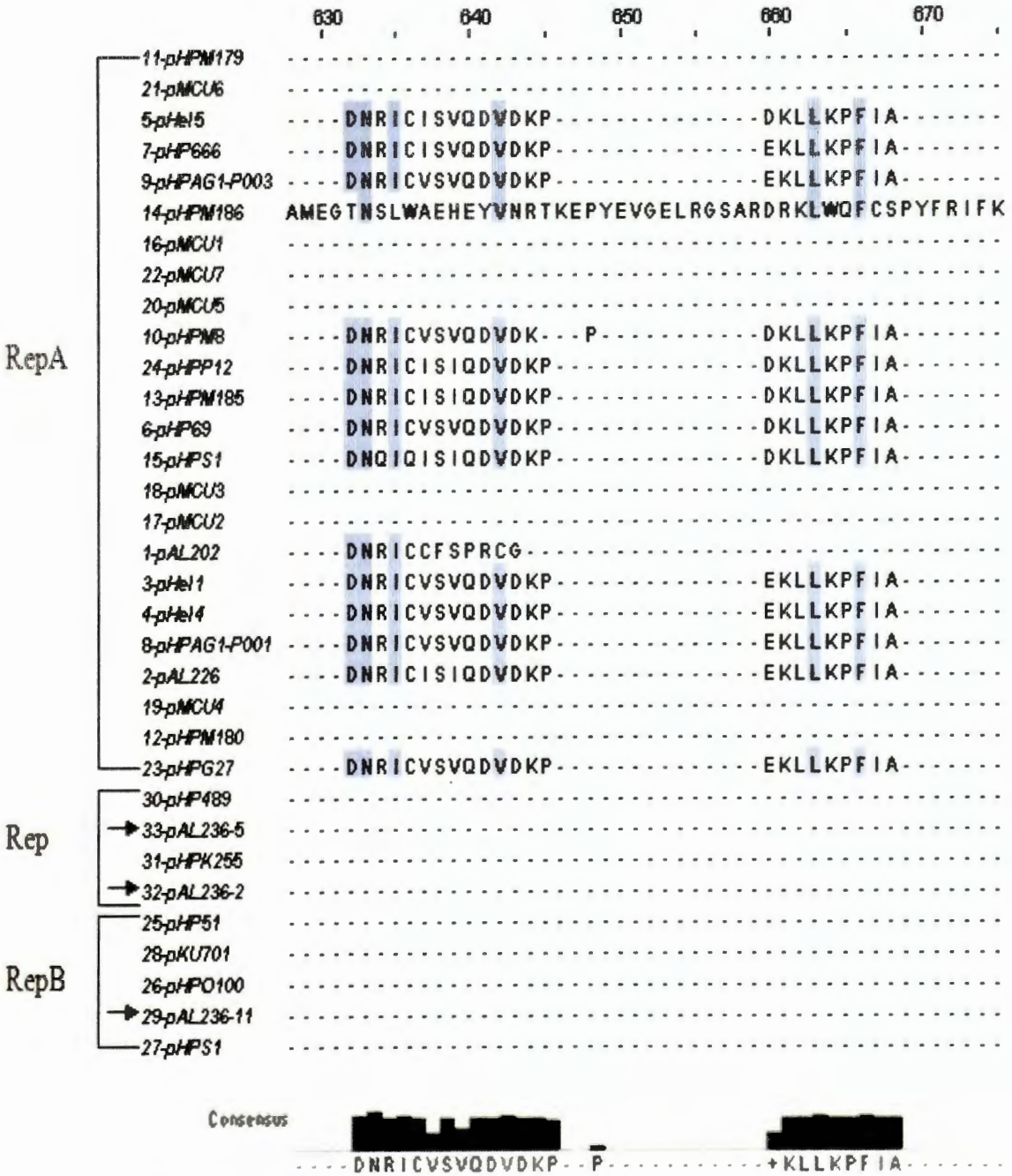

Figure 24 continued.

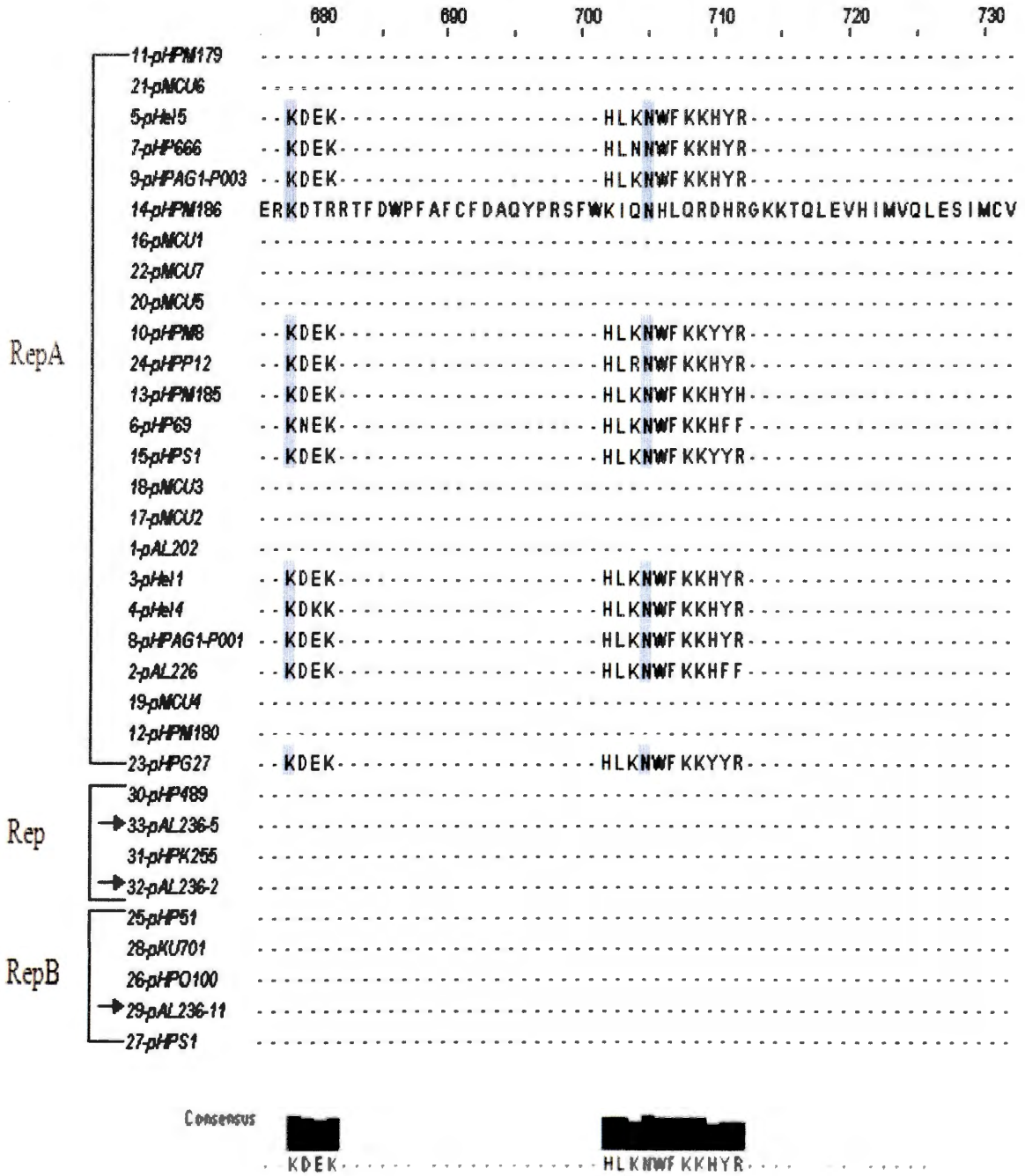

Figure 24 continued.

A cladogram was constructed based on the MSA for all replication proteins of *H. pylori* including gaps (Fig. 25). Plasmids with RepA were organized into three main groups or clades. The first clade consisted of only pHPM179 and pMCU6. The second clade included pHel5, pHP666, pHPAG1 (P003), pMCU1, pHPP12, pMCU7, pMCU5, pHPM8, pHP69, and pHPM185. Within this clade it appears that RepA plasmids were divided into smaller sub-clades.

The first sub-clade consisted of pHel5, pHP666, and pHPAG1 (P003). A second sub-clade contained pMCU1, pHPP12, pMCU7, pMCU5, pHPM8, pHP69, and pHPM185. The third and final clade consisted of pMCU3, pAL202, pMCU2, pHPM180, pHel1, pHel4, pMCU4, pHPAG1 (P001), pAL226, and pHPG27. However, the third clade also revealed that these plasmids belong to several sub-clades. The first sub-clade consisted of pMCU3, pAL202, pMCU2, and pHPM180. Plasmids of the second sub-clade included pHel1, pHel4, pMCU4, pHPAG1 (P001), pAL226 and pHPG27. Interestingly two plasmids pHPS1 and pHPM186 do not fall into any of the three main clades and appeared to be outliers of RepA.

Plasmids within the RepB group, like RepA, replicate via the theta mechanism. The RepB plasmids included pHP51, pKU701, pHPO100, pAL236-11, and pHPS1. There appeared to be one clade which included all of the RepB plasmids. One sub-clade was found with RepB that included pHP51, pKU701, pHPO100, and pAL236-11. However, pHPS1 was an outlier of the sub-clade for RepB proteins.

Besides RepA and RepB another type of replication protein can be found in *H. pylori* involved with RC replication and has been termed Rep. The Rep of RC plasmids

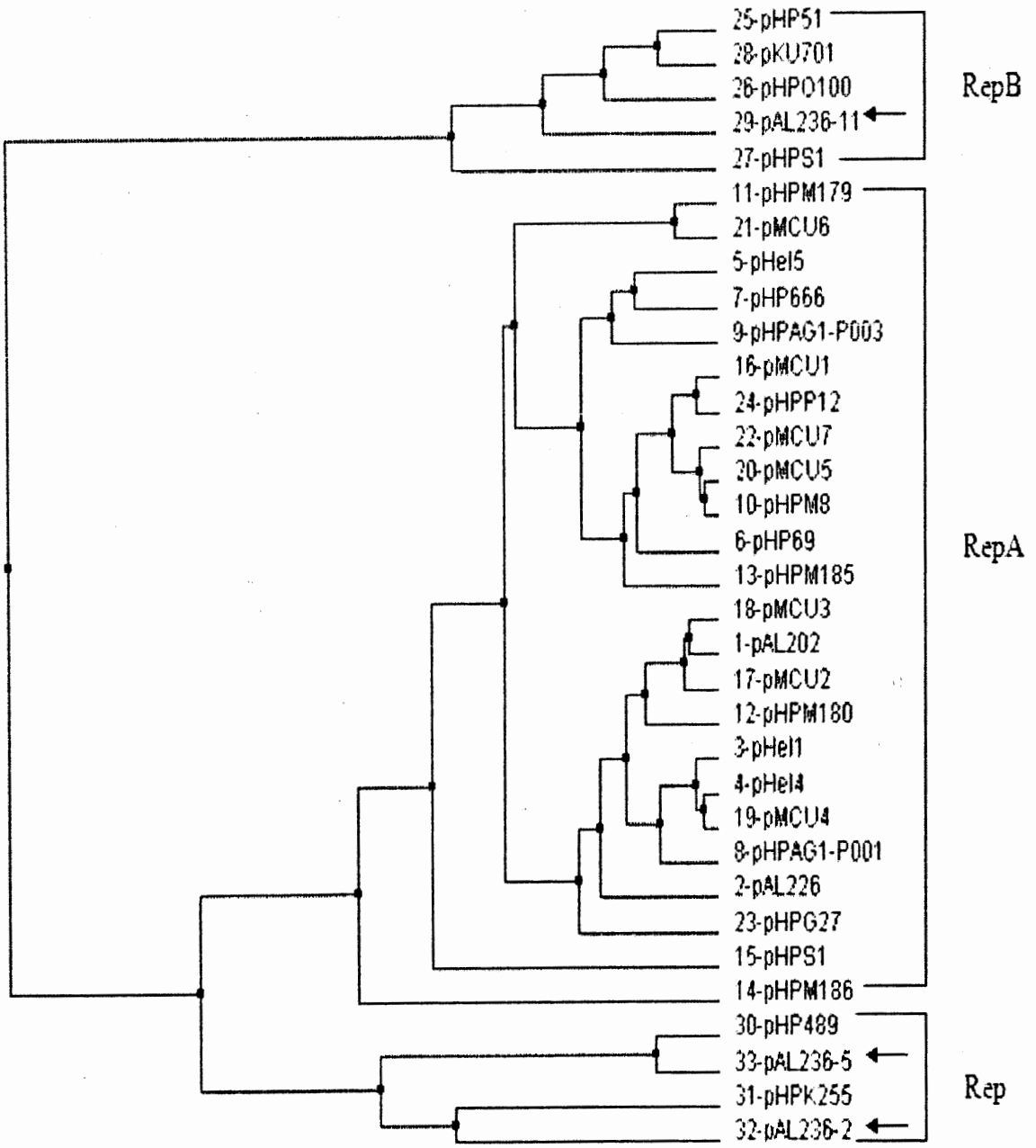

Figure 25. Average distance tree based on percent identity of the multiple sequence alignment of all Rep Proteins from *H. pylori* with gaps. RepA, RepB, and Rep plasmids are indicated by a box. Plasmids pAL236-2, pAL236-5, and pAL236-11 are indicated with arrows.

included pHP489, pHPK255, pAL236-2, and pAL236-5. These plasmids formed one clade that appeared to have two sub-clades. The first sub-clade contained pHP489 and pAL236-5 and the second sub-clade contained pHPK255 and pAL236-2.

PHYLOGENETIC ANALYSIS OF REPLICATION PROTEINS FROM *H. pylori* WITHOUT pMCU1-pMCU7 (#16-22) WITH GAPS INCLUDING pAL236-2, pAL236-5, AND pAL236-11

Since pMCU1 through pMCU7 (#16-22) lacked N-terminus and C-terminus sequences, a second MSA was performed without these plasmids (Fig. 26). MSA results indicated that plasmids with RepA still had a well conserved middle region with the greatest identity to the consensus sequence. However, without pMCU1-pMCU7, there was more identity with the consensus sequence at the N-terminus and C-terminus of the RepA proteins than in the previous MSA. There were still areas of the MSA lacking identity with the consensus sequence. Interestingly, pHPM186 contained the conserved middle region and appeared to have the least similarity with both RepA and consensus sequences of both MSA. Plasmids with RepB and Rep still appeared to have no significant identity with the consensus sequence of the MSA. However, like the previous MSA, RepB did share more amino acid identity with the consensus than did Rep.

A cladogram was constructed based on the MSA lacking pMCU1 through pMCU7 sequences (Fig. 27). Plasmids with RepB and Rep in the cladogram were arranged with each other the exact same as in the previous cladogram. For plasmids with RepA the cladogram showed two main clades, unlike the previous cladogram with three clades. The first clade consisted of pAL202, pHPM180, pHPG27, pHel1, pHel4,

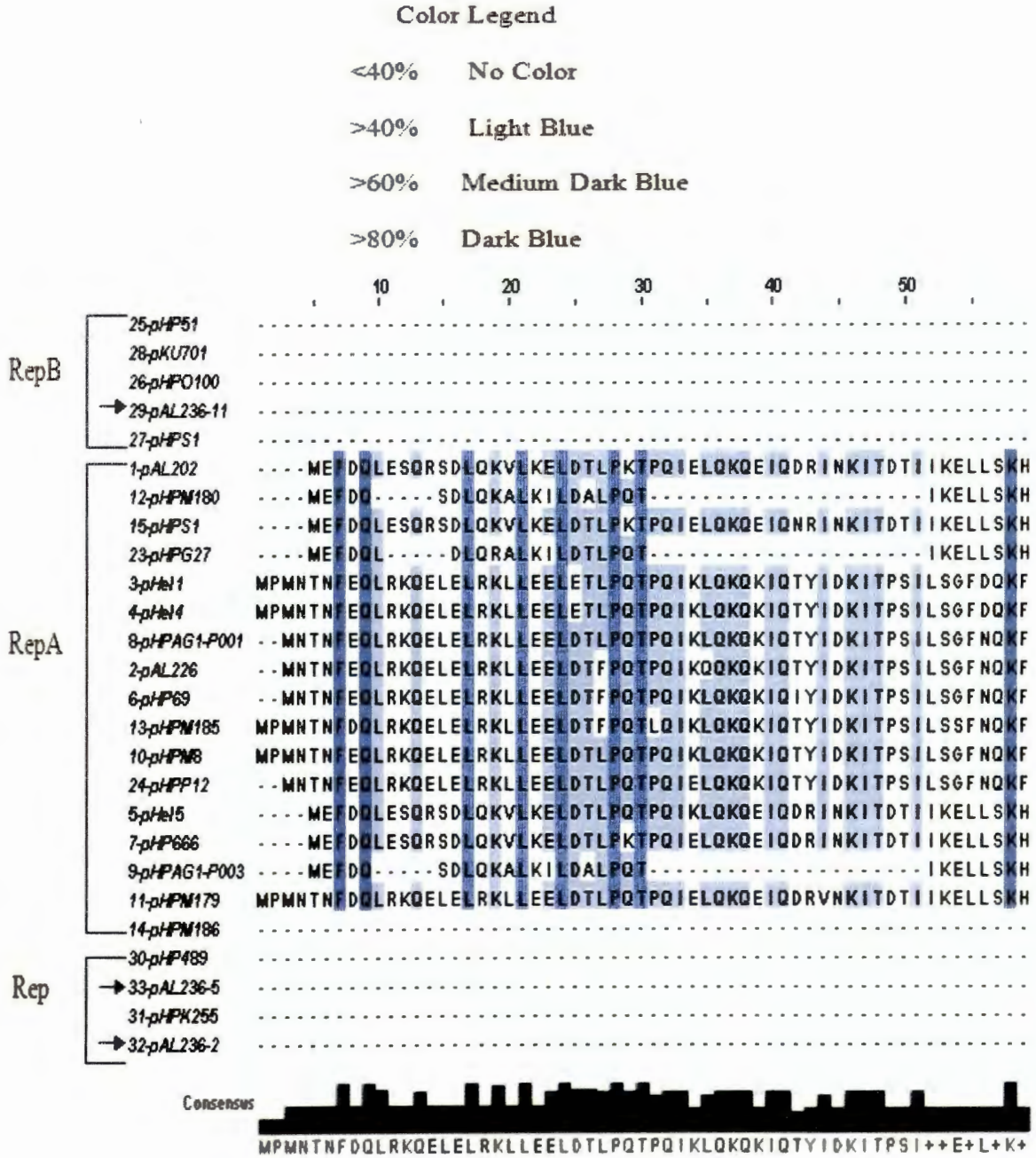

Figure 26. ClustalW2 multiple sequence alignment of Rep Proteins from *H. pylori* Plasmids without pMCU1-7 (#16-22) with gaps. Percent identity of amino acids with consensus are shaded. RepA, RepB, and Rep are indicated by boxes. Plasmids pAL236-2, pAL236-5, and pAL236-11 are indicated with arrows.

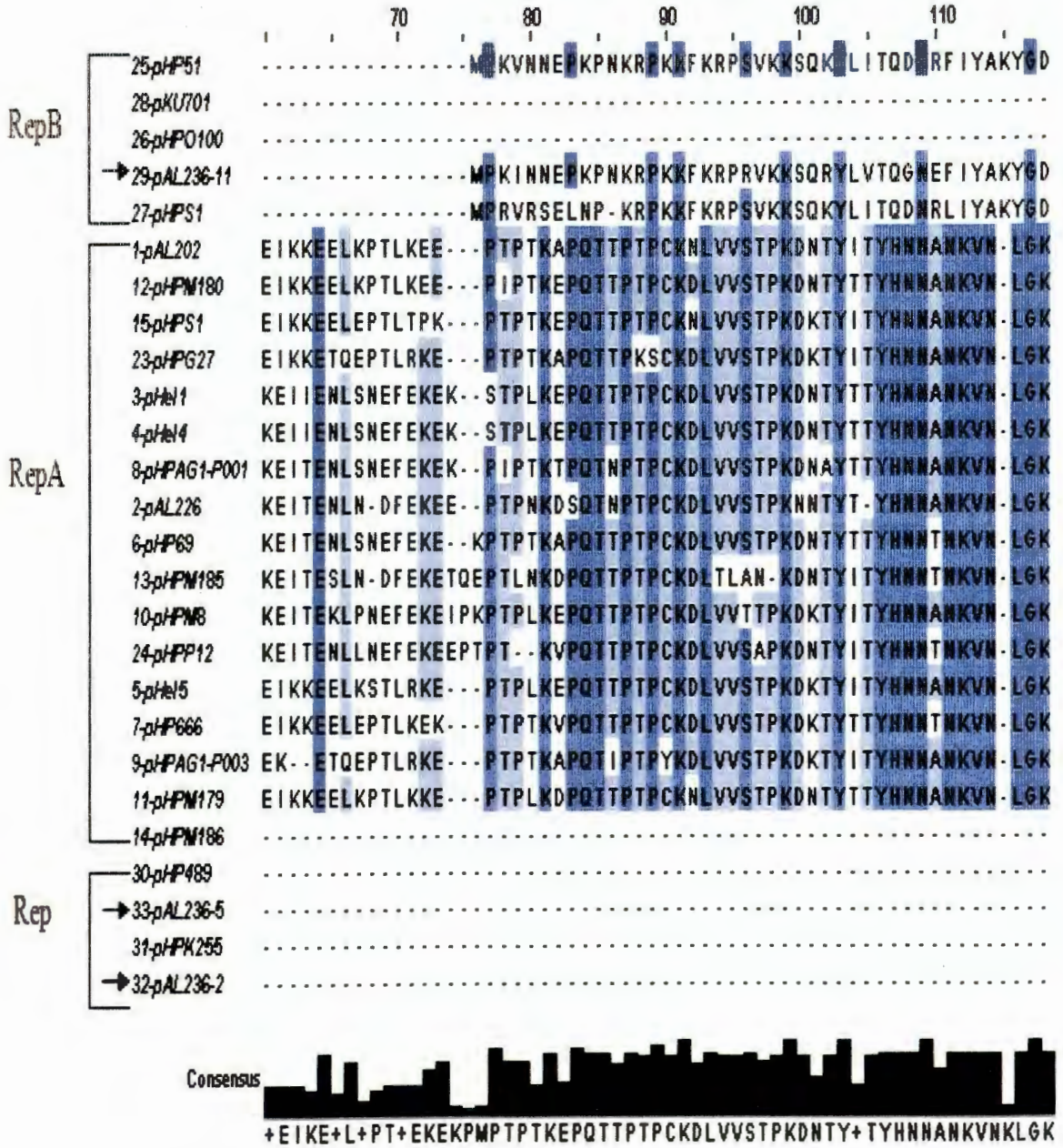

Figure 26 continued.

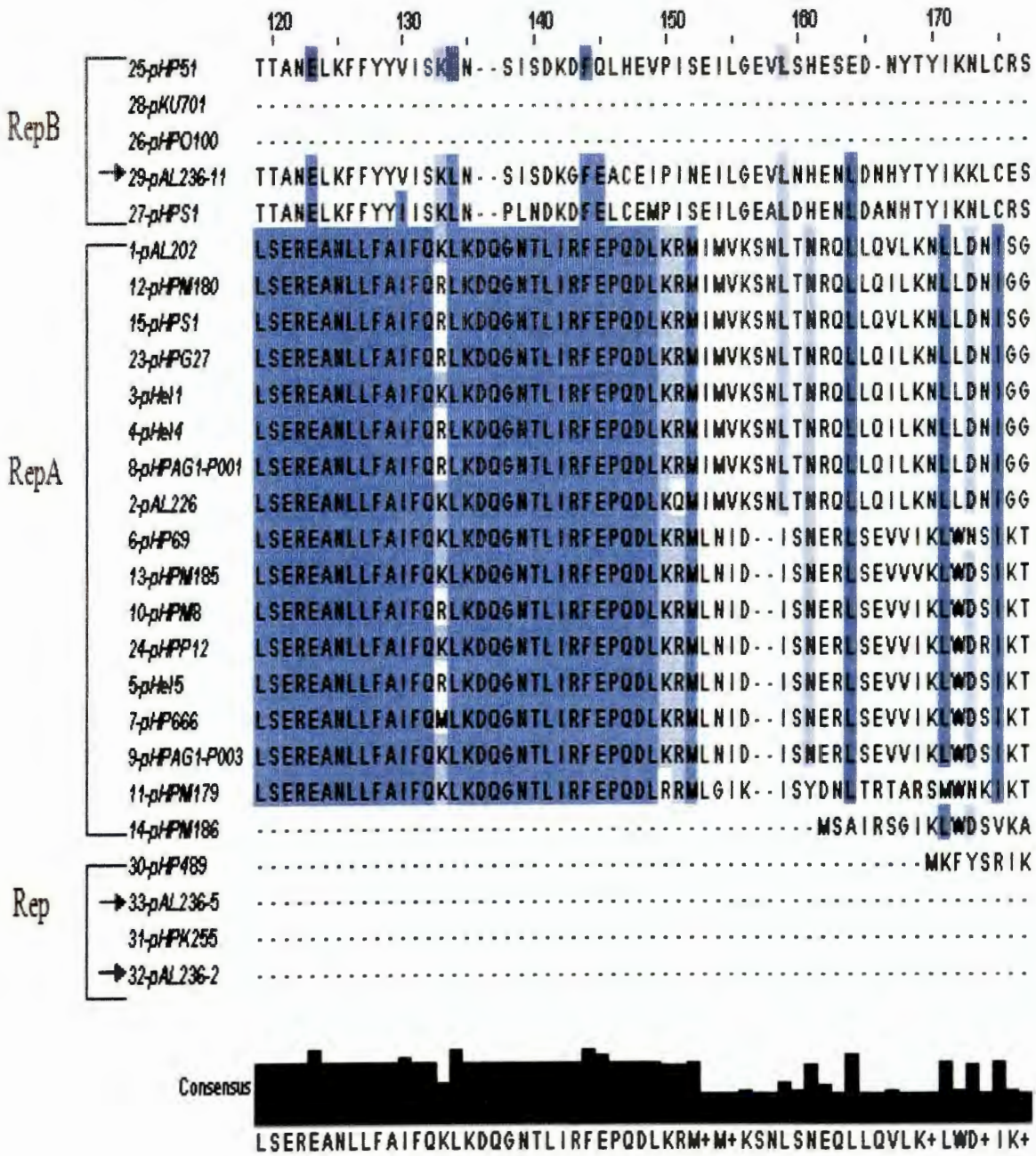

Figure 26 continued.

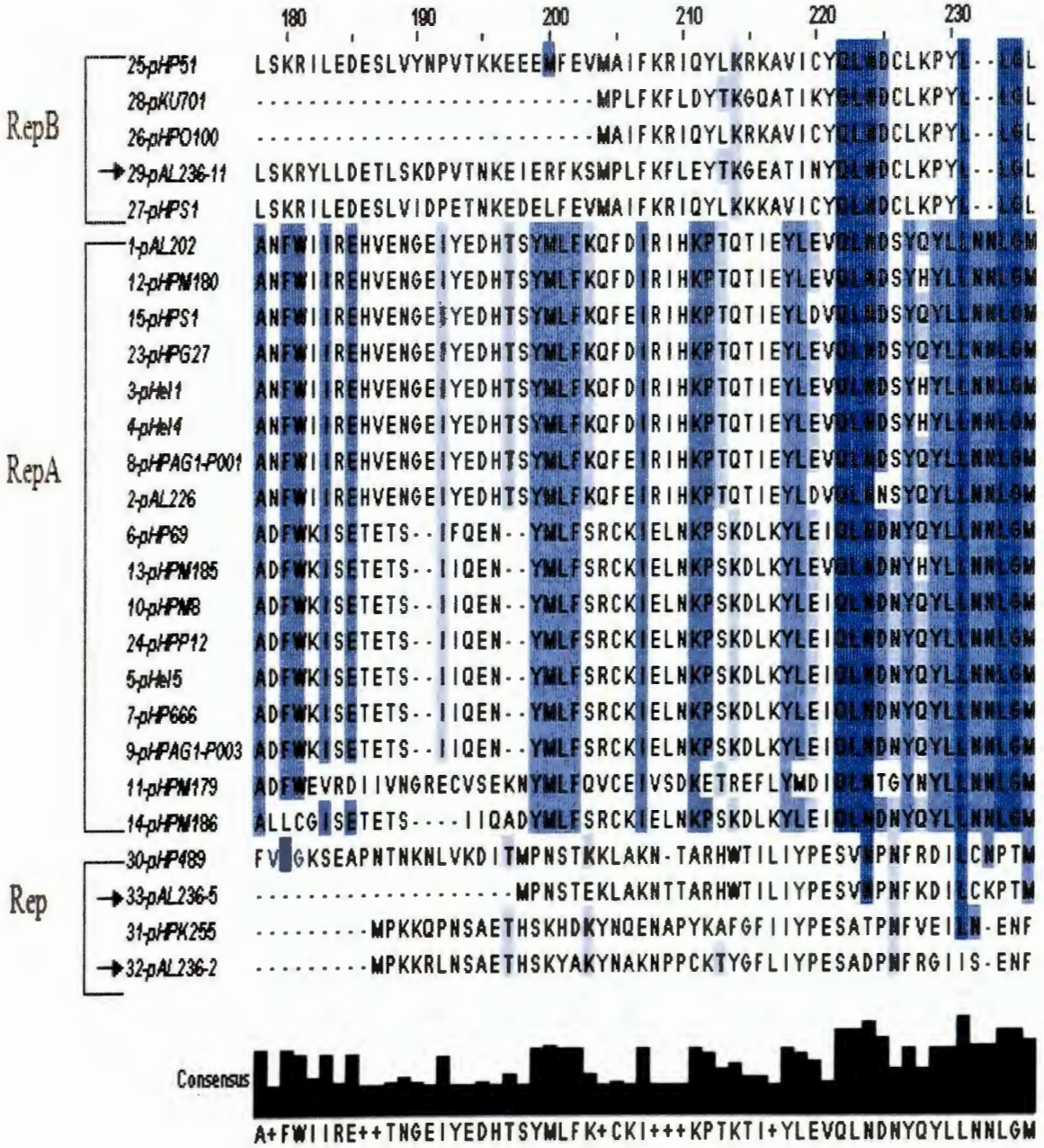

Figure 26 continued.

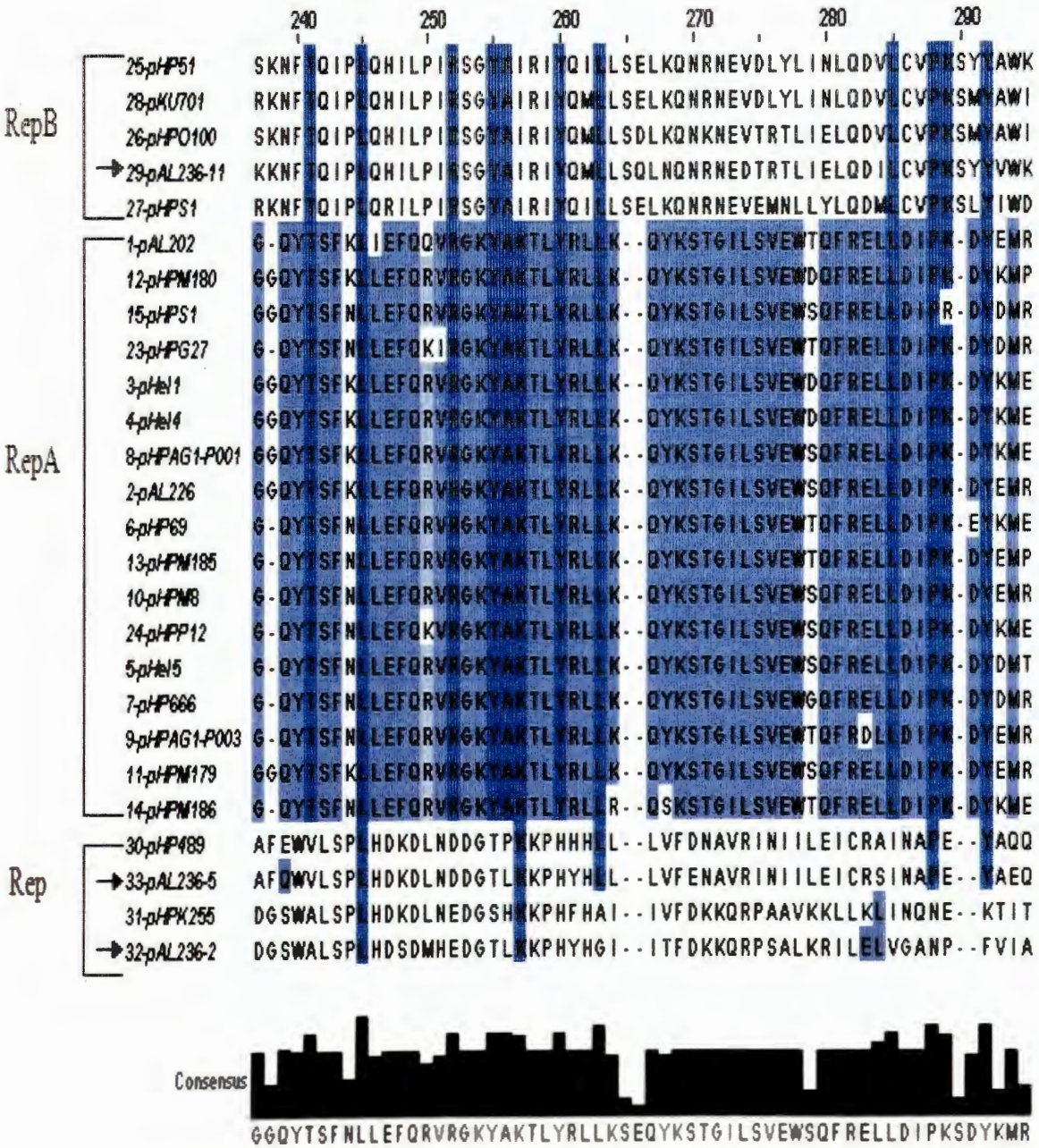

Figure 26 continued.

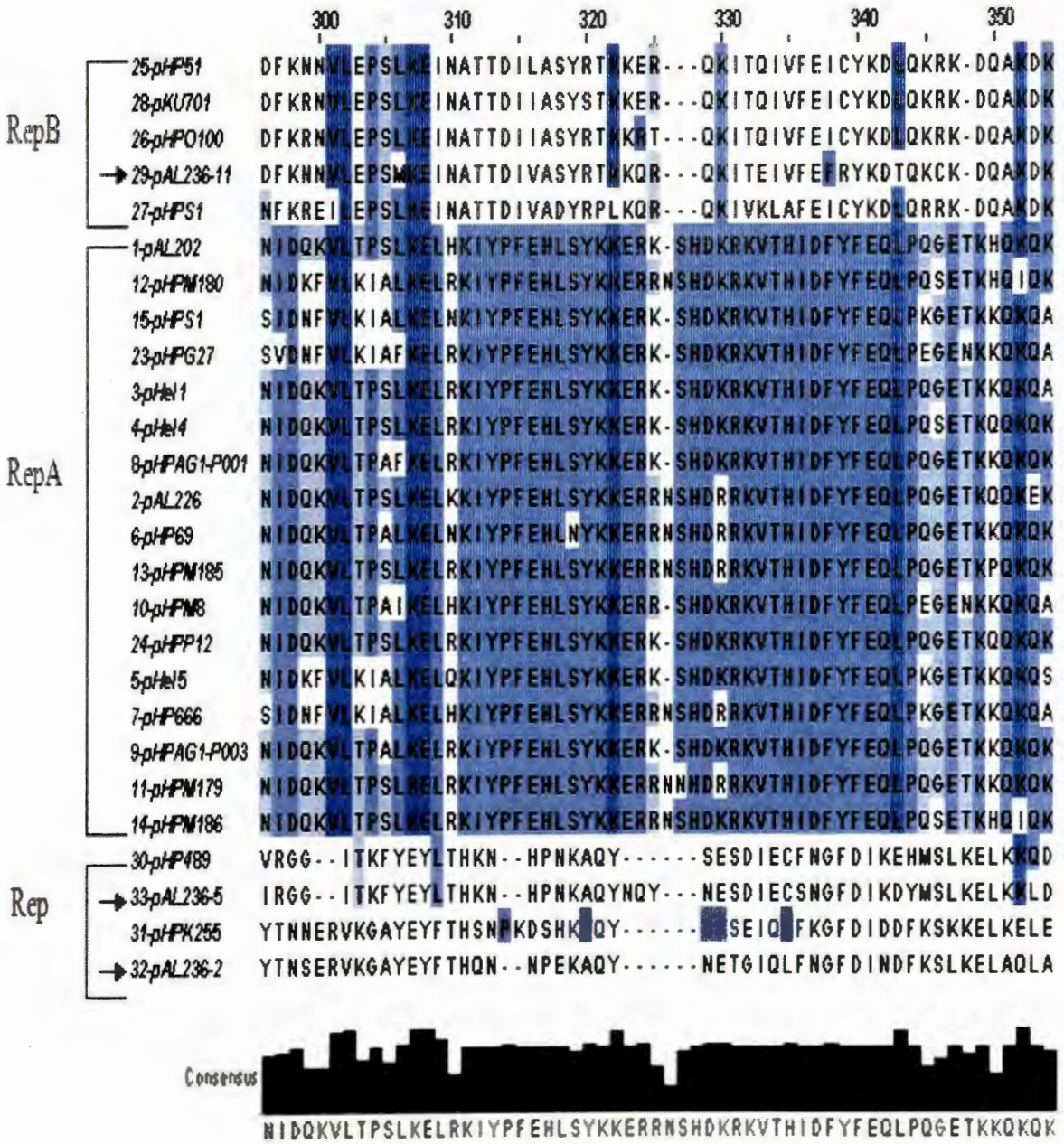

Figure 26 continued.

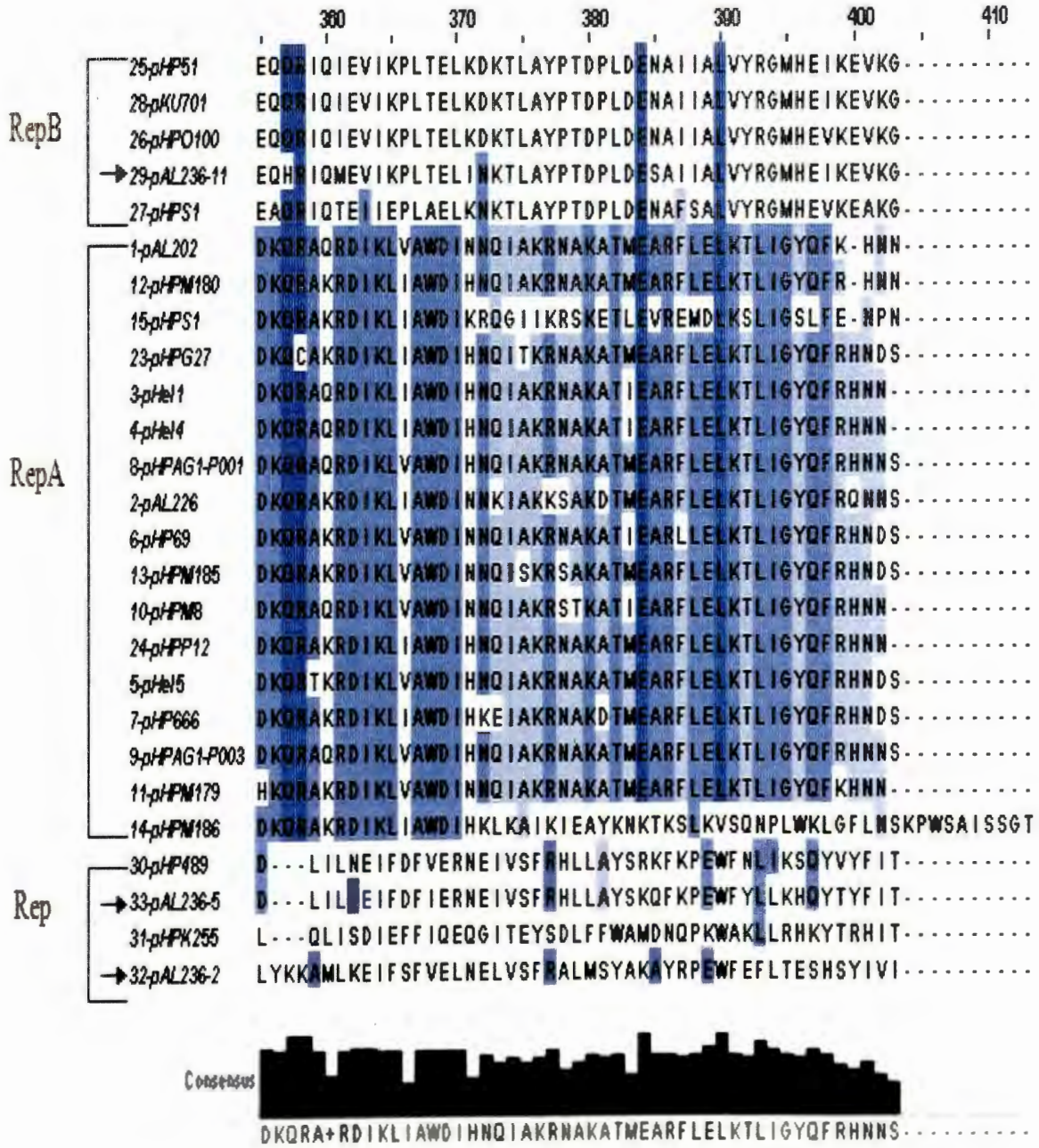

Figure 26 continued.

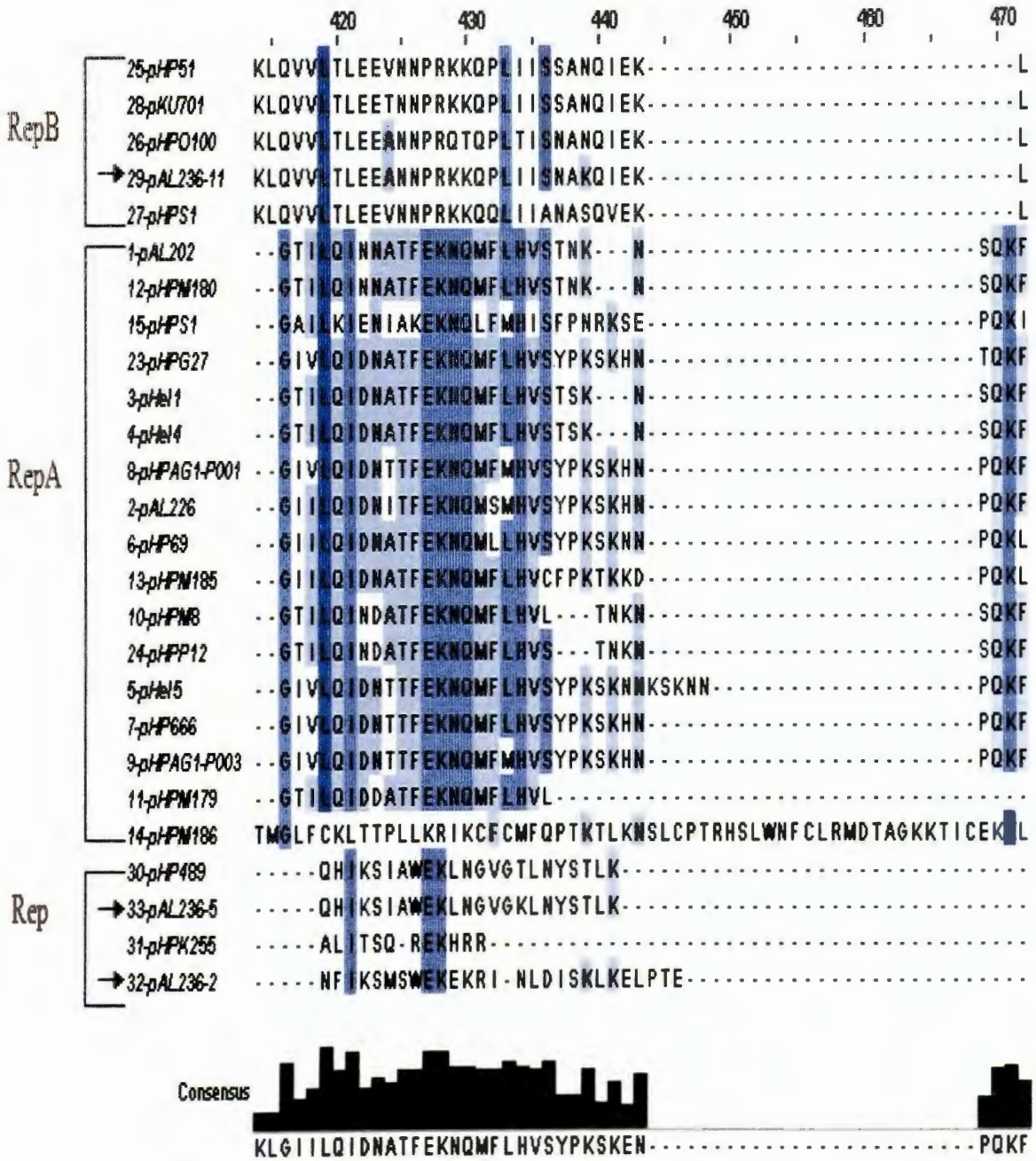

Figure 26 continued.

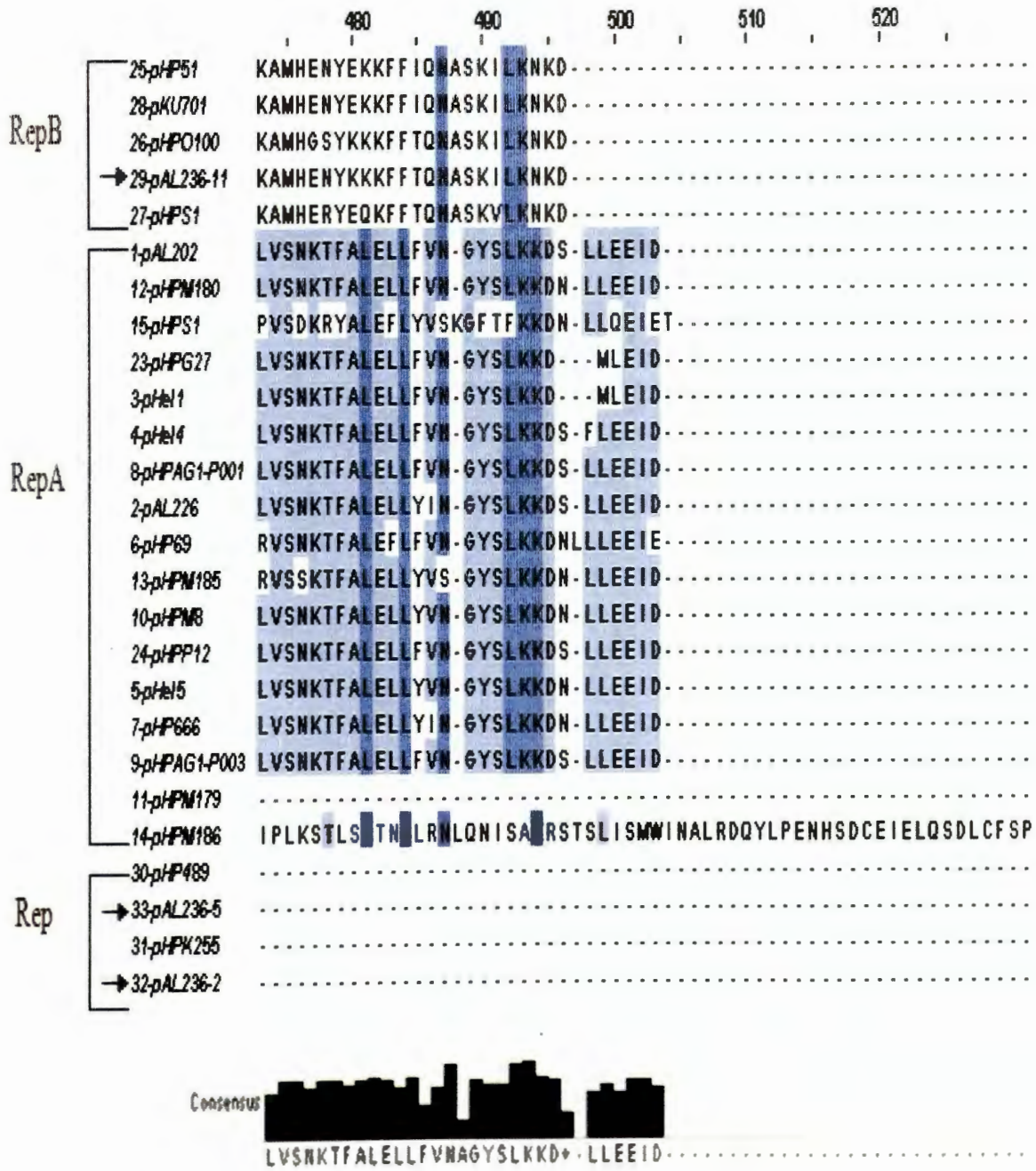

Figure 26 continued.

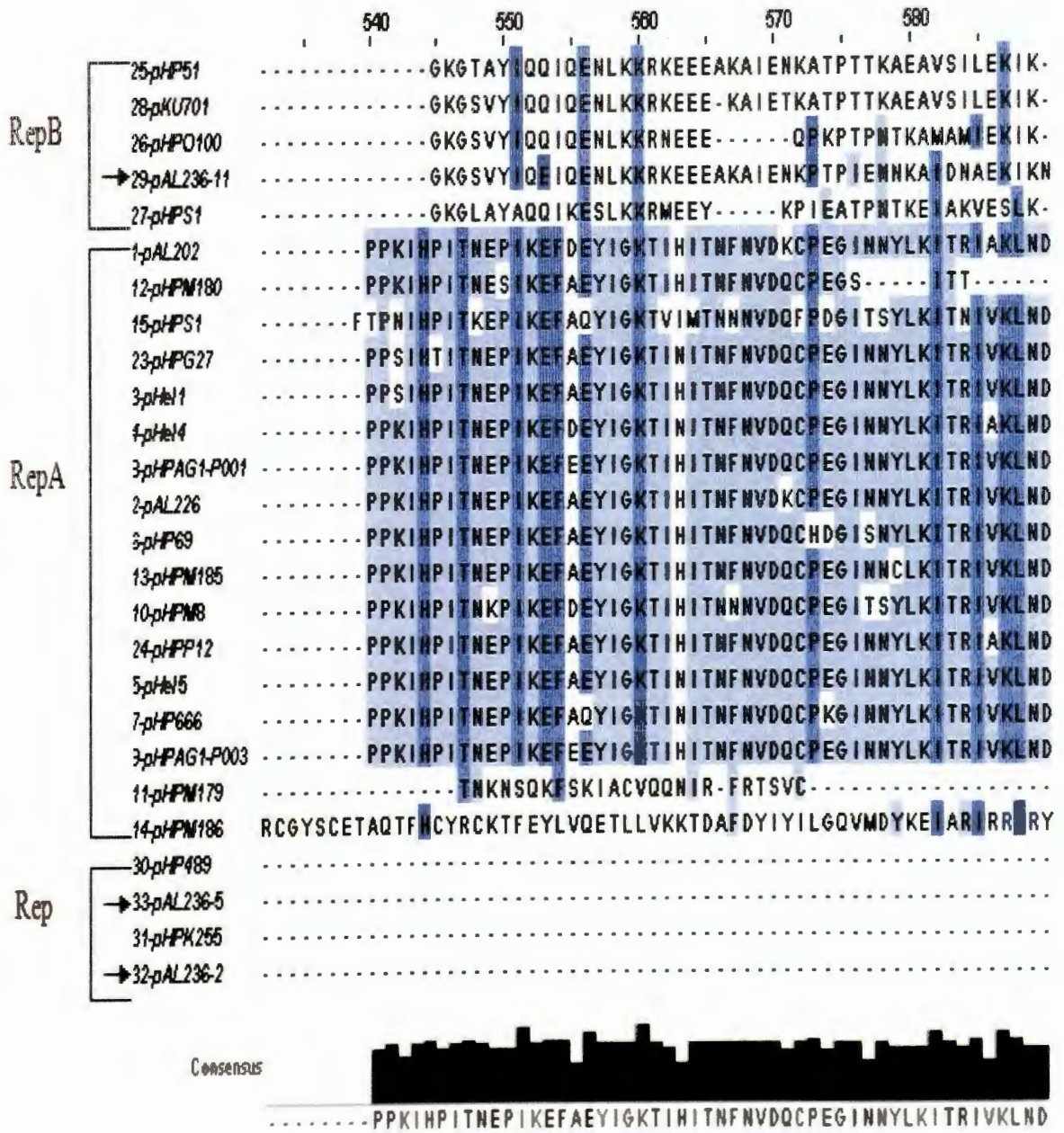

Figure 26 continued.

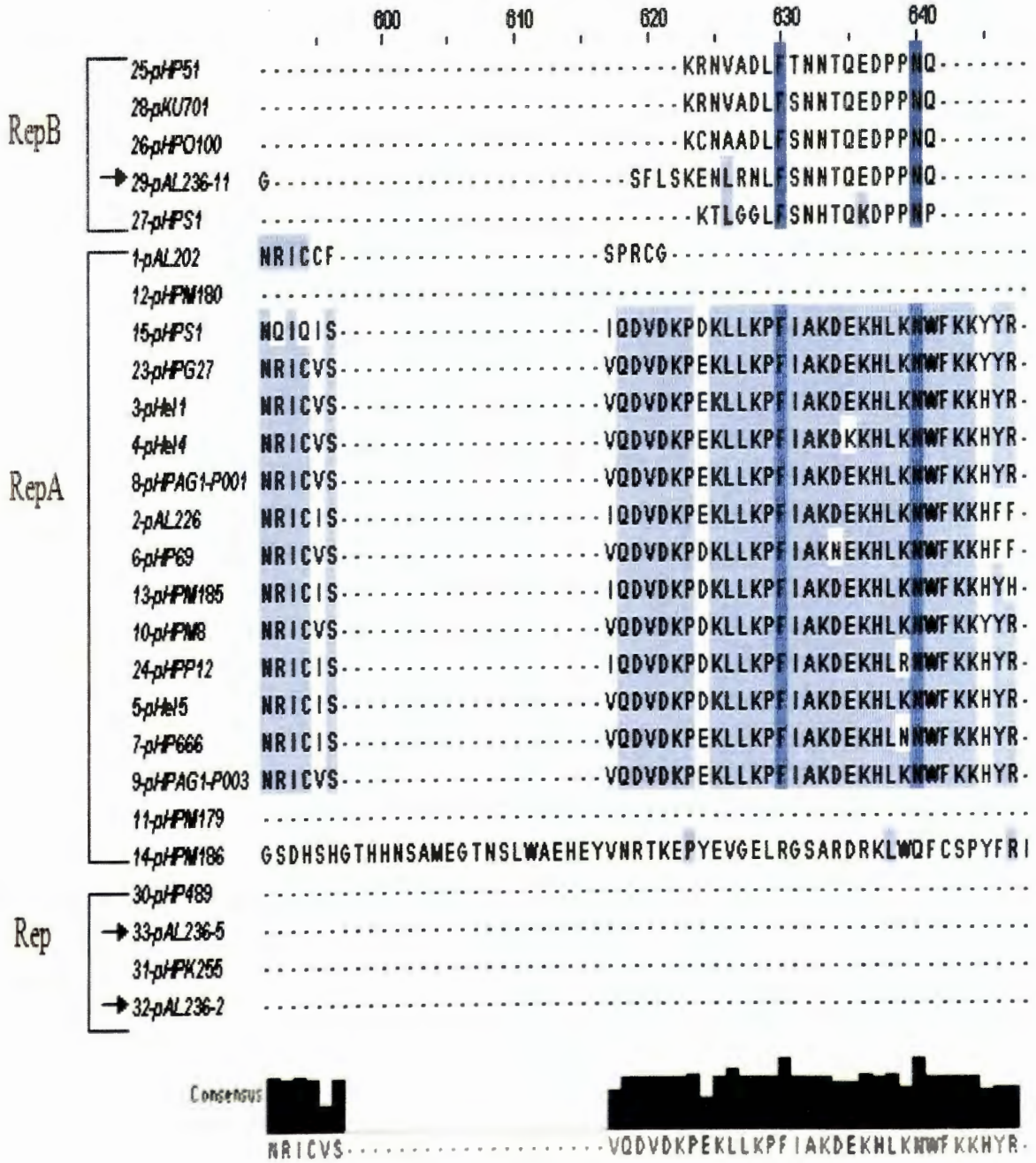

Figure 26 continued.

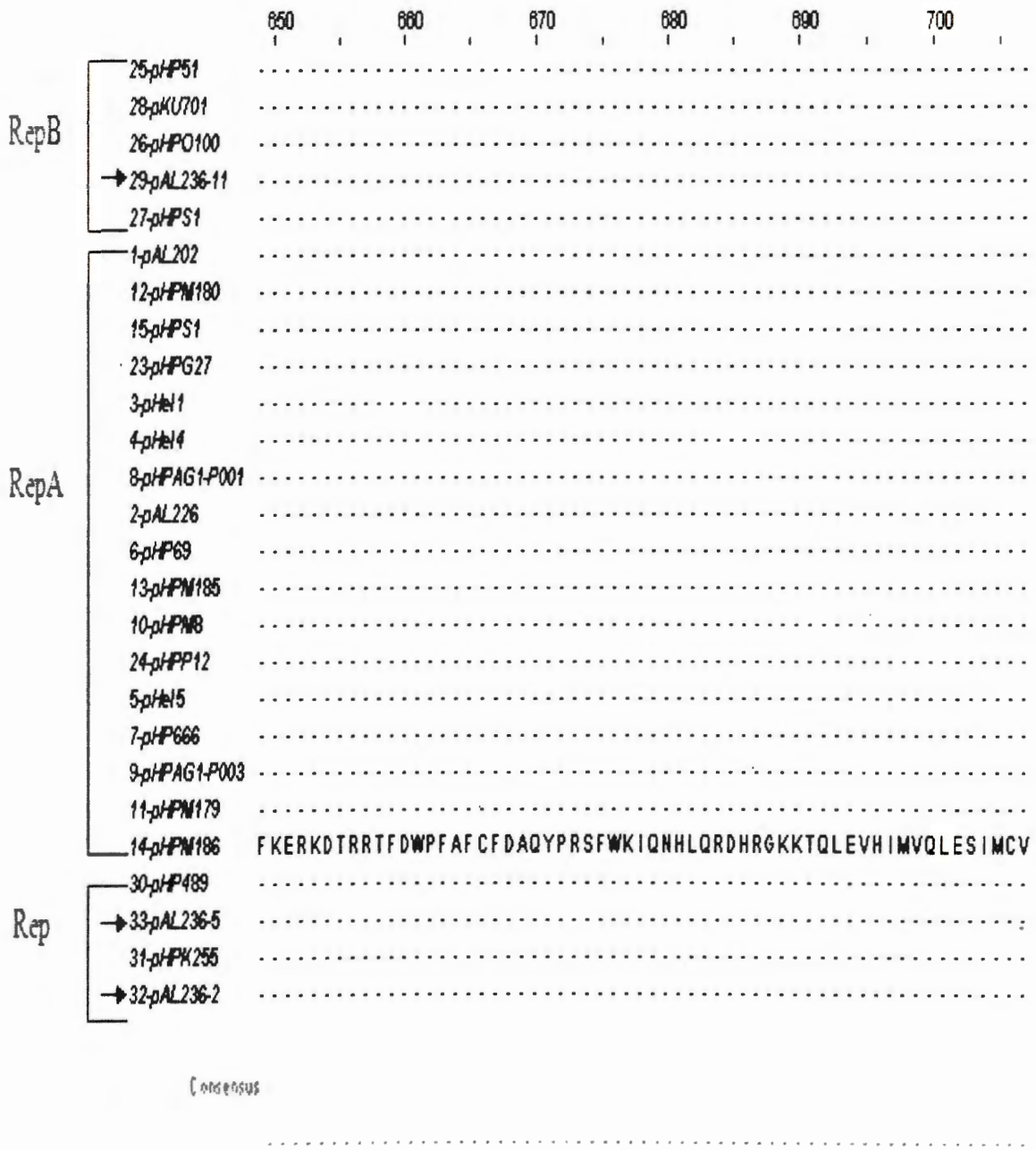

Figure 26 continued.

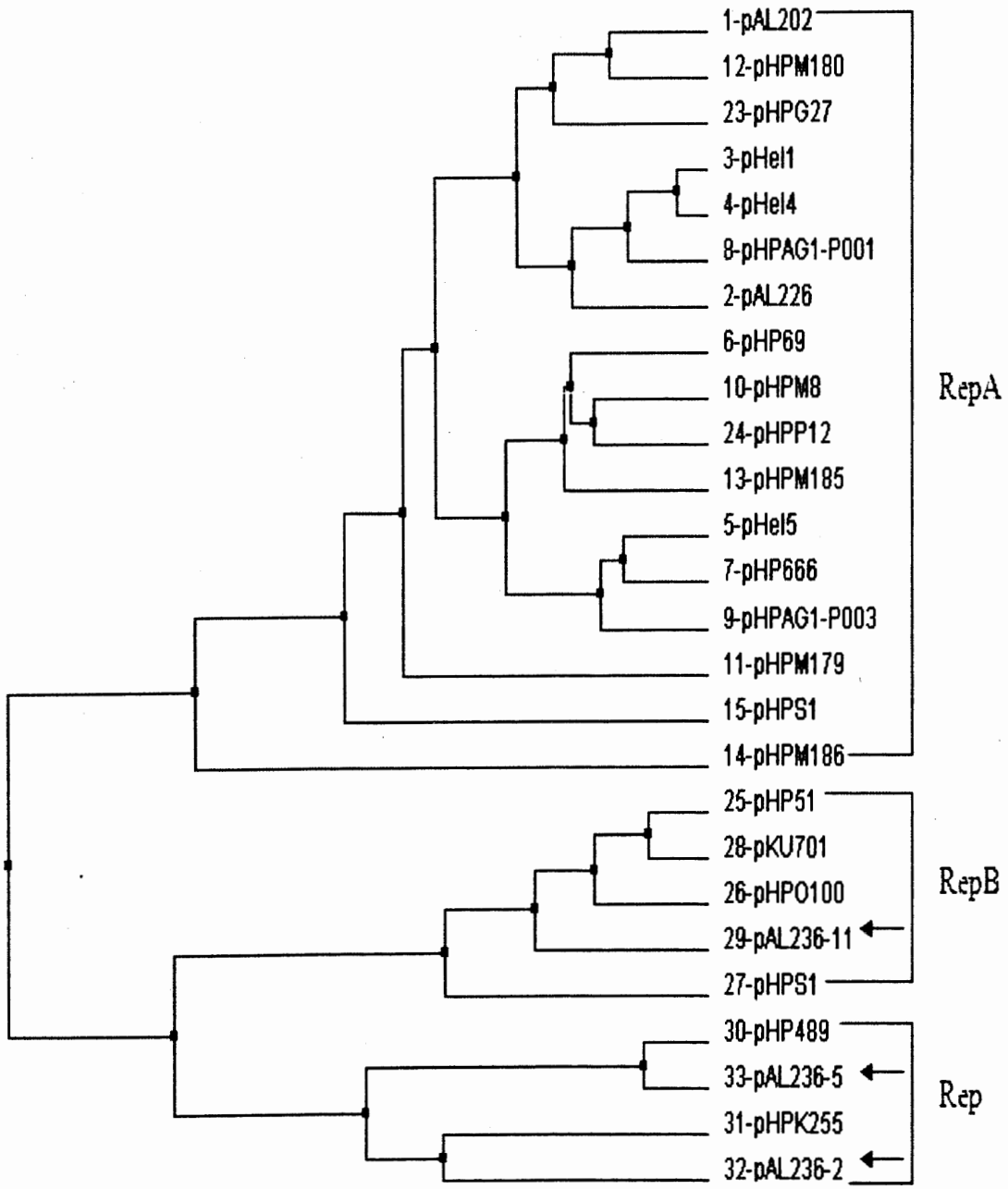

Figure 27. Average distance tree based on percent identity of the multiple sequence alignment of Rep Proteins from *H. pylori* without pMCU1-7 (#16-22) with gaps. RepA, RepB, and Rep plasmids are indicated by a box. Plasmids pAL236-2, pAL236-5, and pAL236-11 are indicated with arrows.

pHPAG1 (P001), and pAL226. However, the first clade also revealed that these plasmids belonged to sub-clades.

The first sub-clade consisted of pAL202, pHPM180 and pHPG27. The second sub-clade contained pHel1, pHel4, pHPAG1 (P001), and pAL226. The second clade contained pHP69, pHPM8, pHPP12, pHPM185, pHel5, pHP666, and pHPAG1 (P003). Within this clade these plasmids were divided into sub-clades. The first sub-clade contained pHP69, pHPM8, pHPP12, and pHPM185. Plasmids of the second sub-clade included pHel5, pHP666, and pHPAG1 (P003). Three plasmids, pHPM179, pHPS1, and pHPM186 were outliers that do not fall within the two main clades for RepA.

PHYLOGENETIC ANALYSIS OF REPLICATION PROTEINS FROM *H. pylori* WITHOUT GAPS INCLUDING pAL236-2, pAL236-5, AND pAL236-11

For the final MSA all of the replication proteins were included with gaps removed (Fig. 28). MSA of the replication proteins showed results similar to the first two MSA analyses. The plasmids with RepA had a central region that has the greatest amino acid identity with the consensus sequence. With plasmids that contained RepB and Rep, the MSA showed no significant amount of amino acid identity with the consensus. However, as the two previous MSA showed RepB had more amino acid identity with the consensus than did Rep.

A cladogram was constructed based on the MSA of the replication proteins of *H. pylori* without gaps (Fig. 29). Plasmids with Rep and RepB were arranged with each other the same way as the previous cladograms showed. As observed with the previous cladograms, the largest arrangement of the replication proteins involved RepA.

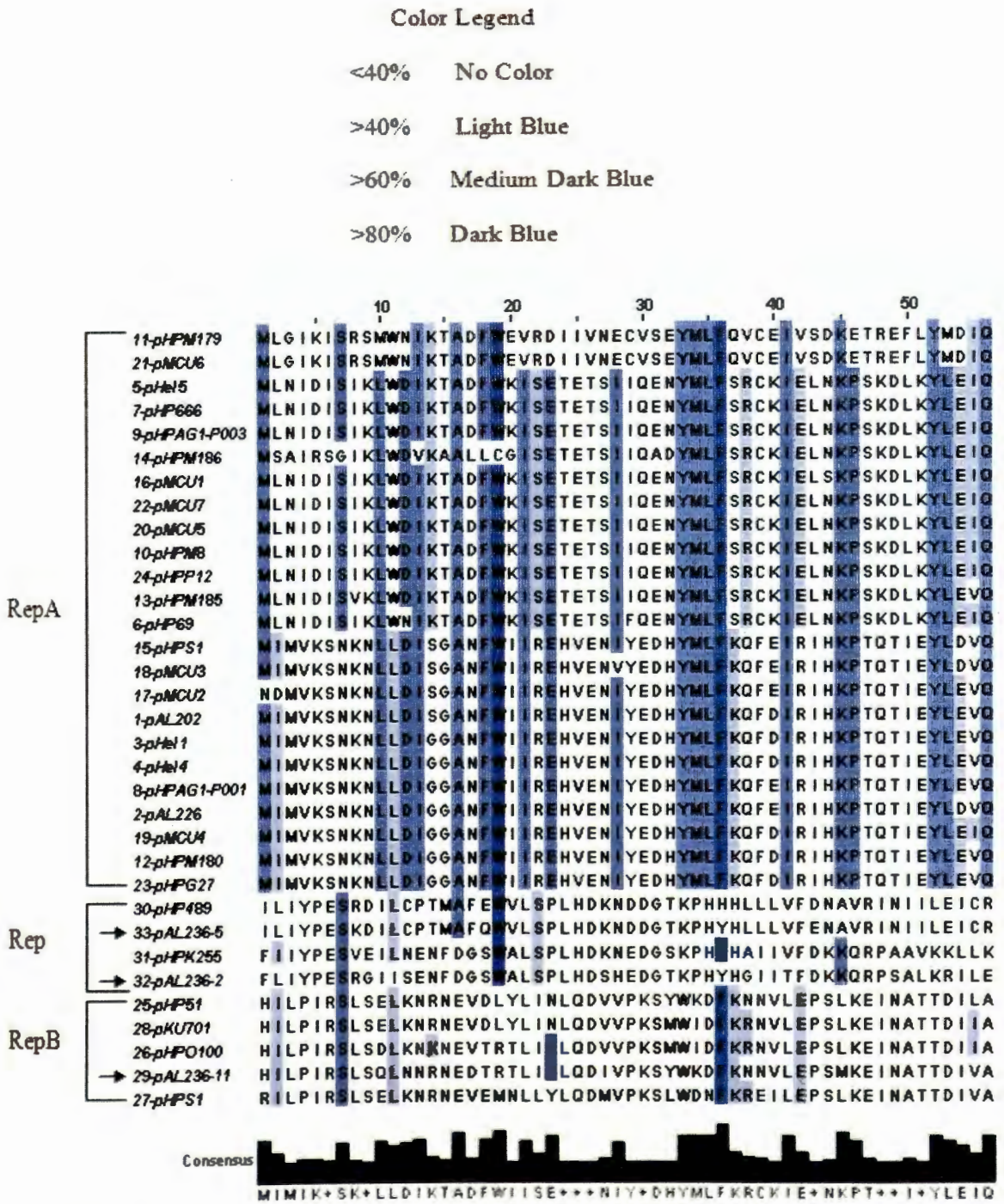

Figure 28. ClustalW2 multiple sequence alignment of all Rep Proteins from *H. pylori* with no gaps. Percent identity of amino acids are shaded. RepA, RepB, and Rep are indicated by boxes. Plasmids pAL236-2, pAL236-5, and pAL236-11 are indicated by arrows.

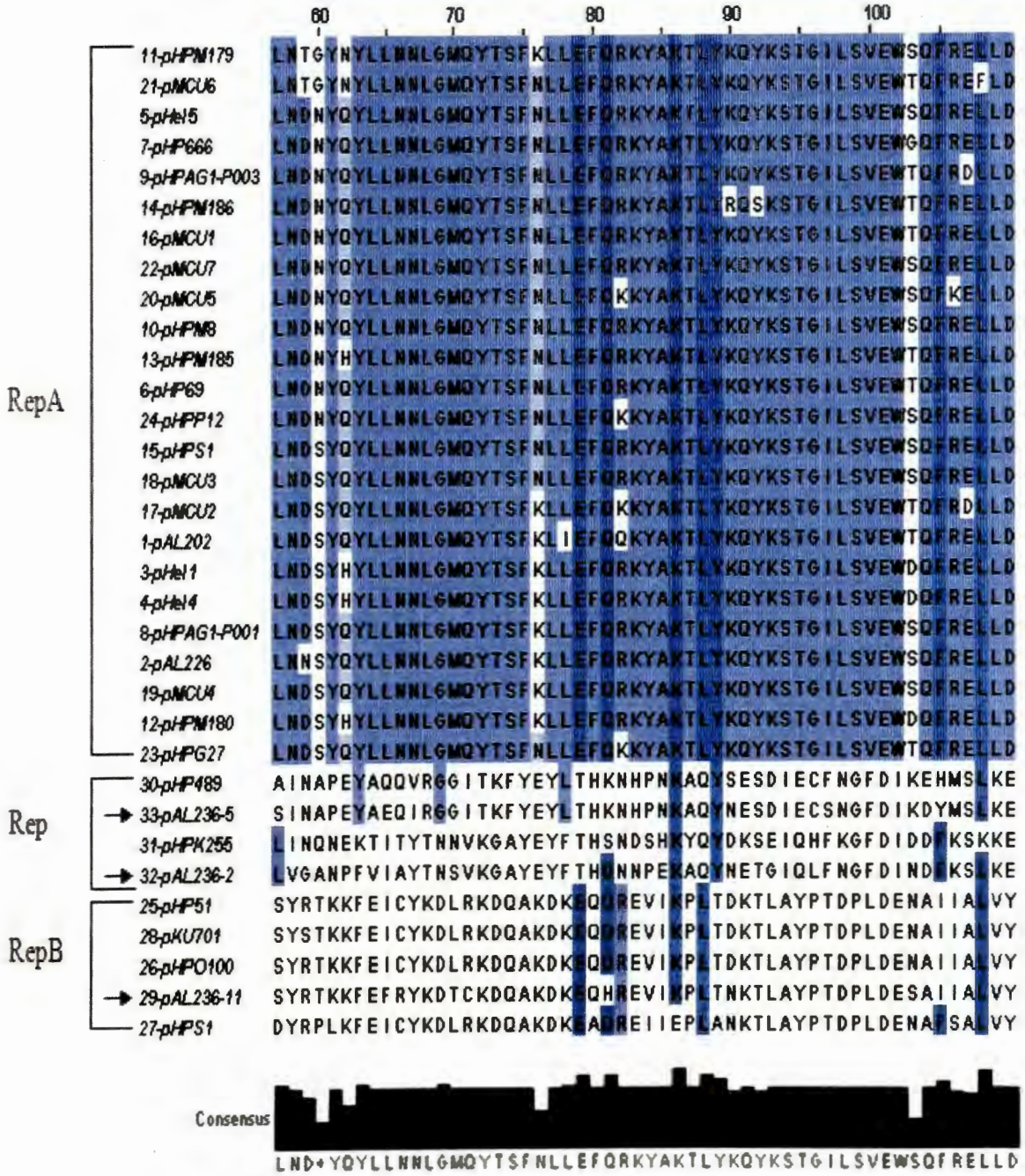

Figure 28 continued.

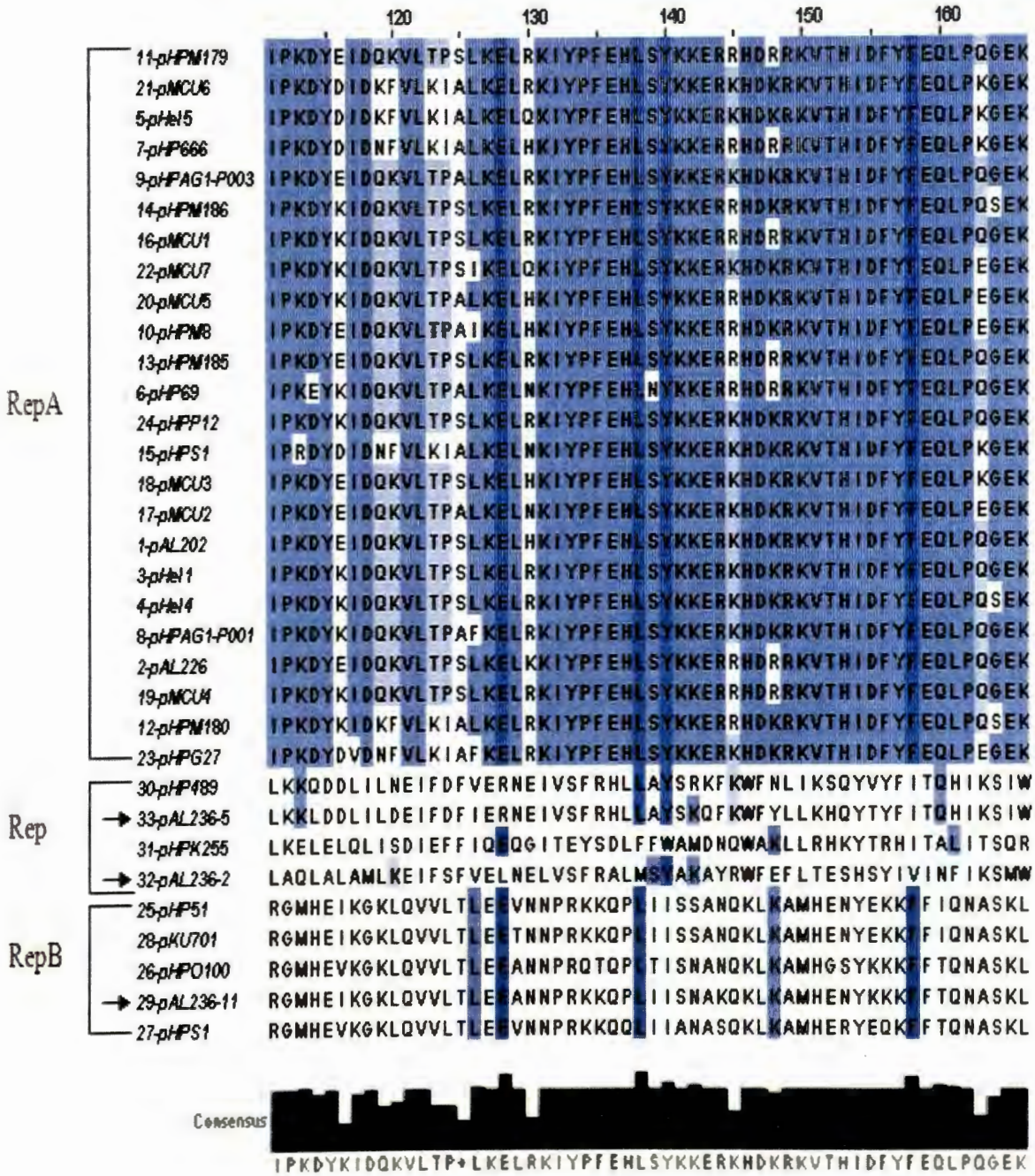

Figure 28 continued.

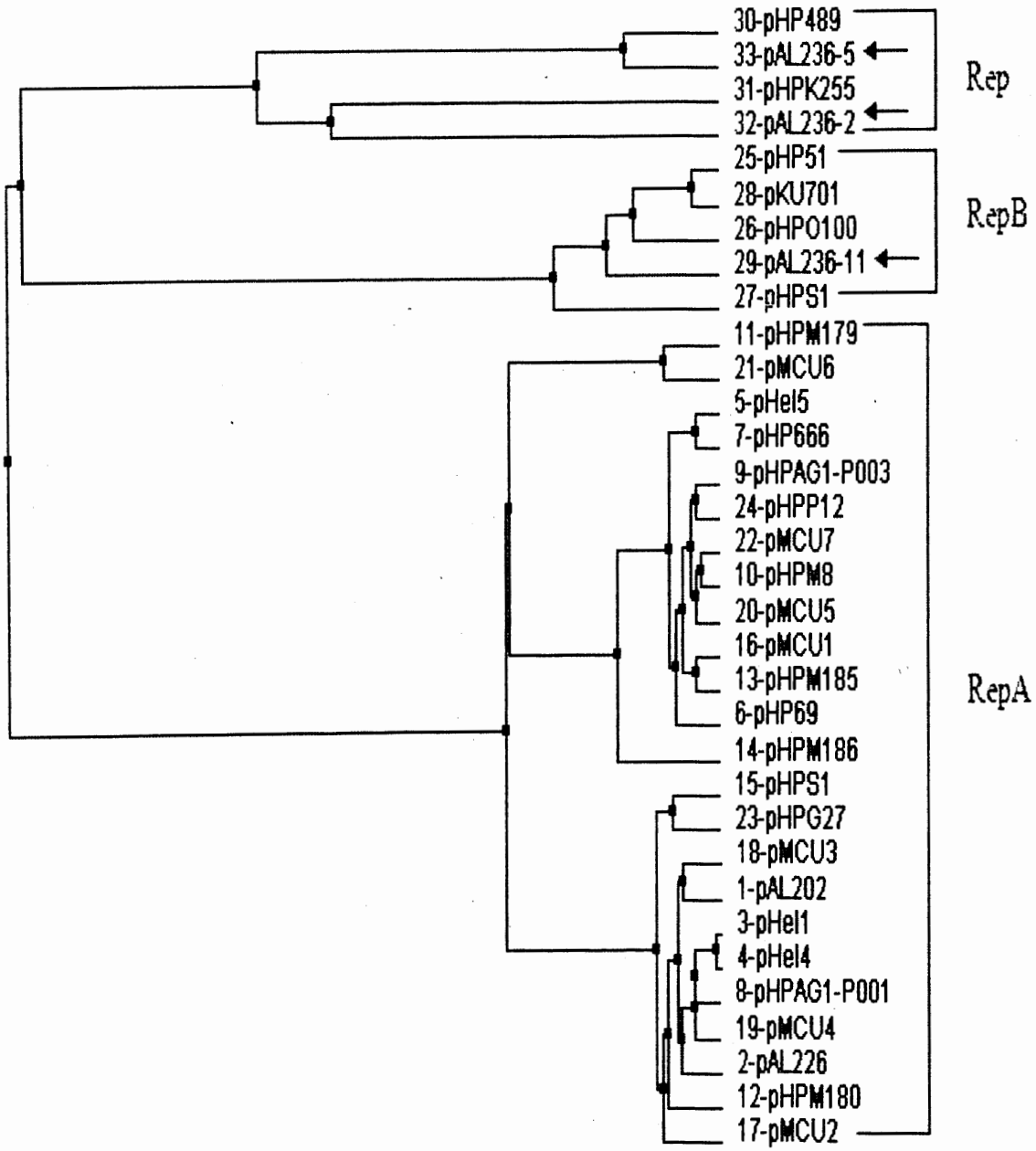

Figure 29. Average distance tree based on percent identity of the multiple sequence alignment of all Rep Proteins from *H. pylori* Plasmids with no gaps. RepA, RepB, and Rep plasmids are indicated by a box. Plasmids pAL236-2, pAL236-5, and pAL236-11 are indicated with arrows.

Each plasmid within the RepA group fell into one of the three clades. The first clade only contained plasmids pHPM179 and pMCU6. The second clade contained plasmids pHel5, pHP666, pHPAG1 (P003), pHPP12, pMCU7, pHPM8, pMCU5, pMCU1, pHPM185, pHP69, and pHPM186. This clade was divided into smaller sub-clades. The first sub-clade had two plasmids pHel5 and pHP666. A second sub-clade included pHPAG1 (P003) and pHPP12. The third sub-clade contained pMCU7, pHPM8, and pMCU5. The fourth sub-clade contained pMCU1 and pHPM185. Plasmid pHP69 appeared as an outlier of the fourth sub-clade, while plasmid pHPM186 was an outlier for the entire second clade.

Finally, the third clade consisted of pHPS1, pHPG27, pMCU3, pAL202, pHel1, pHel4, pHPAG1 (P001), pMCU4, pAL226, pHPM180, and pMCU2. Within this clade several sub-clades were observed. The first sub-clade contained only pHPS1 and pHPG27. The second sub-clade contained plasmids pMCU3, and pAL202. The third sub-clade contained pHel1, pHel4, pHPAG1 (P001), pMCU4, and pAL226. Plasmids pHPM180 and pMCU2 appeared as outliers for the third sub-clade. An interesting feature of this cladogram compared to the previous cladograms for RepA is that no outliers were observed.

CHAPTER IV

DISCUSSION

The plasmids utilized in this study were obtained by Rickets, who randomly inserted the EZ-Tn5™ <R6K*γori*/KAN-2> transposon into the total plasmid DNA of *H. pylori* strain AL236. Rickets identified three recombinant plasmids: pAL236-2; pAL236-5; and pAL236-11, each containing the EZ-Tn5™ <R6K*γori*/KAN-2> transposon. The initial DNA sequence of each plasmid was obtained utilizing transposon specific primers. However, the complete DNA sequence of each plasmid was not obtained. As a result, unlabeled primers were designed to produce M13 PCR products, specific for each recombinant plasmid, containing either a M13 forward or M13 reverse primer site. Each M13 PCR product was produced, purified, and sequenced to obtain the complete DNA sequences of pAL236-2, pAL236-5, and pAL236-11.

Results indicated that pAL236-11 was 3154 bp with a GC content of 35.9%. Analysis revealed pAL236-11 contained two ORFs. *orf1* was determined to encode the replication protein, RepB, previously identified in other *H. pylori* plasmids that replicate via the theta mechanism. Located downstream of the *repB* gene is *orf2*, which encoded an unidentified protein with no significant identity to any *H. pylori* or unrelated species proteins.

Analysis of the replication protein, RepB, of pAL236-11 revealed it was not RepA or Rep. However, both RepA and RepB proteins have been identified with

plasmids that replicate via the theta mechanism in *H. pylori*. Results have shown the greatest amino acid identity of RepB from pAL236-11 was with other RepB proteins previously identified in *H. pylori* plasmids: pKU701; pHPS1; pHP51; and RepA of pHPO100 (discussed below). A unique plasmid of this group is pHPS1 that encodes both RepA and RepB proteins (De Ungria *et al*, 1999). To date, pHPS1 is the only reported *H. pylori* plasmid that encodes both RepA and RepB proteins. As a result, De Ungria *et al.* (1999) suggested that both RepA and RepB belong in separate protein families.

Regardless of their size, RepA proteins all show a strong conserved middle region while the amino acid sequence varies at the N-terminus and C-terminus (Rickets, 2004). When analyzing RepB, the greatest conservation of amino acids appears to occur in the middle region of the protein regardless of size, similar to RepA. Both RepA and RepB, while involved with theta replication, are not the same protein, and the strong conserved middle region appears to be unique to either RepA or RepB. The amino acids of the N-terminus and C-terminus of RepB also vary, however not to the same extent as RepA. This is likely the result of there being more RepA proteins reported than RepB. A better understanding of where the most conserved amino acids are located in RepB might occur once additional RepB sequences are reported in *H. pylori* plasmids. Overall, these observations suggest that both RepA and RepB have areas of amino acid conservation, which encodes the functional portion of each respective protein.

In addition to RepB, pAL236-11 contains R3 iteron sequences located upstream of *repB*. Another iteron region has been identified in *H. pylori* theta replicons as R1, located upstream of *repA*. R1 iterons are four directly repeated sequences of 22 bp, with

the exception of pHPG27, which has four 23 bp direct repeats and pHPAG1 (P001), which consists of three 22 bp direct repeats. While the R3 iterons vary from 3.5-4 repeated sequences of 15-33 bp that have spacers between 5-18 bp. However, pKU701 and pHPS1 R3 iterons are direct repeated sequences with no spacers. Both R1 and R3 iterons are believed to serve as the binding site for either RepA (R1) or RepB (R3) for the initiation of plasmid replication (Minnis *et al.*, 1995; De Ungria *et al.*, 1999; Hosaka *et al.*, 2002). Results of a binding study performed by Vereshlingham (2003) supported these conclusions by showing that the RepA protein of pHPM8 binds to the R1 iterons of pHPM8. No binding studies of the RepB proteins to R3 iterons have been reported.

R3 iterons have been identified on *H. pylori* plasmids pKU701, pHPS1, and pHPO100. Interestingly, plasmid pHPS1 has both R1 and R3 iterons found upstream of their respective *repA* and *repB* genes (De Ungria *et al*, 1999). Another plasmid pHP51, which encodes RepB, was reported as having R1 iterons (Song *et al.*, 2003). Results of the current study showed that the R1 iterons of pHP51 are indeed R3.

R1 iteron sequences from *H. pylori* theta replicons have conserved nucleotide identities (Quiñones, 2000). Analysis of the nucleotide sequences of R3 also indicated conserved nucleotide identities among R3 iterons; however, R1 and R3 are not the same sequence. These observations provide further evidence to support the suggestion by De Ungria *et al.* (1999) in regards to pHPS1, that RepA and RepB probably act independently of each other and have binding sites specific for each replication protein. Identification of either R1 or R3 with their respective RepA or RepB proteins is very important, since each set of iterons appears to be specific for their respective replication

proteins. As a result of pAL236-11 having both RepB and R3 iterons, we propose that this plasmid is a theta replicon.

Several *H. pylori* plasmids have been reported as having an *orf2* gene, which encodes the protein termed ORF2. ORF2 has been suggested to be involved with either plasmid recombination or controlling the copy number of *H. pylori* plasmids (Minnis *et al.*, 1995; De Ungria *et al.*, 1999; Quiñones, 2000). Each *orf2* is located downstream of either *repA* or *repB* in *H. pylori* theta replicons. While not all *H. pylori* theta replicons have *orf2*, no RC plasmids identified in *H. pylori* have been reported as having *orf2*.

Based on the lack of amino acid identity it is clear that *orf2* of pAL236-11 is not the same *orf2* previously reported in *H. pylori* plasmids. In addition, *orf2* of pAL236-11 does not contain the conserved motif HPFXXGNG found with proteins that are members of the cAMP-induced filamentation (*fic*) gene (Utsumi *et al*, 1982; Komano *et al*, 1991). Each of these *H. pylori* ORF2 proteins, with the exception of pHPS1, contained the conserved domain HPFXXGNG. Since the ORF2 in pAL236-11 had no significant amino acid identity with any previously reported protein, the exact role of this protein remains unknown.

Results indicated that pAL236-2 and pAL236-5 were 1448 and 1216 bp, respectively. For plasmid pAL236-2 the GC content was 37.5% while pAL236-5 had a GC content of 33.8 %. Both pAL236-2 and pAL236-5 each encoded a single ORF1. The *orf1* gene was determined to encode the replication protein, Rep, previously identified in other *H. pylori* plasmids that replicate via the RC mechanism.

Analysis of the Rep protein from each plasmid revealed the protein was not RepA or RepB. Further analysis showed the Rep proteins from pAL236-2 and pAL236-5 had identities with the Rep proteins of *H. pylori* RC plasmids pHPK255 and pHP489. Besides pHPK255 and pHP489, the Rep proteins of pAL236-2 and pAL236-5 showed amino acid identities with each other as well. Two important findings were obtained from these results. First, pAL236-2 and pAL236-5 were potential RC plasmids. Second, since the Rep proteins of pAL236-2 and pAL236-5 were not identical, each plasmid had their own Rep protein for RC replication.

Additional evidence indicated that pAL236-2 and pAL236-5 were most likely RC plasmids. First, each plasmid had been identified as having the ability to potentially form hairpin-loop structures needed to form the double-stranded origin (DSO) via inverted repeats (del Solar *et al.*, 1993; Kleanthous *et al.*, 1991; del Solar *et al.*, 1998). Second, regardless of the size of each potential DSO, both pAL236-2 and pAL236-5 contained potential DSO nick sequences, previously identified in other RC plasmids, which are utilized by the Rep protein to initiate leading-strand replication (del Solar *et al.*, 1993; Kleanthous *et al.*, 1991; del Solar *et al.*, 1998). Third, pAL236-2 had iteron sequences and pAL236-5 had short direct repeats near their respective DSOs that could serve as a potential binding site of the Rep protein to initiate RC replication (del Solar *et al.*, 1993; del Solar *et al.*, 1998; Khan, 1997; Khan, 2000; Khan, 2005). Finally, both Rep proteins of pAL236-2 and pAL236-5 contained previously identified protein motifs found with Rep proteins from other RC plasmids including pHPK255 and pHP489 (Kleanthous *et al.*, 1991; Ilyina *et al.*, 1992; del Solar *et al.*, 1993; Song *et al*, 2003).

Three of these motifs included the possible metal-binding domain (motif 3), potential catalytic tyrosine residue (motif 4), and the second tyrosine residue which could be involved in linking the Rep protein to the plasmid DNA (motif 5) (Ilyina *et al.*, 1992; del Solar *et al.*, 1993).

Plasmids pAL236-2 and pAL236-5 were compared against known single-stranded origins needed to convert ssDNA into dsDNA (Boe *et al.*, 1989; Novick, 1989; Gruss *et al.*, 1989; Seegers *et al.*, 1995; Devine *et al.*, 1989; Zaman *et al.*, 1993; Meijer *et al.*, 1995; del Solar *et al.*, 1998; Khan, 1997; Khan, 2000; Khan, 2005). However, neither plasmid appeared to have sequences that could potentially form a SSO. Of all the *H. pylori* RC plasmids, only pHPK255 has both a DSO and SSO (Kleanthous *et al*, 1991). Since pHP489, pAL236-2, and pAL236-5 do not have an identified SSO there is a possibility these plasmids utilize previously unidentified sequence(s) or protein(s) that are involved with the conversion of ssDNA to dsDNA. A possible candidate for the conversion of ssDNA to dsDNA could involve the inverted repeats identified pAL236-2 and pAL236-5 that are not involved with the DSO.

So far there are 17 different RC plasmid families (http://www.essex.ac.uk/bs/staff/osborn/DPR_home.htm) based on the homology of their replication proteins and DSOs (Khan, 2005). Of these 17 groups, we propose that pAL236-2 and pAL236-5 belong to the Group II RC family. This is based on the Rep identities of pAL236-2 and pAL236-5 with other Group II family members including pHPK255 and pHP489. Additional evidence includes the DSO nick sequences found on pAL236-2 (TACACG/A) and

pAL236-5 (TACTACG/A) that were either similar or exactly the same as other members of the Group II family including pHPK255 and pHP489 (TACTACG/A).

Several reports suggest that *H. pylori* plasmids have the ability to be involved with plasmid to plasmid and chromosome to plasmid recombination events involving insertion sequences (IS), R2 repeats, R4 repeats, and short stretches of *H. pylori* genome sequences on several *H. pylori* plasmids (Minnis *et al.*, 1995; Heuermann, *et al.*, 1995; Qasem, 1995; Taylor, 1997; Tomb *et al.*, 1997; Burnham, 1998; Liu, 1998; Alm *et al.*, 1999; Gao, 1999; Hofreuter *et al.*, 1999; De Ungria *et al.*, 1999; Doig, *et al.*, 1999; Quiñones, 2000; Hosaka, *et al.*, 2002; Rickets, 2004; Song *et al.*, 2003; Khatun, 2005; Oh *et al.*, 2006). Hofreuter *et al.* (2002) proposed models in an attempt to show how recombination events might occur. The first model, based on the recombination events involving both R2 and R4 repeats, is an attempt to explain the different sizes of *H. pylori* plasmids. The second model was constructed to help explain the possible exchange of genes between *H. pylori* plasmids and genomes involving IS sequences. The authors also suggest that other proteins encoded by *H. pylori* plasmids or genomes could be involved with recombination events as well.

Analysis of the non-coding regions of pAL236-11 did not reveal any IS or identity with any *H. pylori* genomic sequences. However, two important findings were related to the non-coding regions of pAL236-11. First, the non-coding regions of pAL236-11 had identities with R3 iterons found on plasmids pKU701, pHP51, and the R1 iterons of pHP51, which as mentioned above, led to the discovery that R1 of pHP51 was in fact an R3 iteron. Second, the most interesting result was the discovery of two

potential R2 repeats, previously unidentified, in *H. pylori* plasmids pHPP12 (unpublished, NC_011499) and pHPAG1 (Oh *et al.*, 2006). The R2 repeats for pHPP12 consisted of a two time 683 bp direct repeat separated by 46 bp. While plasmid pHPAG1 had R2 repeats that consisted of a two time 600 bp repeat separated by 2781 bp (2.7 kbp). As a result, these observation of these R2 repeats provided further evidence to the model suggested by Hofreuter *et al.* (2002), that the different sizes of *H. pylori* plasmids is a result of recombination events involving the R2 repeats.

The pAL236-11 non-coding region contained identities with numerous *H. pylori* theta plasmids, including the R2 repeats of pHPS1, pHPAG1, and pHPP12. This region also included identities with the common region (CR), which are areas of identity with other *H. pylori* plasmids, identified on plasmids pHP51 (Song *et al.*, 2003) and pHP69 (unpublished, NC_010884) (De Ungria *et al.*, 1999). One observation was that the area of identity of pAL236-11 CR included the region of identity with the R2 repeats of pHPS1, pHPAG1, and pHPP12. This could indicate that the CR, along with the non-coding regions of other *H. pylori* plasmids with identity with pAL236-11, might be remnants of R2 repeats that are unrecognizable as the result of recombination events. This also suggests that partial or complete sets of R2 repeats identified on plasmids pAL202, pAL226, pHel4, pHel5, pHPM180, pHP51, pHPS1, pHPP12 and pHPAG1 supports the possibility of more than one type of R2 repeat in *H. pylori* plasmids. Since the non-coding regions of pAL236-11 had identity with the R2 repeats of pHPS1, pHPAG1, and pHPP12, pAL236-11 might have had a complete or partial set R2 repeats at one time.

Besides R2, another repeated sequence in several *H. pylori* theta replicons was identified as R4 (Hofreuter *et al.*, 2002). Unlike the R2 repeats, R4 has only been identified in theta replicons that have a *repA* gene. The only exception was pHPS1, which encodes both *repA* and *rep*B genes, where a single copy of the R4 repeat was found at the 3' end of the *repA* gene (Hofreuter *et al.*, 2002). However, as more RepB proteins are reported it may lead to the discovery of the R4 repeats for *H. pylori* plasmids with RepB. Analysis indicated that pAL236-11 had no R4 repeats.

Analyses of the non-coding regions from pAL236-2 and pAL236-5 revealed no IS or R4 repeats. To date, no RC plasmid from *H. pylori* has identified IS or the R4 repeats unlike the theta replicons. Interestingly the non-coding regions from both of these plasmids had various identities with *H. pylori* theta and RC plasmids, along with identities to each other. However, this is not the first observation of identity between a non-coding region of a RC plasmid and a non-coding region of a theta replicon. Minnis *et al.*, 1994, showed a 71% identity of 200 nt from the theta plasmid pHPM180 with the RC plasmid, pHPK255. Interestingly, 85 nt of ORF2 and the first portion of the 232 nt R2 repeat from pHPM180 had identity with the non-coding region of pHPK255.

For plasmid pAL236-2 the regions of identity with the non-coding region was the greatest with the RC plasmid pHPK255, while pAL236-5 has the greatest identity with the RC plasmid pHPK489. Between pAL236-2 and pAL236-5, only pAL236-2 had identity with both pHPK255 and pHP489 non-coding regions. A surprising observation was that the non-coding region of pAL235-5 did not show any significant identity with the non-coding region of pHPK255, except in the Rep protein and DSO nick sequence.

This observation, along with the lack of significant identity with the non-coding region of pAL236-11, gives further evidence that pAL236-2, pAL236-5, and pAL236-11 are separate individual plasmids.

Plasmid pAL236-2 also had identities with IR sequences from both pHPK255 and pAL236-5 that were not involved with the formation of the DSO for both plasmids or the SSO of pHPK255. Interestingly a portion of the non-coding regions from pAL236-5 had identities with the R2 repeat regions of plasmids pHPM180, pHPS1, pHPP12, and pHPAG1, as well as with the CR of pHP69. This observation was similar for pAL236-11, but only pAL236-5 had identity with the R2 repeats from pHPM180. This could indicate that pAL236-5 had complete or partial R2 repeats at some point that are no longer identifiable as the result of recombination events, similar to pAL236-11. Other evidence that recombination could have occurred in pAL236-2 and pAL236-5 was the small identity these plasmids had with the same non-coding region of *H. pylori* genome P12. These observations support De Ungria *et al.* (1998) who suggest that the regions of identity shared between both *H. pylori* plasmids and chromosome sequences most likely are the result of recombination events, indicating that *H. pylori* plasmids could have contributed to the "extensive genetic heterogeneity" of *H. pylori.*

Phylogenetic analysis was performed to organize the known plasmid sequences of *H. pylori.* The replication protein was the best candidate for analysis since it was the one gene conserved among all of the plasmids. Based on the MSA and cladogram results, it is clear that there were three families of replication proteins RepA, RepB, and Rep. Based on the report for pHP489 by Song *et al.* (2003), which called the Rep protein

RepH, we propose that the Rep family described here be named the RepH family. Thus, *H. pylori* plasmids, based on replication proteins, have the ability to replicate either by theta (RepA or RepB) or RC (RepH) mechanism. However, theta replication involving RepA appeared as the predominate form of plasmid replication identified in *H. pylori* plasmids.

Results of the cladograms revealed the possibility that more than one type of RepA exists among *H. pylori* plasmids. A binding study was conducted previously in an attempt to show the binding of RepA from pHPM8 with the R1 iterons of pHPM180 (Veereshlingam, 2003). The results failed to show binding of pHPM8 RepA to the R1 iterons of pHPM180. As a result, Veereshlingam (2003) concluded that iteron-RepA binding is strain specific, based on the ability of RepA of pHPM8 to bind to R1 of pHPM8, while RepA of pHPM8 did not bind to R1 of pHPM180. Cladograms revealed that the RepA of pHPM8 belonged to a different clade than pHPM180, which could explain the lack of binding of RepA of pHPM8 to the R1 region of pHPM180. Since both RepA and R1 iterons contain conserved sequences, Rickets (2004) suggested that binding of RepA to their respective R1 iterons could involve other proteins that contribute to the binding of RepA to the R1 iterons through the recognition of amino acid sequences of RepA not found on other RepA proteins from *H. pylori*. The same suggestion could also apply to both RepB and R3.

Another interesting plasmid is pHPAG1 (Oh *et al.*, 2006) that encodes two RepA proteins located in separate clades of the cladograms. This could be one of the reasons why pHPAG1 contains two different sets of R1 iteron regions that could be specific for

their respective RepA protein. Plasmid pHPAG1 has a complete set of R1 iterons that consist of four 22 bp tandem repeats upstream from one of the *repA* genes encoding RepA (pHPAG1 P003). For the second *repA* gene encoding RepA (pHPAG1 P001), the R1 region consists of only three 22 bp tandem repeats. As a result, one of the *repA* genes could be functional while the other *repA* gene is not.

Currently, pHPO100 (unpublished, AF056496) contains a replication protein identified in GenBank at NCBI as RepA. The MSA and cladograms, along with comparisons to the RepB protein and R3 iterons of pAL236-11, revealed that the replication protein of pHPO100 is RepB and not RepA. This result confirmed the previous suggestion by De Ungria *et al.* (1999) that both RepA and RepB belong in separate plasmid families. Additional support comes from the observation that the iteron region of pHPO100 contains iterons similar to R3 found in plasmids with RepB, and not R1 iterons of plasmids that encode RepA. Our results also confirm that pHPS1 contains both the RepA and RepB replication proteins that were previously identified (De Ungria *et al*, 1999).

These results for plasmids pAL236-2, pAL236-5, and pAL236-11 showed for the first time that *H. pylori* contained plasmids that replicate either by the theta or RC mechanism within the same strain. For plasmid pAL236-11, cladogram results showed the replication protein of this plasmid was grouped with plasmids having RepB and not with RepA or RepH plasmids. Plasmid pAL236-11 not only encoded RepB, but also contained the R3 iterons found with RepB plasmids that are not found with plasmids that encode RepA, which have R1 iterons. However, based on the cladogram results it

appeared that more than one type of RepB protein could be involved with theta replication for RepB plasmids.

Both plasmids pAL236-2 and pAL236-5 contained Rep proteins with identity to RC plasmid replication proteins and were not RepA or RepB. All three cladograms confirmed that the Rep protein of pAL236-2 was more closely related to pHPK255 Rep and pAL236-5 was more closely related to the RepH protein of pHP489. This result also indicated that there could be more than one type of Rep protein involved with RC plasmids in *H. pylori*. This finding also supports the analysis of the non-coding region which identified pAL236-2 as having more identity with the non-coding regions of pHPK255, while the non-coding region of pAL236-5 had more identity with the non-coding region of pHP489. Interestingly, phylogenetic analysis of the replication proteins from *H. pylori* plasmids, regardless of the mode of replication, revealed the importance of properly identifying the replication proteins. As the sequences of additional *H. pylori* plasmids are determined, the possibility exists that a previously unknown replication protein(s) could be identified that is not RepA, RepB, or RepH. Future phylogenetic analysis would be necessary to reveal the relationship that these replication proteins have with either RepA, RepB, or RepH. A summary for the general features of pAL236-2, pAL236-5, and pAL236-11 are shown in Table 12.

A few questions remain regarding the replication proteins of *H. pylori* plasmids. First, could RepA, RepB, or RepH initiate plasmid replication of plasmids within the same clade or sub-clade of their respective replication proteins? Second, does each plasmid encode a strain specific replication protein that will only initiate replication of

Table 12. General Features of pAL236-2, pAL236-5, and pAL236-11.

Plasmid Features	pAL236-2	pAL236-5	pAL236-11
Size	1448 bp	1216 bp	3148 bp
% GC	37.5%	33.8%	35.9%
Number of ORFs	1	1	2
Replication Protein	RepH	RepH	RepB
Mode of Replication	Rolling-Circle	Rolling-Circle	Theta
Nucleotide Sequences for Replication Protein Binding	Iterons consisting of a 2 time 17 bp direct repeat separated by 6 bp	Direct Repeat consisting of a 2 time 5 bp direct repeat separated by 2 bp	R3 Iterons consisting of a 4 time 15 bp direct repeat with 17-18 bp spacer sequence between each repeat
Double Stranded Origin (DSO)	Yes	Yes	No
DSO Nick Sequence	TACACG\A	TACTACG\A	No
Identity with R2 Repeat Sequences	No	Yes	Yes
Inverted Repeats	Yes	Yes	No

their respective plasmid regardless of their clade or sub-clade? These questions remain unanswered and require further investigation.

AL236 was the first strain of *H. pylori* to show multiple plasmids in our lab. This result was exciting since other *H. pylori* strains in our lab only contained a single plasmid with their respective strain. However, *H. pylori* strains with multiple plasmids have been indicated. For example, Lee *et al.* (1997) determined, while examining gastric endoscopic biopsy specimens from Korean patients, that of 108 *H. pylori* strains screened for plasmids only 77% of these strains appeared to contain one or more plasmids. What remains unclear is whether or not the strains that indicated multiple plasmids had separate individual plasmids or were they the same plasmid arranged in different conformations?

This study showed that pAL236-2, pAL236-5, and pAL236-11 are three separate plasmids isolated from *H. pylori* strain AL236. Thus *H. pylori* strains can regulate multiple plasmids regardless of the mode of replication. However, a few questions remain about these plasmids. First, how are these plasmids distributed among the bacterial cells of strain AL236? Second, if multiple plasmids exist in a strain of *H. pylori*, how are the plasmids regulated in regards to replication or expression of potential genes encoded on the plasmid?

In summary, these sequence analyses showed that *H. pylori* plasmids, while genetically diverse, can be organized into three replication families: RepA; RepB; or RepH. Phylogenetic analyses suggest the early evolutionary separation of the RepA family into two or three main branches. The detailed sequence analysis used in this study could assist with the identification of future *H. pylori* plasmid replication proteins

to ensure proper identification as RepA, RepB, and RepH. Similarly, these techniques could aide future investigations in identifying novel replication proteins that are not RepA, RepB, or RepH. This study also showed the importance of analyzing the non-coding regions of *H. pylori* plasmids. Such analysis of the non-coding regions of these plasmids led to the identification of two potential R2 repeats that could be involved with recombination events. Indeed, it was suggested that chromosome recombination events might help *H. pylori* evade the immune system (Blaser *et al.*, 2004). Rickets (2004) suggested that since *H. pylori* plasmids have come from clinical isolates, recombination events involving plasmids could contribute to both the pathogenicity and overall fitness of *H. pylori*. The exact role of these plasmids remains unclear.

REFERENCES

1. Alm, R. A., L. L. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig, D. R. Smith, B. Noonan, B. C. Guild, B. L. deJonge, G. Carmel, P. J. Tummino, A. Caruso, M. Uria-Nickelsen, D. M. Mills, C. Ives, R. Gibson, D. Merberg, S. D. Mills, Q. Jiang, D. E. Taylor, G. F. Vovis, and T. J. Trust. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397:176-180.

2. Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

3. Altschul, S. F., W. Gish, W. Miller, and E. W. Myers. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-10.

4. Altshcul, S. F., J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schäffer, and Y-K Yu. 2005. Protein database searches using compositionally adjusted substitution matrices. FEBS J. 272:5101-5109.

5. Baltrus, D. A., M. R. Amieva, A. Covacci, T. M. Lowe, D. S. Merrell, K. M. Ottemann, M. Stein, N. R. Salama, and K. Guillemin. 2009. The complete genome sequence of *Helicobacter pylori* strain G27. J. Bacteriol. 191:447-448.

6. Blaser, M. J. and D. E. Berg. 2001. *Helicobacter pylori* genetic diversity and risk of human disease. J. Clin. Invest. 107:767-773.

7. Blaser, M. J., and J. C. Atherton. 2004. *Helicobacter pylori* persistence: biology and disease. J. Clin. Invest. 113:321-333.

8. Boe, L., M. F. Gros, H. te Riele, S. D. Ehrlich, and A. Gruss. 1989. Replication origins of single-stranded-DNA plasmid pUB110. J. Bacteriol. 171:3366-3372.

9. Bourzac, K. M., C. M. Botham, and K. Guillemin. 2007. *Helicobacter pylori* CagA induces AGS cell elongation through a cell retraction defect that is independent of Cdc42, Rac1, and Arp2/3. Infect. Immun. 75:1203-1213.

10. Brown, L. M. 2000. *Helicobacter pylori*: Epidemiology and routes of transmission. Epidem. Rev. 22:283-297.

11. Burnham, K. D. 1998. Plasmid contains unique *Helicobacter pylori* insertion sequence, IS605. Dissertation. Texas Woman's University, Denton, TX.

12. Censini, S., C. Lange, Z. Xiang, J. E. Crabtree, P. Ghiara, M. Borodovsky, R. Rappuoli, and A. Covacci. 1996. *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. Proc. Natl. Acad. Sci. USA 93:14648-14648.

13. Chalker A. F., H. W. Minehart, N. J. Hughes, K. K. Koretke, M. A. Lonetto, K. K. Brinkman, P. V. Warren, A. Lupas, M. J. Stanhope, J. R. Brown, and P. S. Hoffman. 2001. Systematic identification of selective essential genes in *Helicobacter pylori* by genome prioritization and allelic replacement mutagenesis. J. Bacteriol. 183:1259-1268.

14. Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton. 2004. The Jalview Java alignment editor. Bioinformatics 20:426-7.

15. Couturier, M. J., E. Tasca, C. Montecucco, and Markus Stein. 2006. Interaction with CagF is required for translocation of CagA into the host via the *Helicobacter pylori* Type IV secretion system. Infect. Immun. 74:273-281.

16. Cover, T. L., U. S. Krishna, D. A. Israel, and R. M. Peek, Jr. 2003. Induction of gastric epithelial cell apoptosis by *Helicobacter pylori* vacuolating cytotoxin. Can. Res. 63:951-957.

17. De Ungria, M. C. A., T. Kolesnikow, P. T. Cox, and A. Lee. 1999. Molecular characterization and interstrain variability of pHPS1, a plasmid isolated from the Sydney Strain (SS1) of *Helicobacter pylori*. Plasmid 41:97-109.

18. De Ungria, M. C. A., D. Tiller, B. A. Neilan, P. T. Cox, and A. Lee. 1998. A novel method of extracting plasmid DNA from *Helicobacter* species. Helicobacter 3:269-277.

19. del Solar, G. H., R. Diaz, and M. Espinosa. 1987. Initiation signals for the conversion of single stranded to double stranded DNA forms in the streptococcal plasmid pLS1. Nucleic. Acids Res. 15:5561-5580.

20. del Solar, G. H., R. Giraldo, M. J. Ruiz-Echevarria, M. Espinosa, and R. Diaz-Orejas. 1998. Replication and control of circular bacterial plasmids. Microbiol. Mol. Biol. Rev. 62:434-464.

21. del Solar, G. H., M. Moscoso, and M. Espinosa. 1993. Rolling-circle replicating plasmids from Gram-positive and Gram-negative bacteria: a wall falls. Mol. Microbiol. 8:789-796.

22. Devine, K. M., S. T. Hogan, D. G. Higgins, and D. J. McConnell. 1989. Replication and segregational stability of the Bacillus plasmid pBAA1. J. Bacteriol. 171:1166-1172.

23. Doig, P., B. L. de Jonge, R. A. Alm, E. D. Brown, M. Uria-Nickelsen, B. Noonan, S. D. Mills, P. Tummino, G. Carmel, B. C. Guild, D. T. Moir, G. F. Vovis, and T. J. Trust. 1999. *Helicobacter pylori* physiology predicted from genomic comparison of two strains. Microbiol. Mol. Biol. Rev. 63:675-707.

24. Dubois, A. 1995. Spiral bacteria in the human stomach: The gastric *Helicobacters*. Emerg. Infect. Dis. 1:79-83.

25. Dunn, E. B., H. Hartley, and M. J. Blaser. 1997. *Helicobacter pylori*. Clin. Microbiol. Rev. 10:720-741.

26. Fan, F., H. Z. Gunasena, R. Cheng, S. E. Espejo, E. Crowe, P. B. Ernst, and V. E. Reyes. 2000. *Helicobacter pyori* urease binds to class II MHC on gastric epithelial cells and induces their apoptosis. J. Immun. 165:1918-1924.

27. Gao, X. 1999. Characterization of pHPM185F plasmid from *Helicobacter pylori*. Thesis. Texas Woman's University, Denton, TX.

28. Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R.Wilkins, R. D. Appel, and A. Bairoch. 2005. Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press pp. 571-607.

29. Gauthier, N. C., P. Monzo, V. Kaddai, A. Doye, R. Ricci, and P. Boquet. 2005. *Helicobacter pylori* VacA Cytotoxin: a probe for a clathrin-independent and Cdc42-dependent pinocytic pathway routed to late endosomes. Mol. Biol. Cell. 16:4852-4866.

30. Goodwin, C. S., J. A. Armstrong, T. Chilvers, M. Peters, D. M. Collins, L. Sly, W. McConnell, and W. E. S. Harper. 1989. Transfer of *Campylobacter pylori* and *Campylobacter mustelae* to *Helicobacter* gen. nov. *Helicobacter pylori* comb. nov. and *Helicobacter mustelae* comb. nov., respectively. Int. J. Syst. Bacteriol. 39:397-405.

31. Gruss A, D., H. F. Ross, and R. P. Novick. 1987. Functional analysis of a palindromic sequence required for normal replication of several staphylococcal plasmids. Proc. Natl. Acad. Sci. USA 84:2165-2169.

32. Gruss, A. D., and S. D. Ehrlich. 1989. The family of highly interrelated single-stranded deoxyribonucleic acid plasmids. Microbiol. Rev. 53:231-241.

33. Heurermann, D., and R. Haas. 1995. Genetic organization of a small cryptic plasmid of *Helicobacter pylori*. Gene 165:17-24.

34. Higashi, H., R. Tsutsumi, A. Fujita, S. Yamazaki, M. Asaka, T. Azuma, and M. Hatakeyama. 2002. Biological activity of the *Helicobacter pylori* virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. Proc. Nat. Acad. Sci. USA 99:14425-14433.

35. Hirokawa T., S. Boon-Chieng, and S. Mitaku. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14:378-9.

36. Hofreuter, D., and R. Haas. 2002. Characterization of two cryptic *Helicobacter pylori* plasmids: a putative source for horizontal gene transfer and gene shuffling. J. Bacteriol. 184:2755-2766.

37. Hosaka, Y., R. Okamoto, K. Irinoda, S. Kaieda, W. Koizumi, K. Saigenji, and M. Inoue. 2002. Characterization of pKU701, a 2.5 kb plasmid, in a Japanese *Helicobacter pylori* isolate. Plasmid 47:193-200.

38. Ilyina, T. V., and E. V. Koonin. 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by a diverse replicons from eubacteria, eukaryotes and archaebacteria. Nucleic Acids Res. 20:3279-3285.

39. Israel, D. A., N. Salama, U. Krishna, U. M. Rieger, J. C. Atherton, S. Falkow, and R. M. Peek, Jr. 2001. *Helicobacter pylori* genetic diversity within gastic niche of a single human host. Proc. Natl. Acad. Sci. USA 98:14625-14630.

40. Kalia, A., A. K. Mukhopadhyay, G. Dailide, Y. Ito, T. Azuma, B. C. Y. Wong, and D. E. Berg. 2004. Evolutionary dynamics of insertion sequences in *Helicobacter pylori*. J. Bacteriol. 186:7508-7520.

41. Karianne, T., S. M. Williams, L. Connolly, and K. M. Otteman. 2005. Chemotaxis plays multiple roles during *Helicobacter pylori* animal infection. Infect. Immun. 73:803-811.

42. Kersulyte, D., N. S. Akopyants, S. W. Clifton, B. A. Roe, and D. E. Berg. 1998. Novel sequence organization and insertion specificity of IS*605* and IS*606*: chimaeric transposable elements of *Helicobacter pylori*. Gene 223:175-186.

43. Kersulyte, D., H. Chalauskas, and D. E. Berg. 1999. Emergence of recombinant strains of *Helicobacter pylori* during human infection. Mol. Microbiol. 31:31-34.

44. Kersulyte, D., A. Kalia, M. Zhang, H. Lee, D. Subramaniam, L. Kiuduliene, H. Chalkauskas, and D. E. Berg. 2004. Sequence organization and insertion specificity of the novel chimeric IS*Hp609* transposable element of *Helicobacter pylori*. J. Bacteriol. 186:7521-7528.

45. Kersulyte, D., A. K. Mukhopadhyay, M. Shirai, T. Nakazawa, and D. E. Berg. 2000. Functional organization and insertion specificity of IS*607*, a chimeric element of *Helicobacter pylori*. J. Bacteriol. 182:5300-5308.

46. Kersulyte, D., A. K. Mukhopadhyay, B. Velapatino, W. Su, Z. Pan, C. Garcia, V. Hernandez, Y. Valdez, R. S. Mistry, R. H. Gilman, Y. Yuan, H. Gao, T. Alarcon, M. Lopez-Brea, G. B. Nair, A. Chowdhury, S. Datta, M. Shirai, T. Nakazawa, R. Ally, I. Segal, B. C. Y. Wong, S. K. Lam, F. O. Olfat, T. Boren, L. Engstrand, O. Torres, R. Schneider, J. E. Thomas, S. Czinn, and D. E. Berg. 2004. Differences in genotypes of *Helicobacter pylori* from different human populations. J. Bacteriol. 182:3210-3218.

47. Kersulyte, D., B. Velapatino, G. Dailide, A. K. Mukhopadhyay, Y. Ito, L. Cahuayme, A. J. Parkinson, R. H. Gilmann, and D. E. Berg. 2002. Transposable element IS*HP608* of *Helicobacter pylori*: nonrandom geographic distribution, functional organization, and insertion specificity. J. Bacteriol. 184:992-1002.

48. Khan, S. A. 1997. Rolling-circle replication of bacterial plasmids. Microbiol. Mol. Biol. Rev. 61:442-455.

49. Khan, S. A. 2000. Plasmid rolling-circle replication: recent developments. Mol. Microbiol. 37:477-484.

50. Khan. S. A. 2005. Plasmid rolling-circle replication: highlights of two decades of research. Plasmid 53:126-136.

51. Khatun, I. 2005. Analysis of plasmid DNA from AL226, an Alaskan Indian strain of *Helicobacter pylori*. Thesis. Texas Woman's University, Denton, TX.

52. Kleanthous, H., C. L. Clayton, and S. Tabaqchali. 1991. Characterization of a plasmid from *Helicobacter pylori* encoding a replication protein common to plasmids in Gram-positive bacteria. Mol. Micro. 5: 2377-2389.

53. Komano, T., R. Utsumi, and M. Kawamukai. 1991. Functional analysis of the *fic* gene involved in regulation of cell division. Res. Microbiol. 142:269-277.

54. Kusters, J. G., A. H. M. van Vliet, and E. J. Kuipers. 2006. Pathogenesis of *Helicobacter pylori* infection. J. Clin. Micro. 19:449-490.

55. Larkin M.A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. ClustalW and ClustalX version 2. Bioinformatics 23:2947-2948.

56. Lee, W-K., Y-S. An, K-H. Kim, S-H. Kim, J-Y. Song, B-K. Ryu, Y-J. Choi, Y-H. Yoon, S-C. Baik, K-H. Rhee, and M-J Cho. 1997. Construction of a *Helicobacter pylori-Escherichia coli* shuttle vector for gene transfer in *Helicobacter pylori*. Appl. Env. Microbiol. 63: 4866-4871.

57. Liu, H-H. 1998. Characterization of a 6 kbp *Eco*RI fragment of *Helicobacter pylori* plasmid pHPM185F. Thesis. Texas Woman's University, Denton, TX.

58. Marchler-Bauer, A., and S. H. Bryant. 2004. CD-search: protein domain annotations on the fly. Nucleic Acids Res. 32:327-331.

59. Marchler-Bauer, A., J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitez, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamshita, D. Zhang, N. Zhang, and S. H. Bryant. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37:205-10.

60. Marshall, B. J., and J. R. Warren. 1984. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet i:1311-1315.

61. Meijer, W. J. J., G. Venema, and S. Bron. 1995. Characterization of the single strand origins of cryptic rolling-circle plasmids from *Bacillus subtilis*. Nucleic Acids Res. 23:612-619.

62. Minnis, J. A., T. E. Taylor, J. E. Knesek, W. L. Peterson, and S. A. McIntire. 1995. Characterization of a 3.5-kbp plasmid from *Helicobacter pylori*. Plasmid 34:27-36.

63. Mitaku S., T. Hirokawa, and T. Tsuji. 2002. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. Bioinformatics 18: 608-616.

64. Mitaku, S. and T. Hirokawa. 1999. Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length. Protein Eng.12:953-957.

65. Nguyen, V. Q., R. M. Caprioli, and T. L. Cover. 2001. Carboxy-terminal proteolytic processing of *Helicobacter pylori* vacuolating toxin. Infect. Immun. 69:543-546.

66. Nilsson, C., A. Sillen, L. Eriksson, M. L. Strand, H. Enroth, S. Normark, P. Falk, and L. Engstrand. 2003. Correlation between *cag* Pathogenicity island composition and *Helicobacter pyori*-associated gastroduodenal disease. Infect. Immun. 71:6573-6581.

67. Novick, R. P. 1989. Staphylococcal plasmids and their replication. Annu. Rev. Microbiol. 43:537-565.

68. Oh, J. D., H. Kling-Backhed, M. Giannakis, J. Xu, R. S. Fulton, L. A. Fulton, H. S. Cordum, C. Wang, G. Elliott, J. Edwards, E. R. Mardis, L. G. Engstrand, and J. I. Gordon. 2006. The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. Proc. Natl. Acad. Sci. USA 103:9999-10004.

69. Oliveira, G. A., A. Santos, J. B. Guerra, G. A. Rocha, A. M. M. F. Nogueira, and D. M. M. Queiroz. 2003. *babA2*- and *cagA*-positive *Helicobacter pylori* strains are associated with duodenal ulcer and gastric carcinoma in Brazil. J. Clin. Micro. 41:3964-3966.

70. Paster, B. J., A. Lee, J. G. Fox, E. F. Dewhirst, L. A. Tordoff, G. J. Fraser, J. L. O'Rourke, N. S. Taylor, and R. Ferrero. 1991. Phylogeny of *Helicobacter felis* sp. nov., *Helicobacter mustelae*, and related bacteria. Int. J. Syst. Bacteriol. 41:31-38.

71. Qasem, J. A. A. 1995. Cloning and sequence analysis of plasmid DNA from *Helicobacter pylori*. Dissertation. Texas Woman's University, Denton, TX.

72. Quiñones, M. 2000. Characterization of a 7.8 kbp plasmid from *Helicobacter pylori* strain HPM8. Thesis. Texas Woman's Univeristy, Denton, TX.

73. Rickets, I. Y. 2004. Sequencing and characterization of *Helicobacter pylori* plasmid pAL202. Dissertation. Texas Woman's University, Denton, TX.

74. Schwarz, S., G. Morelli, B. Kusecek, A. Manica, F. Balloux, R. J. Owen, D. Y. Graham, S. van der Merwe, M. Achtman, and S. Suerbaum. 2008. Horizontal versus familial transmission of *Helicobacter pylori*. PLoS Pathog. 4:1-9.

75. Seegers, J. F. M. L., A. C. Zhao, S. A. Khan, L. E. Pearce, W. J. J. Meijer, G. Venema, and S. Bron. 1995. Structural and functional analysis of the single-strand origin of replication from the lactococcal plasmid pWVO1. Mol. Gen. Genet. 249:43-50.

76. Selbach, M., S. Moese, R. Hurwitz, C. R. Hauck, T. F. Meyer, and S. Backert. 2002. The *Helicobacter pylori* CagA protein induces cortactin dephosphorylation and actin rearrangement by c-Src inactivation. EMBO 22:515-528.

77. Song, J-Y., S-H. Choi, E-Y. Byun, S-G. Lee, Y-H. Park, S-G. Park, S-K. Lee, K-M. Kim, J-U. Park, H-L. Kang, S-C. Baik, W-K. Lee, M-J. Cho, H-S. Youn, G-H. Ko, D-W. Bae, and K-H. Rhee. 2003. Characterization of a small cryptic plasmid, pHP51, from a Korean isolate of strain 51 of *Helicobacter pylori*. Plasmid 50:145-151.

78. Song, J-Y., S-G. Park, H-L. Kang, W-K. Lee, M-J. Cho, J-U. Park, S-C. Baik, W-K. Youn, G-H. Ko, and K-H. Rhee. 2003. pHP489, a *Helicobacter pylori* small cryptic plasmid, harbors a novel gene coding for a replication initiation protein. Plasmid 50:236-241.

79. Suerbaum, S., J. M. Smith, K. Bapumia, G. Morelli, N. H. Smith, E. Kunstmann, I. Dyrek, and M. Achtman. 1998. Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA 95:12619-12624.

80. Taylor, T. E. 1997. Unusual Structure of a *Helicobacter pylori* plasmid. Thesis. Texas Woman's University, Denton, TX.

81. Terebiznik, M. R., C. L. Vazquez, K. Torbicki, D. Banks, T. Wang, W. Hong, S. R. Blanke, M. I. Colombo, and N. L. Jones. 2006. *Helicobacter pylori* VacA toxin promotes bacterial intracellular survival in gastric epithelial cells. Infect Immun. 74:6599-6614.

82. Tomb, J-F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, E. K. Hickey, D. E. Berg, J. D. Gocayne, T. R. Utterback, J. D. Peterson, J. M. Kelley, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W. S. Hayes, M. Borodovsky, P. D. Karp, H. O. Smith, C. M. Fraser, and J. C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539-547.

83. Torres, J., G. Perez-Perez, K. J. Goodman, J. C. Atherton, B. D. Gold, P. R. Harris, A. Madrazo-de la Garza, J. Guarner, and O. Munoz. 2000. A comprehensive review of the natural history of *Helicobacter pylori* infection in children. Arch. Med. Res. 31:431-469.

84. Tusnady, G. E., and I. Simon. 1998. Principles governing amino acid composition of integral membrane proteins: applications to topology prediction. J. Mol. Biol. 283:489-506.

85. Tusnady, G. E., and I. Simon. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849-850.

86. Utsumi, R., Y. Nakamoto, M. Kawamukai, M. Himeno, and T. Kamano. 1982. Involvement of cyclic AMP and its receptor protein in filamentation of an *Escherichia coli fic* mutant. J. Bacteriol. 151:807-812.

87. Vandamme, P., E. Falsen, R. Rossau, B. Hoste, P. Segers, R. Tytgat, and J. De Ley. 1991. Revision of *Campylobacter*, *Helicobacter*, and *Wolinella* taxonomy emendation of generic descriptions and proposal of *Arcobacter* gen. nov. Int. J. Syst. Bacteriol. 41:88-103.

88. Veereshlingam, L. 2003. Interaction of pHPM encoded RepA protein from *Helicobacter pylori* with pHPM iteron sequence. Thesis. Texas Woman's University, Denton, TX.

89. Versalovic, J., and J. G. Fox. 2001. Taxonomy and Phylogeny of *Helicobacter pylori*. Helicobacter pylori: Molecular and Cellular Biology. United Kingdom: Horizon Scientific Press.

90. Warren, J. R., and B. J. Marshall. 1983. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet i:1273-1275.

91. Waterhouse, A.M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. 2009. Jalview version 2: A multiple sequence alignment and analysis workbench. Bioinformatics 25:1189-1191.

92. Zaman, S., L. Radnedge, H. Richards, and J. M. Ward. 1993. Analysis of the site for second-strand initiation during replication of the *Streptomyces* plasmid pIJ101. J. Gen. Microbiol. 139:669-676.