Gene essentiality, miniature inverted-repeat transposable elements and pigmentation in *Porphyromonas gingivalis*: Convergence at a unique glycosyltransferase

A thesis

submitted by

Brian A. Klein

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Molecular Microbiology

TUFTS UNIVERSITY

Sackler School of Graduate Biomedical Sciences

August, 2015

Advisor: Linden T. Hu

ACKNOWLEDGEMENTS

First and foremost I would like to thank my thesis advisor, Linden. He allowed me to join his laboratory, start a project from the ground floor and follow my ideas throughout all five years in his laboratory. The free-reign and ownership that he allowed me to have over the project helped me to develop as a student, researcher, thinker and collaborator.

I would like to thank my thesis committee, Dr. Andrew Camilli, Dr. Carol Kumamoto, Dr. Michael Malamy and Dr. Margaret Duncan for their guidance and support throughout my graduate career. Their expertise, mentorship and collaboration has greatly furthered my skill set, nurtured my abilities and prepared me to continue on in my science career. Dr. Malamy frequently helped hone or flesh out ideas throughout the development of my project; a level of mentorship not frequently gained for a Ph.D. student from their own PI, let alone a PI of another laboratory. Additionally, I would like to thank Dr. Andrew Goodman (Yale University School of Medicine) for providing the mutagenesis strains and plasmids and his help with technical aspects of mutant library construction semi-random PCR. Furthermore, I'd like to thank Dr. Rachel Dutton (UCSD) for serving as my outside examiner for my thesis defense.

For getting me to my graduate position I must thank Dr. Pam Baker and Dr. Lee Abrahamsen from Bates College (Lewiston, ME). Drs. Baker and Lee first introduced me to 'microbes', teaching the courses microbiology, immunology and virology during college. Dr. Baker gave me my first laboratory research positions involving microbiology. Dr. Baker then allowed me to complete my undergraduate thesis on a combination of topics that I'm very interested in; oral bacteria and tea.

I would like to thank my family for being supportive of my decision to go to graduate school and for their support and help throughout the six-year process. A special thank you to the soon-to-be newest part of my family, my fiancé Ricki, for her support (and I think interest) of my science and 'knurd' life. Ricki's family, soon-to-be my family too, also deserves a shout out for their support, usually in the form of food, pep talks or 'just because' presents.

To my friends within and the activities of rowing, running, triathlon and November Project, you have kept me happy, engaged, sharp and in shape throughout my time in Boston/Cambridge. Having always believed that the mind cannot be sound without the body, activities outside science, even though they frequently included top tier science buddies, were integral to my development and progress.

To the Hu laboratory, a great group of people who helped make 'work' a place that I enjoyed to come to. Thank you for your help with academic and life questions.

To Tatte, a bakery and café in Boston/Cambridge. Positive outcomes have been associated with thinking and writing processes in coffee shop environments as opposed to libraries or one's usual workplace setting. And importantly, the coffee, tea and pastries that come from your kitchen helped to fuel the mind for thinking and body for writing of this thesis. Additionally, given that tea and *Porphyromonas* brought me to graduate school, I believe that it is fitting that they continue to meet on a daily basis in the mouths of people across the globe.

ABSTRACT

Periodontal disease is a bacterially-induced ailment which results in inflammation, destruction of oral structural integrity, loss of teeth and systemic comorbidities. It is a polymicrobial disease in which no single bacterial species is sufficient or necessary to elicit disease. Yet several periodontopathic species such as *Porphyromonas gingivalis* have been associated with disease occurrence and progression.

P. gingivalis is a Gram-negative anaerobic bacterial species with a subgingival oral cavity niche. *P. gingivalis* elaborates numerous virulence factors that aid in its ability to colonize, evade immune system clearance and cause disease. The most recognizable phenotypic feature, which doubles as an important virulence factor, is black colony pigmentation. Pigmentation of *P. gingivalis* protects against phagocytosis, UV damage and oxidative damage, and non-pigmented mutants have been shown to be avirulent in animal infection models. Another important virulence feature, on a molecular level as opposed to macroscopic colony pigmentation, is lipopolysaccharide. *P. gingivalis* can display at least five different lipid-A versions as well as two distinct O-antigen moieties of lipopolysaccharide, which carry different immunomodulatory capabilities. A third virulence feature, on a genetic and genomic level, is a diverse repertoire of mobile and repetitive elements. Intra-strain chromosomal rearrangement and transcriptional modulation have been

demonstrated for mobile and repetitive elements in *P. gingivalis*, and the potential of interstrain and interspecies genetic transfer allows for clone and population-level adaptation.

We chose to further investigate the virulence factors of *P. gingivalis* by adapting and developing transposon mutagenesis using a transposon system and coupling it with sequencing technology for *P. gingivalis*. After generating the transposon mutant library we characterized the *in vitro* essential genes of *P. gingivalis* and began investigating colony pigmentation and tetrapyrrole metabolism. Through screening the library for colony pigmentation defects and selecting the library under variable nutrient conditions we identified a novel miniature inverted-repeat transposable element that has connections to pigmentation and haem metabolism as well as a lipopolysaccharide-affecting gene that is necessary for proper colony pigmentation.

Understanding the pathways and mechanisms relating to *P. gingivalis* colony pigmentation, as well as the genetic elements capable of modulating the genomic and transcriptomic landscape, will allow for better understanding of a keystone periodontopathic species on a basic genetic and metabolic level along with potential for targeted clinical applications.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LPS	Lipopolysaccharide
O-LPS	O-Antigen lipopolysaccharide
A-LPS	Acidic lipopolysaccharide
CPS	Capsular polysaccharide
COG	Cluster of orthologous groups
DEG	Database of essential genes
Tn-seq	Transposon sequencing
INseq	Insertion sequencing
Amp	Ampicillin
Erm	Erythromycin
DNA	Deoxyribonucleic acid
gDNA	Genomic deoxyribonucleic acid
RNA	Ribonucleic acid
tRNA	Transfer ribonucleic acid
rRNA	Ribosomal ribonucleic acid
CDS	Coding sequence
LB	Luria broth
PBS	Phosphate buffered saline
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
ARB	Arbitrary binding primer
TdT	Terminal deoxynucleotidyl transferase

Pg	*Porphyromonas gingivalis*
Bt	*Bacteroides thetaiotaomicron*
BHI	Brain-heart infusion
BAP	Blood agar plate
RBC	Red blood cell
BROP	Bioinformatics resource oral pathogens
MTD	Microbial transcriptome database
BLAST	Basic local alignment search tool
KEGG	Kyoto encyclopedia of genes and genomes
RE	Repetitive element
TE	Transposable element
TIR	Terminal inverted repeat
MITE	Miniature inverted-repeat transposable element
X-gal	5-bromo-4-chloro-3-indolyl-beta-D-galacto-pyranoside
ONPG	O-Nitrophenyl-β-D-galactopyranoside
CRISPR	Clustered regularly interspaced short palindromic repeats

Chapter 1: Introduction

1.1 Periodontal disease

Periodontal diseases are among the most common and costly diseases throughout the world, affecting people of all socio-economic strata, races, and geographic locations (Pihlstrom, Michalowicz, & Johnson, 2005). The progression and classification of periodontal diseases begins with gingivitis, which is a mild local inflammation of the gingiva, and progresses from early periodontitis to severe periodontitis, described as major gingival inflammation, destruction of the periodontal ligaments, loss of alveolar bone and exfoliation of teeth (Curtis, Zenobia, & Darveau, 2011). General oral care through brushing and flossing of teeth as well as the use of oral care rinses can help prevent the onset of periodontal disease. However, once periodontal disease has begun to progress, treatment typically involves frequent professional cleanings, use of prescription-strength oral rinses, root plane scaling and administration of local and systemic antibiotics. With professional treatment for periodontitis patients may halt progression or have some level of disease regression; however a proportion of individuals will present with refractory periodontitis, which describes individuals who do not respond to many forms of treatment. Periodontitis has been connected to systemic malignancies such as arthritis, pre-term birth, and cardiovascular disease (Darveau, 2010). When considering the direct effects of periodontitis alone the costs associated with treatment in the United States are several billions of dollars per year. The effects of periodontitis connected to the direct tissue destruction and systemic comorbidities can significantly effect disability-adjusted life years, a World Health Organization measure of disease burden expressing the estimated number of years lost from ones' life due to poor health, general disability and early death. Current materials research involving nanoparticles, microparticles and amalgams, as well as microbial research

involving oral ecology and targeted virulence inhibition may lead to novel and potentially less invasive therapies to combat oral disease burden.

1.2 Oral microbiota and microbial ecology

Prior to sequencing studies conducted between the years 2000-2014 the oral microbiota was only thought to contain between 150-250 different species. These species composition estimates were generally determined through experiments based on the ability to culture species *in vitro* and distinguish them based on a limited number of diagnostic phenotypes. However, through additional comparative 16S rRNA sequencing and high-throughput amplicon and metagenomic sequencing using Illumina and 454 pyrosequencing platforms the oral microbiota is now thought to contain an average of 700-1000 distinct species (Aas, Paster, Stokes, Olsen, & Dewhirst, 2005) (Lazarevic et al., 2009). Of the average 700-1000 species known to inhabit human oral cavities, single individuals are estimated to carry between 300-1200. It appears that many individuals share a core species set even when total diversity varies (Hajishengallis, 2015).

Several distinct niche environments have been identified within the oral cavity based on host physiological architecture and environmental parameters. Host and environmental constraints such as space, tissue type, nutrient composition and oxygen levels, as well as the other microbial species present for interactions, cause bacterial species to differentially occupy these niches (Hajishengallis, 2015). The major niches within the oral cavity are supragingival, subgingival, tongue, cheeks, saliva and teeth.

Of the current estimates of oral microbiome species diversity only about sixty percent of the species can be cultivated by common *in vitro* methods. This lack of ability in identifying species present and difficulty of cultivation hinders the understanding of oral

microbiota composition and influence. However, recent advances in co-culture techniques as well as metagenomics may allow for isolation and experimentation on additional species within the oral microbiome through the identification of factors (ie. nutrients or metabolites) necessary for growth.

There is a growing recognition that bacteria within microbial communities communicate with each other and affect growth and composition of the community. In other environments, both ecological and host specific, it is known that commensal microorganisms can either help, have no effect, or hinder the existence of other microorganisms in that same environment (Fig.1-1). Adding further complexity, the type of interaction may vary between two species based on environmental conditions. Effector microorganisms are known to influence several factors such as overall abundance, biofilm formulation, secretion of toxins and expression of outer-membrane adhesion molecules. Microorganisms exerting an effect on resident species, whether commensal or pathogen, do not necessarily need to be endogenous residents of the existing microflora; they can in some instances be transient microorganisms due to a spatial or temporal niche restriction. The oral cavity is essentially one of the most open ecological niches to transient passage of microorganisms. As such, over the course of one's life, eating and breathing alone will allow a significant sampling of Earth's bacterial consortium to transit through. Thus, relative stability of the diversity and metabolic composition in the oral microbiome is intriguing and potentially a good model for studying perturbation of host-associated microbial niches by biotic and abiotic environmental factors.

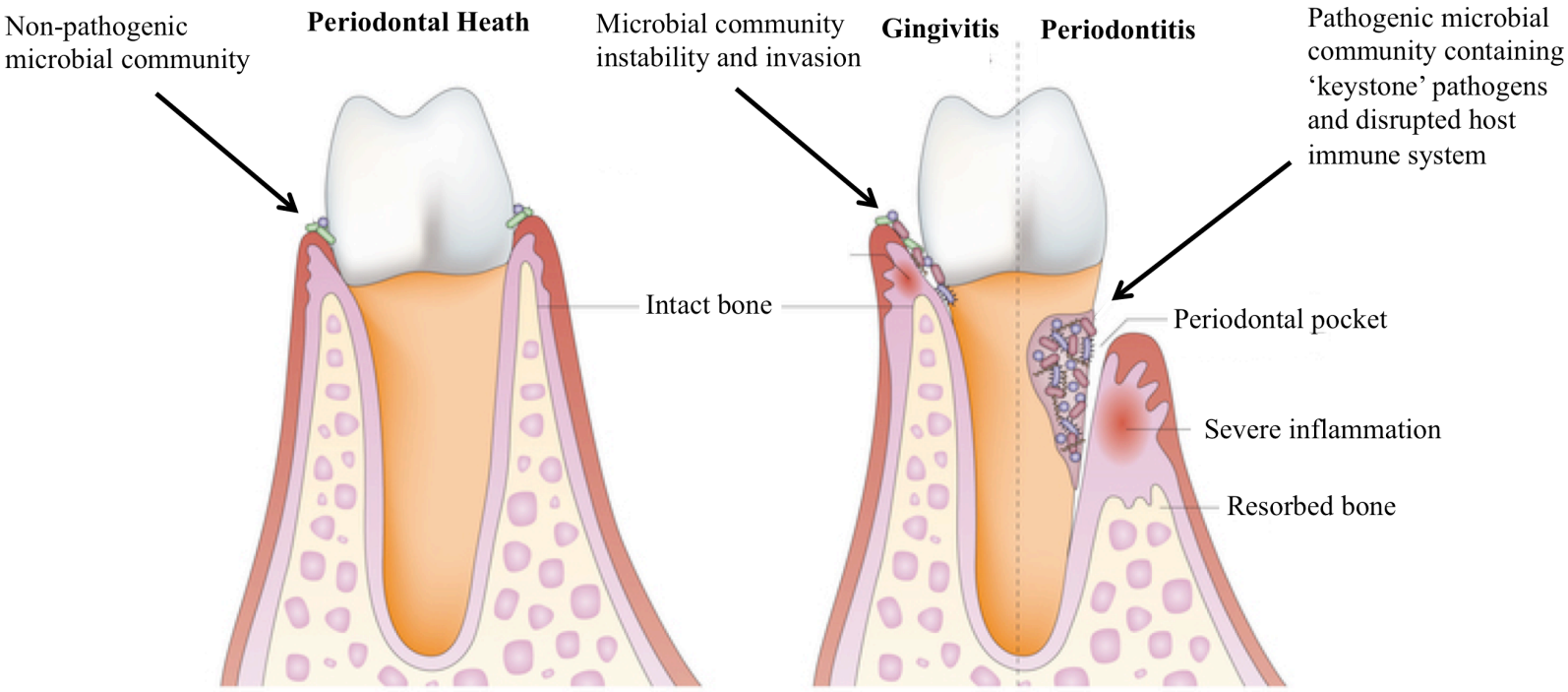

Non-pathogenic microbial community
Periodontal Heath
Microbial community instability and invasion
Gingivitis
Periodontitis
Pathogenic microbial community containing 'keystone' pathogens and disrupted host immune system
Intact bone
Periodontal pocket
Severe inflammation
Resorbed bone

Figure 1-1. Periodontal Disease Pathology.
Model adapted from Hajishengallis,G. **'**Periodontitis: from microbial immune subversion to systemic inflammation' Nat.Rev.Immunol., 2015, 15, 1, 30-44. During periodontal health (left half of model) there is no overt disease; no subgingival pockets, no bleeding, bone surrounding the tooth and no inflammation. During early periodontitis or gingivitis (left half of right part of image) inflammation is apparent and microbial community species composition changes to include more Gram-negative anaerobes. If periodontitis occurs (right half of right part of image) severe inflammation is apparent, bone resorbs and subgingival microbial communities have periodontopathic species within them (Hajishengallis, 2015). The keystone pathogen in the case periodontal disease, *Porphyromonas gingivalis*, serves as a source of nutrient acquisition and immune protection for the microbial community. Thus earning its title of 'keystone', which refers to its ability to hold a structure (the community) together, doing so disproportionately to its abundance. [Permission for use and modification of figure granted by author Dr. George Hajishengallis and the publisher Nature Publishing Group]

Over a dozen bacterial species and several environmental and host genetic factors have been associated with the manifestation and progression of periodontal diseases (Pihlstrom et al., 2005). Periodontal diseases are thought to occur when a shift in an individual's oral microbiota takes place, which can happen due to poor oral hygiene, changes in diet, oral injury, oral surgery, antibiotic use or host-genetics affecting tissue architecture and salivary flow rates. The shift from a commensal-dominated microflora species composition to that of bacterial consortium highly correlated with periodontitis that includes *Porphyromonas gingivalis*, *Aggregatibacter actinomycetemcomitans*, *Fusobacterium nucleatum*, *Tannerella forsythia* and *Treponema denticola* generally progress over a timeframe of a several years. In support of a bacterial progression being necessary for periodontitis, few clinical cases of periodontal disease are observed in children and adolescents, who harbor few if any of the above periodontopathic species, while prevalence and abundance of these species are higher in adult and elderly populations and periodontitis rates are significantly greater. Periodontopathic species such as *P. gingivalis* are frequently isolated or identified by molecular means from human patients with oral disease, yet are rarely isolated from individuals without overt disease.

1.3 *Porphyromonas spp*. and *Porphyromonas gingivalis*

Porphyromonas spp. are members of the *Bacteroidetes* phylum. *Porphyromonas* spp. are most commonly associated with the oral cavity of humans and other animals, however, several *Porphyromonas* spp. can be found throughout the gastrointestinal tract and are occasionally abscess or wound-associated. The genus *Porphyromonas* is comprised of fifteen

species that are anaerobic, Gram-negative, non-spore forming rod or coccobacillus-shaped cells. Of these fifteen species, thirteen canonically display black colony pigmentation on blood agar, while one is variable and another is comprised of solely non-pigmenting strains (Sakamoto, 2013). In the absence of sequencing techniques, sugar fermentation and protease activities are two of the most common used diagnostic phenotypes to differentiate between the phenotypically similar species.

Porphyromonas gingivalis, first described as a distinct species in 1981 and originally names *Bacteroides gingivalis*, was initially isolated from the human oral cavity of patients with severe periodontal disease. *P. gingivalis* has since been demonstrated to be involved in the etiology of periodontal disease (Holt, Ebersole, Felton, Brunsvold, & Kornman, 1988). Human clinical studies using both diagnostic and molecular genetic techniques have also tied the presence, prevalence and 'isolatability' (the viable but non-culturable status) of *P. gingivalis* to increased periodontal disease risk and manifestation.

Instillation of *P. gingivalis* in a number of non-human animal models has demonstrated that *P. gingivalis*, when added to the oral cavity of fully-colonized animals, can induce periodontal disease (Holt et al., 1988). A murine study in 2011 demonstrated that *P. gingivalis* utilizes its host's native microbial flora and parts of the host immune system to generate a disease state; doing so while maintaining itself at low bacterial burden levels (Hajishengallis et al., 2011). *P. gingivalis* itself does not actually increase dramatically in number during disease progression as seen with pathogens such as *V. cholerae* and *Y. pestis*; the bacteria generate an environment where other oral species increase in number and virulence most likely due to increased nutritional sources and a compromised host immune system, which is due to the virulence factors of *P. gingivalis* (*e.*g. gingipain proteases and

variable LPS). Interestingly, inoculation of animal models lacking basal microbial flora do not develop the majority of periodontal disease characteristics, which demonstrates the necessity of other microbial species for colonization and pathogenesis. Studies to determine the exact species required for *P. gingivalis* pathogenesis have yet to be carried out in animal models. A potential confounding factor to such studies would be if specific metabolic and structural characteristics, not exact species, were the necessary factors. Thus, metagenomic approaches may prove useful in determining what genetic factors are necessary for virulent community compositions.

1.4 *Porphyromonas gingivalis* Virulence

Due to its overt association with disease states and different phenotypic presentation in comparison to other oral species, many researchers have attempted to identify and characterize the virulence factors of *P. gingivalis* with hope of 'disarming' the pathogen. Various *in vitro* and *in vivo* models associated with periodontal disease such as RBC hemolysis, haemagglutination, host cell killing, host cell stimulation and host immune evasion have been applied. Confounding progress with these and other assays is that *P. gingivalis* has proven difficult to work with genetically, generally attributed to restriction enzyme systems, atypical promoter structures, aerobiosis limitations and a lack of endogenous plasmids, which has slowed genetic manipulations and high-throughput screening applications of the species. However, in spite of these obstacles several virulence factors have been identified including capsule and lipopolysaccharide variability, toxin production, colony pigmentation, multiple fimbriae, ability to invade host cells, volatile sulfur compounds and several proteases that can modulate the immune system and cellular structure of the host (Holt, Kesavalu, Walker, & Genco, 1999)(Lamont & Jenkinson,

1998)(Nakayama, 2003). Further identification of virulence factors as well as understanding their expression and regulation can be aided by functional genomics.

1.5 *Porphyromonas gingivalis* Genetics and Genomics

Several distinct strains of *P. gingivalis* rose to prominence following clinical isolation and initially became pervasive throughout the field of study; ATCC 33277, FDC-381, W83, W50, A7A1-28, TDC60 and HG66. Of these, the genomes of strains ATCC 33277, W83, TDC60, and HG66 are now completed genome projects (Nelson et al., 2003) (Naito et al., 2008)(Watanabe et al., 2011)(Siddiqui et al., 2014). In addition, strains JCVI-SC001, SJD2 and W50 are sequenced and available as scaffolds, and six other strains are sequenced and available as multiple contigs (McLean et al., 2013) (Liu et al., 2014). Strains FDC-381, W50, A7A1-28 and HG66 fell out of favor due to a combination of issues with genetic tractability, phenotypic abnormalities, genetics and virulence attributes. As such, strains ATCC 33277 and W83 are now the most commonly used laboratory strains in the United States and Europe, while TDC60 accompanies these strains in much of the work done in Asian laboratories.

The five completed genome sequencing and assembly projects wild-type strains are disparate based on origin or lineage: W83 isolated in Germany during the 1950's from an oral lesion; ATCC 33277 was isolated in the USA during the 1980's from subgingival plaque; TDC60 was isolated in Japan in 2011 from an oral lesion; HG66 isolated in the USA in 1989 from a dental school patient; and JCVI SC001 isolated in the USA in 2013 from a hospital sink. Each of these sequencing projects utilized different sequencing and assembly methods, spanning Sanger, 454, Illumina, PacBio and cosmid sequencing and involving various gene annotation pipelines. Importantly, each project provided a *de novo* assembly.

Several phenotypic differences relating to capsule, fimbriae, protease expression, colony pigmentation, nutrient requirements and virulence in animals separate the common *P. gingivalis* laboratory strains. ATCC 33277 is noted as not expressing capsule, whereas all others express capsule but of varying serotypes. W83 is noted as only expressing minor fimbriae, which is in contrast to moderate expression of both major (*fimA*) and minor (*mfa1*) fimbriae from ATCC 33277 and relative overexpression of FDC-381. TDC60 and W83 are frequently referred to as 'more virulent' than ATCC 33277, yet virulence in this case refers to lethality in animal models involving non-oral infections. Pertaining to colony pigmentation and protease expression, strain HG66 displays low, abnormal pigmentation with respect to all other laboratory strains and is suggested to secrete all gingipain proteases into culture supernatant as opposed to retaining certain forms of gingipains on the cell surface as other strains do. The greatest variability between strains may be the transposable element content and the genomic locations of these elements; mainly the Insertion Sequence *Porphyromonas gingivalis* (*ISPg*) and Composite Transposon (*CTn*). The number of each *ISPg* element (*ISPg1-ISPg9*) differ between strains and transpositions and recombinations have reorganized genomes (Chen et al., 2004).

1.6 Porphyromonas gingivalis Transposon Mutagenesis

In 1995 the first transposon mutagenesis system utilizing *Tn4351* was developed for use with *P. gingivalis* (Hoover, Abarbarchuk, Ng, & Felton, 1992; Hoover & Yoshimura, 1994). In 2000 a second transposon mutagenesis system for use in *P. gingivalis* was developed, this system utilizing *Tn4400* (Chen et al., 2000). Both the *Tn4351* and *Tn4400* systems relied on cryptic *Bacteroides* plasmids due to a lack of native plasmids in *Porphyromonas* species and the genetic relatedness and tractability of *Bacteroides*. The only

reported use of the *Tn4351* system studied colony pigmentation. However, the *Tn4400* system has been used to study pigmentation, lipopolysaccharide, biofilm formation, oxygen tolerance and antimicrobial susceptibility. While providing important advances in the field, limitations of the *Tn4400* system hindered its broader application.

In 2012 our laboratory adapted a Mariner transposon based system for mutagenesis of *P. gingivalis* (Klein et al., 2012). For other bacterial species multiple laboratories had demonstrated the ability and utility of Mariner transposon based mutagenesis systems in creating saturated sequencing-adapted mutant libraries. These systems and their resultant libraries have led to studies on conditional gene essentiality for a given species, relative fitness assessment of genes under different *in vitro* and *in vivo* growth conditions as well as antimicrobial resistance determinant identification. For *P. gingivalis*, several advantages are afforded by the Mariner-based system, including, use in any *P. gingivalis* strain background, bi-parental as opposed to tri-parental mating, more insertions throughout the genome, a smaller transposon unit inserted into genome, lower risk of secondary transposition or recombination, fewer cryptic or double insertions and a known yet not limiting insertion site preference of the mutagenizing transposon.

Although modified Mainer as well as *Tn5* based transposon systems allowed for saturated transposon mutant libraries to be generated in many species, methods for adapting these mutant libraries to high-throughput screens and selections needed to be developed and employed to utilize the full potential of the mutant pools. Three similar yet independently developed techniques were created for this purpose; HITS, INseq and Tn-seq (Gawronski, Wong, Giannoukos, Ward, & Akerley, 2009)(Goodman et al., 2009)(van Opijnen, Bodi, & Camilli, 2009). Tn-seq, which stands for ‘transposon sequencing’, provides significant

advantages over traditional mutant library screen and selection methods in that it utilizes a more sensitive sequencing system and facilitates the identification of mutants with partial or subtle defects. Above all improved attributes, networks of genetic and molecular interactions can be constructed through the identification of mutant 'fitness' *en masse*. Fitness phenotypes of a given mutant, which can then be traced back to the insertion location in the genome following sequencing, provides phenotype-to-genotype identification that can be applied to both positive and negative selection conditions. Selections can be performed *in vitro* and *in vivo*, which adds to the utility of Tn-seq in identification of physiologically-relevant phenotypes.

1.7 Gene Essentiality

The genetic composition of bacterial genomes can be categorized into accessory, core and essential elements. The use of such categorizations typically refers to *in vitro*, laboratory growth media conditions, thus 'conditionally essential' terminology can arise when dealing with variations from 'normal' environmental conditions as well as when *in vivo* using cell culture or animal models of infection. Essential genes and regions cannot be disrupted and are usually shared among all strains of a species. Core genes are shared among the majority of strains within a species and encompass the essential genes of a species. In addition to the essential gene set, core genes generally include pathways necessary for successful metabolic and physiologic adaptation to environmental factors. Accessory genes are often strain-specific and are frequently associated with virulence or niche specificity. Examples of accessory genes include insertion sequence or transposable elements that carry antibiotic resistance genes, pathogenicity islands and toxins.

Gene essentiality can be determined through several methods that may combine *in vitro* and *in silico* systems. The 'gold standard' for defining gene essentiality is generating clean deletions of a single gene under a defined *in vitro* condition. This has been accomplished for only three organisms to date, and a small false positive rate has now been identified (De Berardinis et al., 2008) (Kobayashi et al., 2003) (Yamamoto et al., 2009). The majority of gene essentiality studies employ transposon mutagenesis followed by high-throughput sequencing and identify genes lacking insertions. *In silico* analyses require seed information from previous *in vitro* studies to generate a model and make predictions.

Brunner et al. described a core genome for *P. gingivalis* in 2010 (Brunner et al., 2010). Our laboratory added to the understanding of accessory, core and essential genes in *P. gingivalis* by describing the putative essential gene set for *P. gingivalis* strain ATCC 33277 in 2012 (Klein et al., 2012). The core genome analysis study utilized a DNA microarray hybridization method for determining presence, absence or aberrance at a given locus. Our gene essentiality analysis utilized the Illumina sequencing platform to perform Tn-seq. A re-analysis and comparison of methods for core genome determination could now be carried out using the completed genome sequencing and assembly projects of the aforementioned strains.

By determining the genes groupings and mechanisms involved in various systems, understanding the basic biological workings of *P. gingivalis* as well as progress toward identifying molecular targets for clinical application may be made.

1.8 Transposons and Repetitive Elements

Numerous transposable and repetitive elements have been identified in the *P. gingivalis* genomes. With respect to transposable elements, *CTnPg*, *TnPg17*, *ISPg1* through *ISPg11* as well as four distinct miniature inverted-repeat transposable elements have been

identified in at least one, but usually to some degree in all strains of *P. gingivalis*. Concerning repetitive elements, single nucleotide tracts as well as repeats ranging from 3-41 nucleotides have been found within *P. gingivalis* genomes. For pathogenic species that inhabit host-associated niches the presence of multiple types of numerous copies of transposable and receptive elements is not uncommon. However, the lack of endogenous plasmids and bacteriophage for *P. gingivalis* in light of the substantial foreign genomic content is puzzling.

Transposable Elements (TEs) are non-fixed DNA sequences that can change locus or multiply and insert into new loci within a genome or between genomes via excision, potentially replication, and insertion. TEs can insert into chromosomes, plasmids and bacteriophages. Class I TEs are retrotransposons, which require reverse-transcriptase activity to transpose. Class II TEs are DNA transposons, which require a transposase or a replicase to transpose.

Repetitive Elements (REs) are DNA sequences present in multiple copies throughout a genome, chromosome or vector. Such elements can be classified into 'terminal', 'tandem' and 'interspersed' repeats, yet, each broad classification encompasses several sub-types as well. Repeats are classified as either identical or non-identical based on nucleic acid homology. They can then be further classified as either micro, mini or macro satellites based on size of the repeat. Of note, repetitive elements can either be localized at a single site where a motif is sequentially recurrent, or at many loci as reiterations.

Transposable and repetitive elements are frequently associated with virulence, whether through carrying drug resistance, serving as modulators of gene regulation or causing genomic rearrangements. In *P. gingivalis* TEs have been shown to modulate

expression of gingipain proteases, to facilitate large chromosomal rearrangements and produce functional conjugation systems (Lewis & Macrina, 1998) (M. Naito et al., 2011).

The variability of transposable and repetitive element content between species, as well as the functions of the various elements that comprise such a large percentage of the genome require further attention in order to understand their biological and evolutionary roles.

1.9 Pigmentation, Iron, Heme and Gingipains

P. gingivalis was first recognized as a periodontal pathogen by its pigmentation on blood containing media; when cultures were taken from patients with periodontal disease black-pigmented anaerobic bacteria could be isolated. This black-pigmenting species was rarely isolated from healthy patients.

Pigmentation has been categorized as a virulence factor of pathogens such as *Cryptococus neoformans*, *Staphylococcus aureus*, *Vibrio* spp., *Plasmodium* spp., *Pseudomonas aeruginosa*, *Serratia* spp. and *Yersinia pestis* (Liu & Nizet, 2009). The pigment of oral anaerobes, some of which generate black pigments, has been suggested to aid against DNA and oxidative damage, phagocytosis and antimicrobial interactions.

Genes involved in pigmentation of numerous bacteria including *P. gingivalis* have been shown to play important roles in virulence. For example, the staphyloxanthin golden pigment of *S. aureus* is capable of detoxifying reactive oxygen species (Liu & Nizet, 2009). The melanin (black) pigment of *C. neoformans* protects against oxidative stress, blocks phagocytosis and inhibits multiple antimicrobials (Liu & Nizet, 2009). Green and brown pigments from *P. aeruginosa* are involved in colonization and immunomodulation of the host during lung infections (Lau, Ran, Kong, Hassett, & Mavrodi, 2004).

Pigmentation was first demonstrated to be a key virulence property of *P. gingivalis* following the discovery that spontaneous non-pigmented *P. gingivalis* mutants showed a marked decrease in mouse models of lethality as well as lower hemolytic, haemagglutination, and proteolytic activities (McKee, McDermid, Wait, Baskerville, & Marsh, 1988). These initial spontaneous mutants, brown and beige coloured as opposed to black, were isolated during extended chemostat growth experiments. Other pigmentation mutants were then isolated using the aforementioned *Tn4351* and *Tn4400* transposon mutagenesis systems and by targeted mutagenesis. Of note, the loci, singular or multiple, involved in the chemostat experiments were never determined.

Pigmentation of *P. gingivalis* is due to the binding, uptake and metabolism of hemin that results in μ-oxo bis-heme molecules displayed on the bacterial cell surface (Smalley, Silver, Marsh, & Birss, 1998). The lysine-gingipain (Kgp) was the first *P. gingivalis* gene confirmed to be necessary for pigmentation (Chen, Dong, Yong, & Duncan, 2000). Kgp liberates hemin from iron-containing compounds via its proteolytic activity as well as binds and transports hemin. Other genes associated with pigmentation in *P. gingivalis* that have also been found to serve as virulence determinants include the proteases RgpA and RgpB, glycosyltransferases involved in protease and lipopolysaccharide biogenesis, and the Por Secretion System (Sato, 2011).

Iron is an essential element for all eukaryotes and most prokaryotes. Very little free iron is found within eukaryotic host cells or fluids; iron is frequently bound or chelated by proteins that have the specific function of binding iron either for sequestration, transfer or to limit toxicity within the organism (Lewis, 2010). When considering bacterial pathogens of mammals, iron must be accumulated passively or actively taken from the host, since free iron

is tightly restricted in mammals as an antimicrobial strategy. All bacterial species, commensal and pathogenic, have systems for iron acquisition such as hemoproteins, hemolysins, siderophores and proteases that function to degrade host-associated iron.

Heme, a porphyrin ring tetrapyrrole structure that forms functional prosthetic groups of proteins involved in a myriad of molecular processes, amounts to a significant percentage of iron-associated complexes. Heme molecules contain coordinated iron; the heme structure lacking a coordinated iron is protoporphryin IX, and hemin compounds are the chloride-modified version frequently used as a growth medium supplement. Cytochromes, catalyases and peroxidases are usually the majority of heme-containing proteins within a cell/organism, however, other proteins involved in translation, transcriptional regulation and secondary metabolite biosynthesis (including pigments) harbor heme as well.

Iron accumulation and heme accumulation or biosynthesis are important to the physiology, metabolism and regulatory systems of bacteria. A defect in the ability to accumulate or store iron, as well as a lack of heme uptake or biosynthesis, can halt cellular growth due to energy production defects and non-functional metalloproteins that are required for numerous cellular processes. Uptake of iron from the environment can be accomplished via siderophore-dependent or siderophore-independent mechanisms; involving an outer-membrane receptor, periplasmic binding protein, inner-membrane transport (for Gram-negative species) and potentially cytoplasmic binding proteins. Uptake of heme from the environment (in Gram-negative species) occurs through either direct heme uptake systems, bipartite systems or hemophore-mediated systems. All heme uptake systems are currently believed to require energy provided via ExbB/ExbD/TonB pump systems (or analogous systems). The major differences between the three systems lie in the type of outer-membrane

receptor for the heme/heme-containing protein and how the heme is transported across the inner-membrane (Lewis, 2010)(Wilks & Burkhard, 2007)(Wandersman & Delepelaire, 2004).

The species *P. gingivalis* has previously been shown to require iron, and in most cases based on strains or environmental conditions heme as well, for growth *in vitro*. Without exogenous supplementation via blood, serum or individual heme components *P. gingvialis* growth is generally weak or absent. Thus, *P. gingivalis* is characterized as a heme auxotroph. Previous publications and current databases suggest that *P. gingivalis* only contains the genes necessary for three of the eight steps in heme biosynthesis, as well as only part of the cobalamin biosynthesis pathway (Roper et al., 2000). However, this information was generated using potentially outdated or not all-encompassing methods. For example, data mining for gene annotations without attempting to generate or check annotations de novo, relying on single genomes to represent a species, assaying specific enzymatic reactions under limited conditions, and attempting degenerate PCR with too stringent conditions and not closely enough related base sequence. Importantly, *P. gingivalis* does contain a functional ferrochelatase, thus it can utilize protoporphyrin IX from the host after adding in the metal ion.

Three distinct heme uptake systems and a single iron uptake system have been characterized in *P. gingivalis;* Hmu*,* Tlr*,* Iht, and Feo (Olczak, Simpson, Liu, & Genco, 2005)(Lewis, 2010). Although the main genes for these systems have been identified, the mechanisms of action and regulation are not fully understood. Each of these three heme uptake systems identified in *P. gingivalis* have been confirmed to be present and presumable functional in all sequenced strains. However, given that there are drastic growth and

regulatory differences between these strains under iron deplete and replete conditions other genetic factors must be at play. Also, the systems are presumably functionally redundant under *in vitro* laboratory conditions because deletions of any one system only gives a modest iron/heme uptake deficiency; none of which lead to pigmentation defects.

Importantly, it has yet to be determined if host heme is directly incorporated into *P. gingivalis* cellular proteins or if modifications occur first. Additionally, it has not been determined if the pigment coating of *P. gingivalis* can be resorbed into the cell for use during heme or iron limiting conditions. The cellular decision whether to incorporate heme directly likely involves sensing iron and heme levels within and outside the bacterial cell, and balancing that with the energetic and component costs for heme biosynthesis.

Two significant characteristics of *P. gingivalis* are asaccharolytic metabolism and an abundance of proteinases; attributes that fit well together for an oligopeptide-rich metabolism. Greater than 80% of the proteolytic activity of *P. gingivalis* is attributed to the cysteine proteases, gingipains. There are two arginine-specific gingipains and a single lysine-specific gingipain, encoded by *rgpA*, *rgpB* and *kgp*, respectively. RgpA and Kgp contain haemagglutinin domains while RgpB does not. The haemagglutinin domains are identical to domains found on the protein HagA (haemagglutinin A), which is conserved in *P. gingivalis* (Potempa, Sroka, Imamura, & Travis, 2003). Many strains also manufacture proteins HagB and HagC, however, there is strain-specific variability of the presence and isotype. The ability of *P. gingivalis* to agglutinate and lyse erythrocytes is attributed to the haemagglutinin and proteolytic domains of the gingipains and haemagglutinins. Gingipains serve as the main method for initial protein breakdown, prior to more specific breakdown via oligo- and di- or

tri-peptidases. Gingipains can digest various serum components, albumin, immunoglobulin, transferrin and multiple defensins (Imamura, 2003)(O'Brien-Simpson, Veith, Dashper, Reynolds, 2003) (Guo, Nguyen, & Potempa, 2010) (Hajishengallis & Lamont, 2014).

CHAPTER 2: TRANSPOSON MUTAGENESIS AND ESSENTIAL GENE DETERMINATION

[The content of the following chapter has been published. All figures are included, however, due to size/length several tables have been omitted. References to the initial publication and a subsequent methods article are below:

Klein BA, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ, Hu LT. (2012) Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*. BMC Genomics. 13:578. doi:10.1186/1471-2164-13-578. PMCID: PMC3547785. PMID: 23114059

BAK conceived of the study, participated in its design and coordination, carried out molecular genetics, carried out bioinformatic analyses and drafted the manuscript. ELT participated in study design and coordination and drafted the manuscript. LTH conceived of the study, participated in its design and coordination and drafted the manuscript. DWL participated in study design and coordination and drafted the manuscript. AC participated in study design and coordination and drafted the manuscript. MJD participated in study design and coordination and drafted the manuscript.

and

Klein BA, Duncan MJ and Hu LT. (2015) Defining essential genes and identifying virulence factors of *Porphyromonas gingivalis* by massively-parallel sequencing of transposon libraries (Tn-seq). Methods in Molecular Biology, Microbial Gene Essentiality. Volume 1279, 2015, pp 25-43. DOI:10.1007/978-1-4939-2398-4_3. PMID: 25636611]

2.1 Abstract

Background

Porphyromonas gingivalis is a Gram-negative anaerobic bacterium associated with periodontal disease onset and progression. Genetic tools for the manipulation of bacterial genomes allow for in-depth mechanistic studies of metabolism, physiology, interspecies and host-pathogen interactions. Analysis of the essential genes, protein-coding sequences necessary for survival of *P. gingivalis* by transposon mutagenesis has not previously been attempted due to the limitations of available transposon systems for the organism. We adapted a Mariner transposon system for mutagenesis of *P. gingivalis* and created an insertion mutant library. By analyzing the location of insertions using massively-parallel sequencing technology we used this mutant library to define genes essential for *P. gingivalis* survival under *in vitro* conditions.

Results

In mutagenesis experiments we identified 463 genes in *P. gingivalis* strain ATCC 33277 that are putatively essential for viability *in vitro*. Comparing the 463 *P. gingivalis* essential genes with previous essential gene studies, 364 of the 463 are homologues to essential genes in other species; 339 are shared with more than one other species. Twenty-five genes are known to be essential in *P. gingivalis* and *B. thetaiotaomicron* only. Significant enrichment of essential genes within Cluster of Orthologous Groups 'D' (cell division), 'I' (lipid transport and metabolism) and 'J' (translation/ribosome) were identified. Previously, the *P. gingivalis* core genome was shown to encode 1,476 proteins out of a possible 1,909; 434 of 463 essential genes are contained within the core genome. Thus, for the species *P. gingivalis* twenty-two, seventy-seven and twenty-three percent of the genome respectively are devoted

to essential, core and accessory functions.

Conclusions

A Mariner transposon system can be adapted to create mutant libraries in *P. gingivalis* amenable to analysis by next-generation sequencing technologies. *In silico* analysis of genes essential for *in vitro* growth demonstrates that although the majority are homologous across bacterial species as a whole, species and strain-specific subsets are apparent. Understanding the putative essential genes of *P. gingivalis* will provide insights into metabolic pathways and niche adaptations as well as clinical therapeutic strategies.

2.2 Background

Porphyromonas gingivalis is an oral Gram-negative, anaerobic, asaccharolytic and black-pigmented bacterium that is highly correlated with the development and progression of periodontal diseases and systemic comorbidities (Pihlstrom et al., 2005) (Curtis et al., 2011) (Darveau, 2010) (Holt et al., 1988) (Hajishengallis et al., 2011). The organism has been characterized as a 'keystone' pathogen whose interactions with other bacteria and the host are critical for the development of periodontitis (Hajishengallis et al., 2011). *P. gingivalis* utilizes multiple virulence factors, many of which have been identified and studied *in vitro* and *in vivo* such as proteinases (e.g. gingipains), fimbriae, non-canonical lipopolysaccharide (LPS), capsular polysaccharide (CPS), and cytotoxic and hemolytic molecules (Holt et al., 1999) (Lamont & Jenkinson, 1998) (Nakayama, 2003). The identification of genes and proteins involved in pathogenesis has most commonly relied on analyzing genetic variations between strains or by directly isolating then genetically and biochemically characterizing specific mutants (Chen et al., 2004) (Igboin, Griffen, & Leys, 2009) (McKee et al., 1988)

(Ozmeric, Preus, & Olsen, 2000). In contrast, the identification of essential genes in this organism has lagged. Essential genes can be used as targets for antimicrobial drug development, and through bioinformatic and experimental study of essential gene, may reveal unique aspects of the physiology and metabolism of *P. gingivalis.* High-throughput strategies to screen for genetic determinants of virulence and identify essential genes have been limited due to a paucity of tools for genetic manipulation (Kuramitsu, 2003).

Transposon mutagenesis has been used to identify genes involved in pathogenesis and other bacterial functions (Hayes, 2003) (Hensel & Holden, 1996) (Mazurkiewicz, Tang, Boone, & Holden, 2006) (Picardeau, 2010) (Reznikoff & Winterberg, 2007). The utility of transposon mutagenesis depends on the ability of the transposable element to insert randomly into different sites in the host genome in a one insertion per strain manner. Two previous transposon mutagenesis systems for *P. gingivalis* were based on *Tn4435* and *Tn4400* (Chandad, Mayrand, Grenier, Hinode, & Mouton, 1996) (Chen et al., 2000) (T. Chen, Dong, Yong et al., 2000) (Hoover & Yoshimura, 1994) (Takada & Hirasawa, 1998). While the use of the mutant libraries generated with these systems led to important insights into *P. gingivalis* pathogenesis, both elements inserted preferentially into 'hot-spots' in the genome thus limiting the distribution of interrupted genes and were also limited to which strains could be mutagenized. The lack of insertion saturation with these transposons into *P. gingivalis* genes resulted in libraries that were not suitable for the genome-wide identification of essential genes.

Mariner-family transposons have been used to generate highly-saturated mutant libraries in numerous phylogenetically distinct bacterial species (Picardeau, 2010). Mariner transposons preferentially insert into 'TA' nucleotide sequences, which are abundant

throughout genomes; the *P. gingivalis* ATCC 33277 genome only has four NCBI annotated genes that lack a TA site, all of which are hypothetical proteins and are less than 40 amino acids in length (Lampe, Churchill, & Robertson, 1996) (Lampe, Akerley, Rubin, Mekalanos, & Robertson, 1999) (Bryan, Garza, & Hartl, 1990). No other constraints or preferences for Mariner transposon insertion have been identified. Recently, several investigators paired mutagenesis with mini-transposon derivatives of the Himar 1 Mariner transposon with massively-parallel sequencing technology in strategies variously named Tn-seq, IN-seq and HITs (van Opijnen et al., 2009) (van Opijnen & Camilli, 2010) (Gawronski et al., 2009) (Goodman et al., 2009). These strategies allow for quantitative assessment of individual mutants in a library by sequencing the transposon-genome junctions. Complex Mariner transposon libraries, in some cases saturating, have been used to define essential genes of several bacterial species including *Bacteroides thetaiotaomicron, Campylobacter jejuni*, *Haemophilus influenzae*, *Staphylococcus aureus* and *Streptococcus pneumoniae* (Gawronski et al., 2009) (Goodman et al., 2009) (Stahl & Stintzi, 2011) (Metris, Reuter, Gaskin, Baranyi, & van Vliet, 2011) (Molzen et al., 2011) (Chaudhuri et al., 2009). Data from these studies have been collated into a Database of Essential Genes (DEG) (Zhang, Ou, & Zhang, 2004) (Zhang & Lin, 2009) (Zhang & Zhang, 2007). Comparison of essential genes between bacterial species included in the DEG reveals that a large fraction of essential genes are species-specific. *P. gingivalis* essential genes cannot simply be inferred from studies in other bacteria, and such studies in *P. gingivalis* have not been performed to date, although a 'core' genome has been described by comparing ten strains by DNA/DNA-hybridization using microarray technology (Brunner et al., 2010). However, while there is likely to be overlap, a core genome does not equate with the set of genes essential for survival, and likely includes

both essential and non-essential genes.

We have successfully adapted a Mariner-based transposon mutagenesis system to create highly-saturated mutant libraries in *P. gingivalis*. Here we describe our construction and analysis of this mutant library to identify essential genes in *P. gingivalis* and compare these genes to those identified in other bacteria.

2.3 Results and Discussion

2.3.1 Generation of the mutant library

We generated transposon insertion libraries in *P. gingivalis* using a Himar 1 Mariner mini-transposon system created for use in *Bacteroides thetaiotaomicron* (Goodman et al., 2009). The *B. thetaiotaomicron* promoter of BT1331 that drives expression of *himar1c9a* transposase is recognized by *P. gingivalis*, allowing us to use the *B. thetaiotaomicron* plasmid vector pSAM_Bt with modifications in growth media and antibiotic selection conditions. This mini-transposon is constructed with two transcription terminators downstream of the gene for antibiotic selection, thus eliminating read-through downstream from the insertion.

We performed mutagenesis using pSAM_Bt with *P. gingivalis* strain ATCC 33277. The 4.6 kb pSAM_Bt vector containing the Mariner mini-transposon cannot replicate in *P. gingivalis* and, in addition, the plasmid lacks sequence homology with the *P. gingivalis* genome. Therefore, after the plasmid enters *P. gingivalis* by transformation, transposition from the plasmid into the genome occurs without significant background insertion of the plasmid into the genome by illegitimate recombination. This system allows for single, stable transposition events since transposase activity is lost along with the plasmid. We collected 54,000 transposon insertion strains (individual colonies) from six separate transformation

experiments. Variable colony sizes were observed among the mutants harvested and pooled following 14 days of growth. However, the majority of macroscopically visible colonies were similar in size to those of wild-type *P. gingivalis* strain ATCC 33277 after 14 days of growth. To confirm that the strains contained transposon insertions and not cryptic or full plasmid integrations, we performed PCR specific for the transposon (*ermG*) as well as for two portions of the vector backbone (*himar1c9a* and *bla*) (Fig.2-1A). Of 100 colonies that were screened, all showed positive PCR reactions for the transposon gene and negative reactions for the vector backbone, indicating 'correct' transposition. 'Incorrect' transpositions can include portions of the vector backbone inserting with the transposon, the vector being stably maintained within the bacterium extra-chromosomally or multiple insertions within the same genome; such transposition events were not detected in the subset of mutants tested (Fig.2-1). To determine whether the transposon inserted into different genes and not preferentially into genetic 'hot-spots', we performed nested semi-random PCR followed by sequencing which confirmed that insertions occurred in multiple locations across the genome (Fig.2-1D). This traditional sequencing method is effective for targeted sequencing a subset of mutants from the mutant library if massively-parallel high-throughput are neither desired nor necessary.

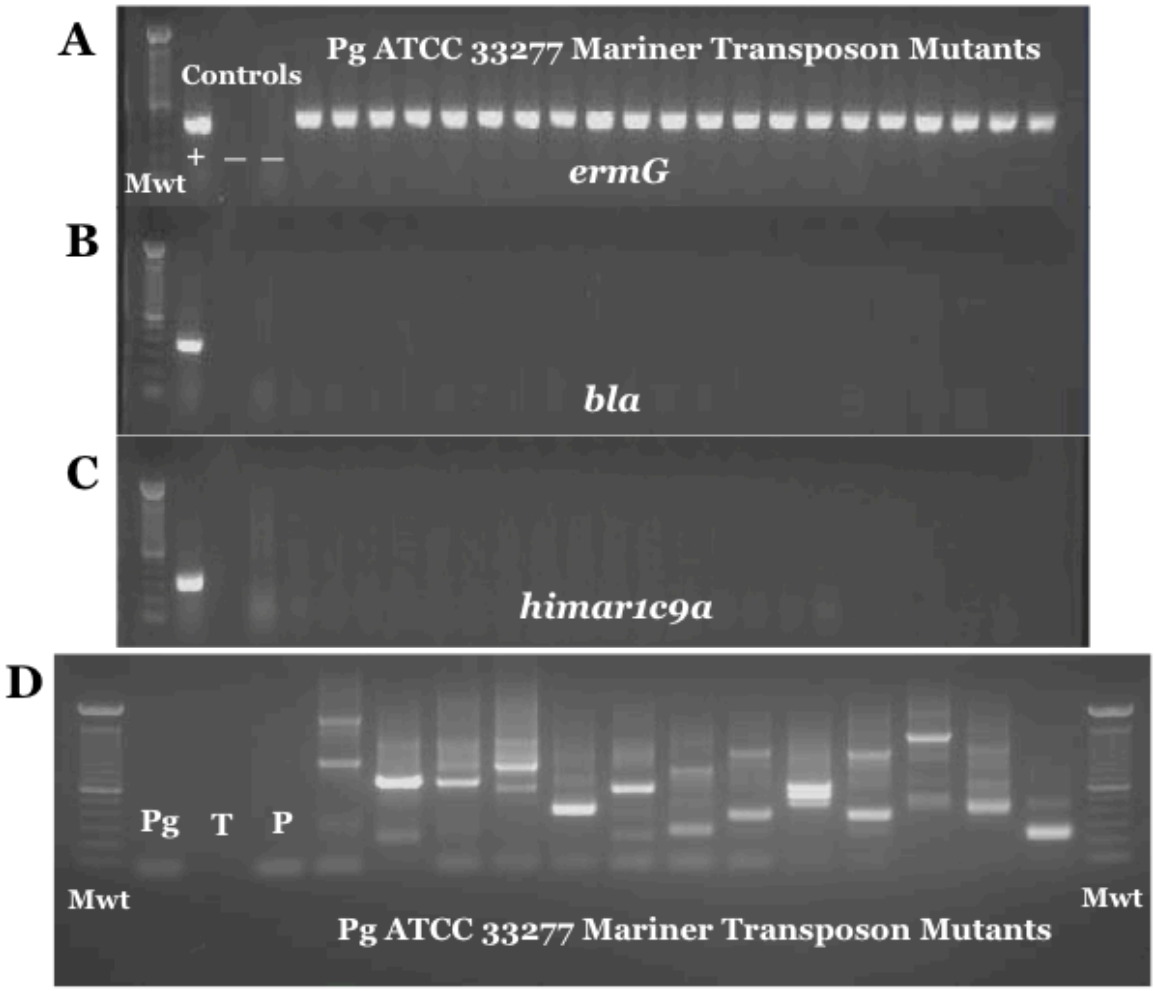


Figure 2-1. Determination of proper transposon insertion. Confirmation of transposon insertions was performed by PCR for presence of transposon (*ermG*) (A). "Mwt" = molecular weight marker, "+" = positive control (gDNA from *E. coli*/pSAM_Bt); "-"= a negative control (*P. gingivalis* ATCC 33277). All other lanes contain amplicons from PCR of individual colonies of transformed *P. gingivalis*. Panels (B) and (C) show PCR of the same samples using primers for the *bla* and *himar1c19a* genes respectively that are present in the plasmid, but which should be lost with proper insertion of the transposon. These three panels are a combination of separate gels; all which were run using identical PCR gDNA template for each of the separate reactions. (D) Nested semi-random PCR for individual mutant sequencing preparation. PCR from individual colonies was performed using primers to Mariner transposon and random primers ARB1 and ARB2 (Additional file 6: Table S6). Two rounds of nested PCR were performed. Negative controls of wild-type *P. gingivalis* strain ATCC 33277 (*Pg*), template only (T) and primer only (P) lanes precede thirteen individual mutants.

2.3.2 Validation of Tn-seq of the P. gingivalis library

Having confirmed via nested semi-random PCR and subsequent sequencing that the libraries contained different transposon insertions scattered throughout the genome, we identified the location of each insertion in the library by Tn-seq analysis (van Opijnen et al., 2009) (van Opijnen & Camilli, 2010). This method couples transposon mutagenesis with massively-parallel, next-generation sequencing technology to identify the location of each insertion and quantitate the relative abundance of each insertion mutant in the library.

Prior Tn-seq experiments using Mariner libraries have taken advantage of an engineered *Mme*I restriction site that cuts 18–20 base pairs away from the recognition site into the genomic DNA. This method has been successfully employed to evaluate library sequences in a variety of settings, however, it suffers from a number of disadvantages including: 1) Yielding small sequencing reads limited to 16–18 nucleotides in length. 2) Requiring the use of a mutant transposon and hence existing transposition vectors must be mutated. 3) *MmeI* generates 2 base pair 3' overhangs at adjacent sequences to which adapters are ligated. It is possible that the enzyme cleaves these adjacent sequences with varying efficiency. Furthermore, T4 DNA ligase is likely to join adapters to these varying overhangs with differential efficiency (for instance GG should be more efficient than AA). Such variations in efficiencies, if they exist, would lead to unequal representation of sequenced insertions. An alternate method for sequencing from junctions in transposon libraries involves the ligation of adapters to sequences near transposon junctions (Gawronski et al., 2009). However, the method is labor intensive, requires gel purification of ligated products, and is prone to the unintended creation of inhibitory adapter dimers and other side products.

Here we report a new method, without the abovementioned disadvantages, for the

construction of high-throughput sequencing libraries from transposon, retrovirus or repetitive element insertions sites in any genome. For details see the Materials and Methods section. In brief, genomic DNA containing the insertion element of interest is sheared, and then Terminal deoxynucleotidyl Transferase (TdT) is used to add an average of twenty deoxycytidine nucleotides to the 3' ends of all molecules. Two rounds of PCR using a poly-C-specific and an insertion element-specific primer pair are then used to amplify one of the two insertion element-genomic DNA junctions and append all user-defined sequences needed for high-throughput sequencing and indexing. This new method does not require a ligation reaction, does not produce adapter dimers, does not require gel purification and is compatible with long sequencing reads the size of which is only limited by the length of library fragments and the sequencing technology. Here, in contrast to the 16–18 nucleotide reads obtained with the *MmeI* method we used 50 nucleotide reads allowing for significantly more effective and precise mapping of sequences to regions with nucleotide repeats as well as genes that contain nucleotide homology (Fig.2-2A). This is particularly important given that the current Illumina HiSeq2000 base-calling algorithm gives poor quality scores for the first few bases (Fig.2-2A).

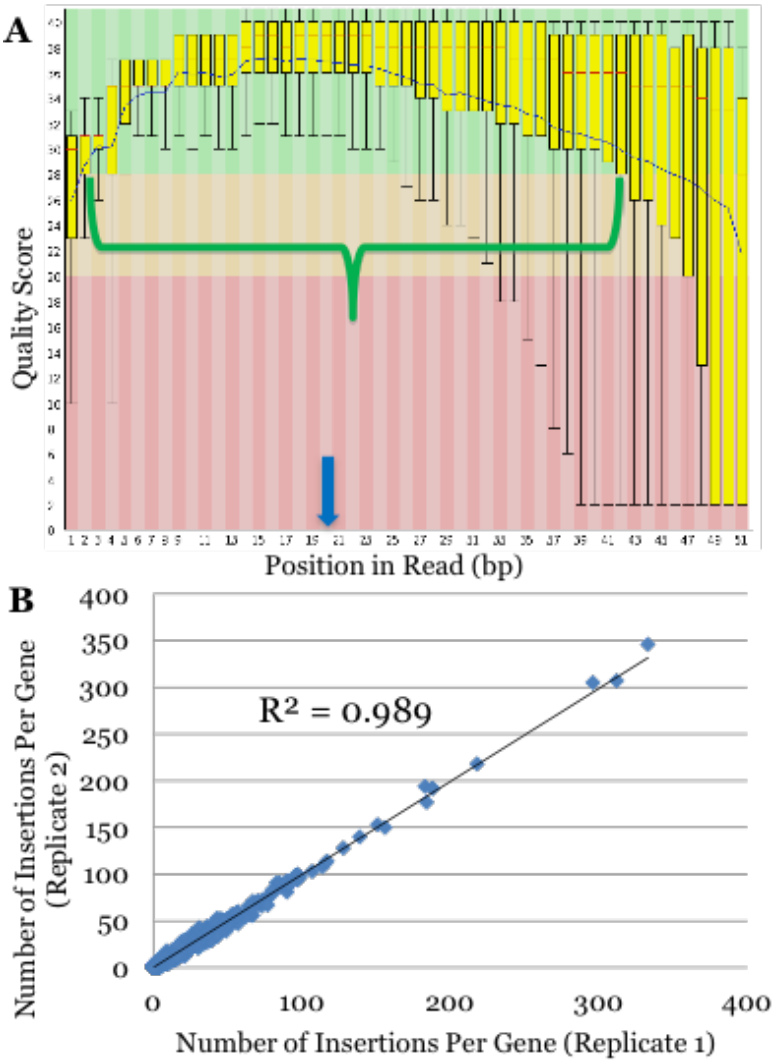


Figure 2-2. Sequencing quality control and reproducibility. Panel A shows quality scores of the Illumina sequencing reads for mapping. Fifty base pair single-end reads were obtained with 'high' quality out to ~42 base pairs and 'good' quality out to ~47 base pairs. The green background corresponds to high quality reads, the yellow background to intermediate quality reads and the red background to poor quality reads. Data shown are for the number of high, intermediate and low quality reads at a specific number of base pairs away from the transposon. The yellow bar encompasses the 25-75th percentile and the red horizontal bar indicates the mean. The green bracket identifies the base pair position where high quality reads comprise the over 75% of the total reads. The blue arrow signifies the amount of sequencing that can be obtained when preparing DNA using the *MmeI* restriction site, demonstrating superior mapping and analysis ability of C-tailing method. No sequencing reads shorter than 20 bp were used for analyses. B) Replicates of the same library were sequenced in separate experiments. The graph compares the number of insertions per gene for technical replicates 1 and 2 of *P. gingivalis* strain ATCC 33277 Mariner mutant library and showed excellent correlation between the replicates (R^2=0.9892). Median number of insertions when excluding genes containing zero is 9 while the mean is 17. Sixteen genes have 100 insertions or greater.

Two replicate samples derived from the same master mutant library, but processed separately for sequencing, were compared. Sequencing revealed 35,937 and 35,732 distinct insertions (mutants) respectively (Fig.2-2B). Of the total insertions, 7,230 and 7,193 in the respective replicate runs were in putative intergenic regions. After quality filtering sequencing reads an average 6,310,573 reads could be attributed to an average of 35,835 insertions mapped to the genome. Of note, during multiplexed Illumina sequencing runs between 10–20 percent of sequencing reads are 'thrown out' during quality control analyses. This level of 'discarded' read data is seen by all groups performing permutations of Tn-seq, RNA-seq, ChIP-seq and other massively-parallel adapted methods. Sequencing data removed during our quality control analyses was within the 10–20 percent range previously noted. The number of insertions per gene and the number of reads per gene when comparing the technical replicates gave R^2 values of 0.989 and 0.998 respectively (Fig.2-2B). The similarity between the two technical replicates demonstrates that aliquot production from the master library, processing of the samples as well as sequencing and analyses are highly reproducible.

2.3.3 Identifying putative essential genes of P. gingivalis by Tn-seq

The genome of *P. gingivalis* strain ATCC 33277 comprises 2.35 Mb of chromosomal DNA and no plasmids. With a GC-content of 48.4%, there are 2,155 genes, 2,090 protein-coding sequences, 53 tRNA, and 12 rRNA [GenBank: AP009380.1]. An important factor for Mariner transposition is that the genome contains 117,742 informative 'TA' sites, the only known specific motif 'required' for Mariner transposition. In previous studies and in agreement with our sequencing results, approximately 98% of Mariner insertions occur at TA

sites (unpublished results).

The presence in our library of a mutant bacterium, in which a gene or intergenic region has been interrupted by insertion of the Mariner transposon, would indicate that it is unlikely that the gene or region is essential for growth *in vitro* on blood agar plates, provided that the insertion was likely to have inactivated the function. Insertions into the first or last five percent of a gene have a higher likelihood of generating a functional gene product relative to insertions in the middle portions of a gene, therefore these mutants were eliminated from consideration. In addition, we required a minimum of three sequencing 'reads' of the mutant locus be present in both technical replicates to exclude nonexistent insertions introduced by mapping of incorrectly sequenced reads, and lower rates may be due to mis-assignment by the reference assembly software. By these criteria, we determined that 1,639 genes are non-essential for growth *in vitro*. Sixteen of these genes contained 100 or greater insertions, notably the proteinases/adhesins *kgp* (310), *rgpA* (300), *rgpB* (152) and *hagA* (340) as well as the minor fibrilin *mfa1* and four of the twelve 23S rRNA genes. Eighty-eight genes contained 50 or more insertions and 837 contained 10 or more insertions, with a median number of 10 insertions per gene. There is a direct, but not completely exclusive, correlation between number of insertions and sequencing reads as 9 of the top 10 highest reads from the library belong to genes with more than 100 insertions; thus *kgp*, *rgpA*, *rgpB* and *mfa1* are in the top ten most-read genes. Nine hundred and twenty genes had transposon insertions in at least 25% of their reported TA sites, while a remaining 528 genes had insertions in at least 10% of their reported TA sites. The average number of TA sites per gene, when including all 2155 genes (CDS, tRNA and rRNA), is 55. A total of 87 genes were fully saturated with at least one mutant insertion into every available TA site in the gene. Full

saturation results in a TA insertion ratio (actual number of different insertions into the TA sites of a gene divided by the theoretical maximum number of different insertions into the TA sites of a gene) of 1. A TA site to insertion ratio of greater than 1 demonstrates that at low frequency Mariner will insert into non-TA sites, most likely due to medium composition such as salt concentration that alleviate transposon specificity, local DNA structure, nucleotide composition and/or DNA-binding proteins. Of the 87 fully saturated genes having on average 49 TA sites, 64 are present in multiple copies throughout the genome (Additional file 1: Table S1) and include rRNA genes, *ISPg1*, *ISPg3*, gingipains, and hypothetical proteins (Additional file 1: Table S1-1). All of the rRNA genes (12 in total) are located in spatially separated clusters of three and are fully saturated. Efforts are currently underway to determine whether there is conservation among non-TA insertion sites.

For the remaining putative essential genes we applied the following rules to identify those most likely to be essential for *in vitro* survival. The rules are similar to those used in previous essential gene analyses of other bacteria and contend that (Christen et al., 2011) (Gawronski et al., 2009) (Goodman et al., 2009) (van Opijnen et al., 2009): 1) A gene must contain at least 10 TA sites. Genes with less sites could be under-inserted due to random chance. Of the 204 genes in the *P. gingivalis* ATCC 33277 genome with less than 10 TA sites, 189 (93%) are annotated as hypothetical. 2) Genes found to have an actual to theoretical insertion ratio of 50-fold or greater under-insertion were considered putatively essential (actual:theoretical ≤ 0.020). Applying these rules, out of a total 2,102 genes in the ATCC 33277 genome (all protein coding sequences and rRNA genes combined minus the 53 tRNAs), we identified 463 (22.0%) genes as putatively essential for *in vitro* survival (described below) (Fig.2-3) (Additional file 1: Table S1). Twenty-two percent of a 2.35 Mb

genome containing 2,102 genes is within the range of essential genes determined by transposon mutagenesis, single gene deletions and *in silico* analyses of other bacterial genomes, as described below (Baba et al., 2006) (Christen et al., 2011) (Gawronski et al., 2009) (Goodman et al., 2009) (Chaudhuri et al., 2009) (van Opijnen et al., 2009) (Gerdes et al., 2003) (Khatiwara et al., 2012) (Glass et al., 2006) (Kobayashi et al., 2003) (De Berardinis et al., 2008) (Knuth, Niesalla, Hueck, & Fuchs, 2004) (Metris et al., 2011) (Molzen et al., 2011) (Salama, Shepherd, & Falkow, 2004) (Wong & Akerley, 2007).

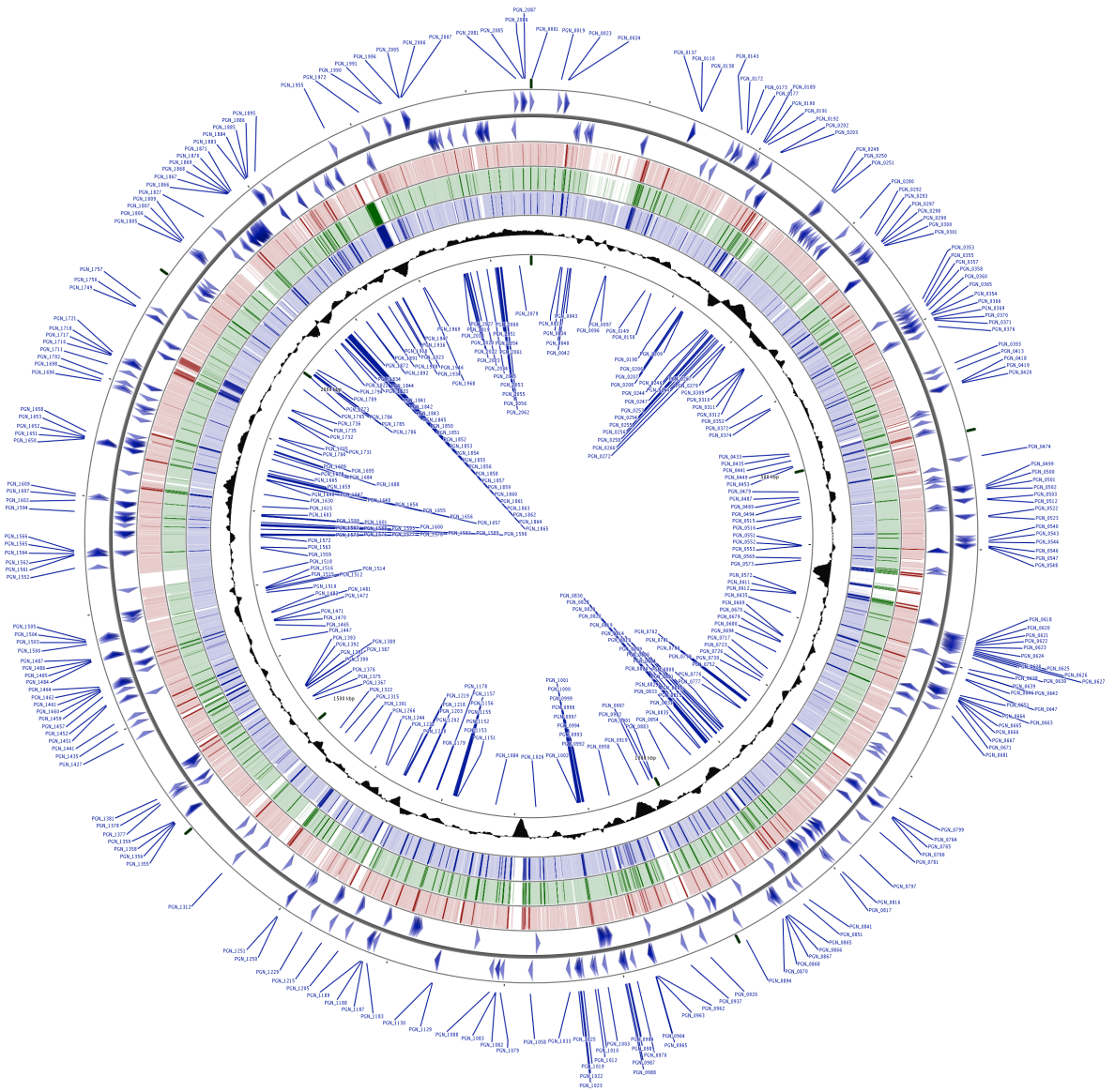


Figure 2-3. Mapping of the *P. gingivalis* essential genes. In blue in the outermost ring are shown the putative *P. gingivalis* essential genes identified by transposon mutagenesis of strain ATCC 33277. Blue arrows depict the orientation of the essential genes. Genetic loci for positive strand (outer set of arrows) are shown in the outermost circle and genetic loci for the negative strand are shown in the innermost ring. In red tick marks are coding sequences for *P. gingivalis* strain W83, in green are coding sequences for strain W50 and in blue are coding sequences for strain TDC60. Shaded black area represents GC-content for given regions. CGViewer (http://stothard.afns.ualberta.ca/cgview_server/) was utilized to visualize the entire circular genome of *P. gingivalis* strain ATCC 33277 with the putative essential genes labeled. NCBI FASTA files of *P. gingivalis* strains W83, W50 and TDC60 were used for BLAST matching to the ATCC 33277 base genome.

Prior to applying any cutoffs described above we found that 273 putative essential genes contained zero insertions. Given that these genes were a minimum of 200 base pairs in length and contained at least 10 TA sites the confidence level for concluding these as essential is high. Of the remaining 190 putative essential genes, 64 were found to have a ratio of between 0.001-0.010, 100-fold or less under-inserted, and 76 had a ratio between 0.010-0.020, 50-fold or less under-inserted. In most cases these genes had a single insertion over a gene length of 1.5-3.0 kb. Fifty genes had ratios between 0.020-0.050, however, these insertions were found to fall under the constraints outlined above and also met our qualifications for putative essentiality as well. Of note, of these 50 genes the majority (72%) have homology to genes of other bacteria identified in previous essential gene studies (Zhang et al., 2004)(Zhang & Lin, 2009).

In addition to identifying the essential nature of a gene, more detailed analysis, specifically mapping domains of proteins and intergenic regions, can provide valuable information about protein functional domains, promoter regions, mis-annotations, operon structure and regulatory RNAs (Fig.2-4/Fig.2-5). Simply mapping the insertions onto the genome to view saturation of specific genes provides a qualitative understanding of library complexity (Fig.2-4A). Annotations of genomes identify gene/coding-sequence start and stop codons, spatial relationships to other genes, operon structure, number of possible amino acids and amino acid composition. Such bioinformatic analyses are not perfect because they are based on coding-sequences from model organisms, e.g. *Escherichia coli,* and not adapted to less well-known bacterial species. Detailed insertion mapping allows for the determination of essential genes on a visual scale (Fig.2-4B). In addition, transposon mutagenesis mapping may reveal previously mis-annotated start and stop sites for genes as well as putative internal

start sites, providing information on potential operon structure. Furthermore, essentiality of function domains can be determined (Fig.2-5A/2-5B). Although intergenic regions are far less abundant in prokaryotic genomes, mapping of insertions, or a lack thereof, to a specific intergenic region within the genome can provide insights on regulatory features within non-coding DNA sequences.

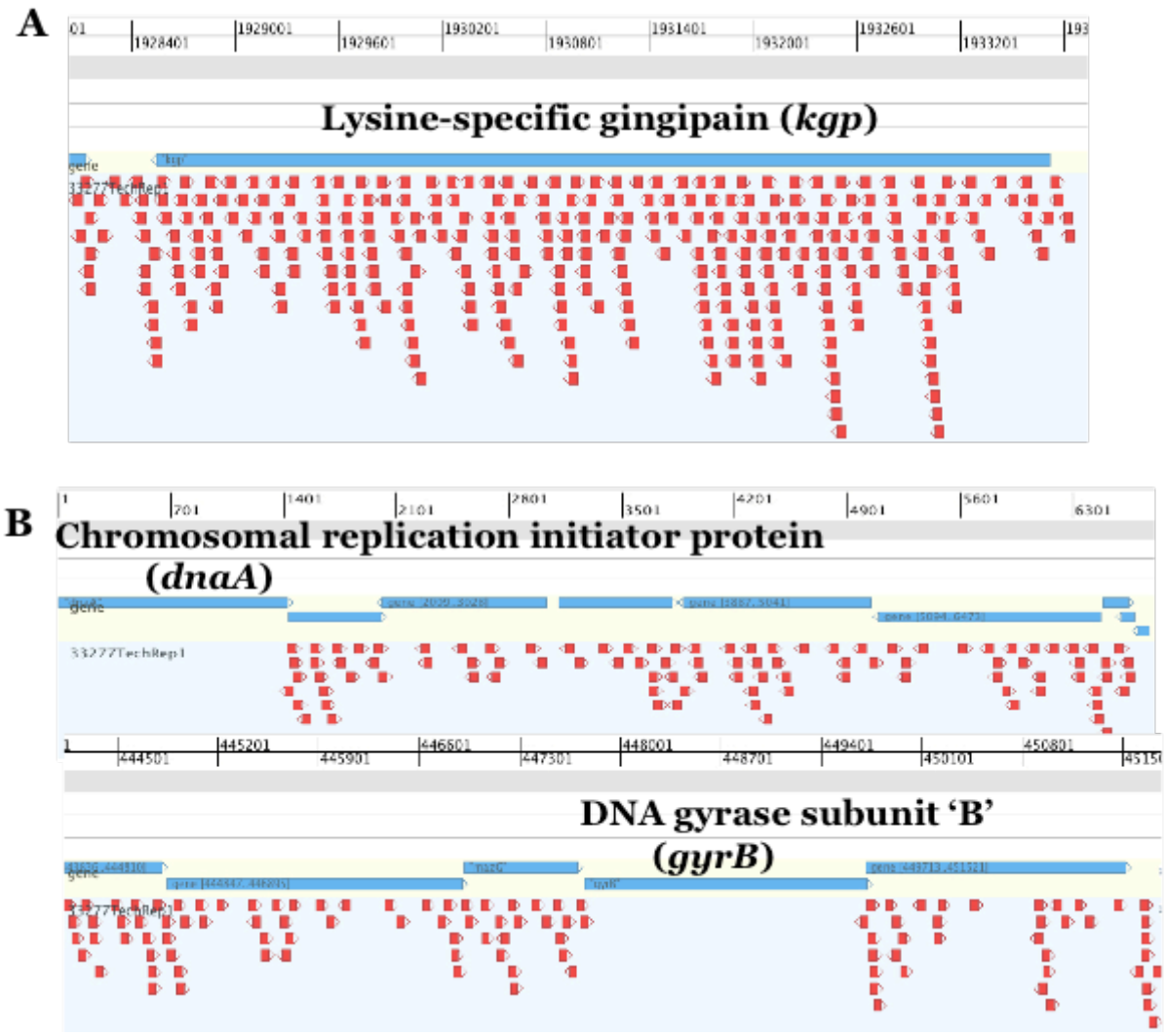


Figure 2-4. Examples of transposon insertion distribution into highly-saturated and essential genes. Panel A shows the insertion positions of the transposon for a very highly inserted gene, lysine gingipain, *kgp*. The blue bars represent the gene sequence. Each red arrow represents the location and orientation of a single insertion in the library. In panel B, we show two examples of essential genes, chromosomal replication initiator protein, *dnaA* and DNA gyrase subunit 'B', *gyrB*. As shown, there are numerous insertions in the flanking genes extending from the stop of *dnaA* and to the start and stop of *gyrB*.

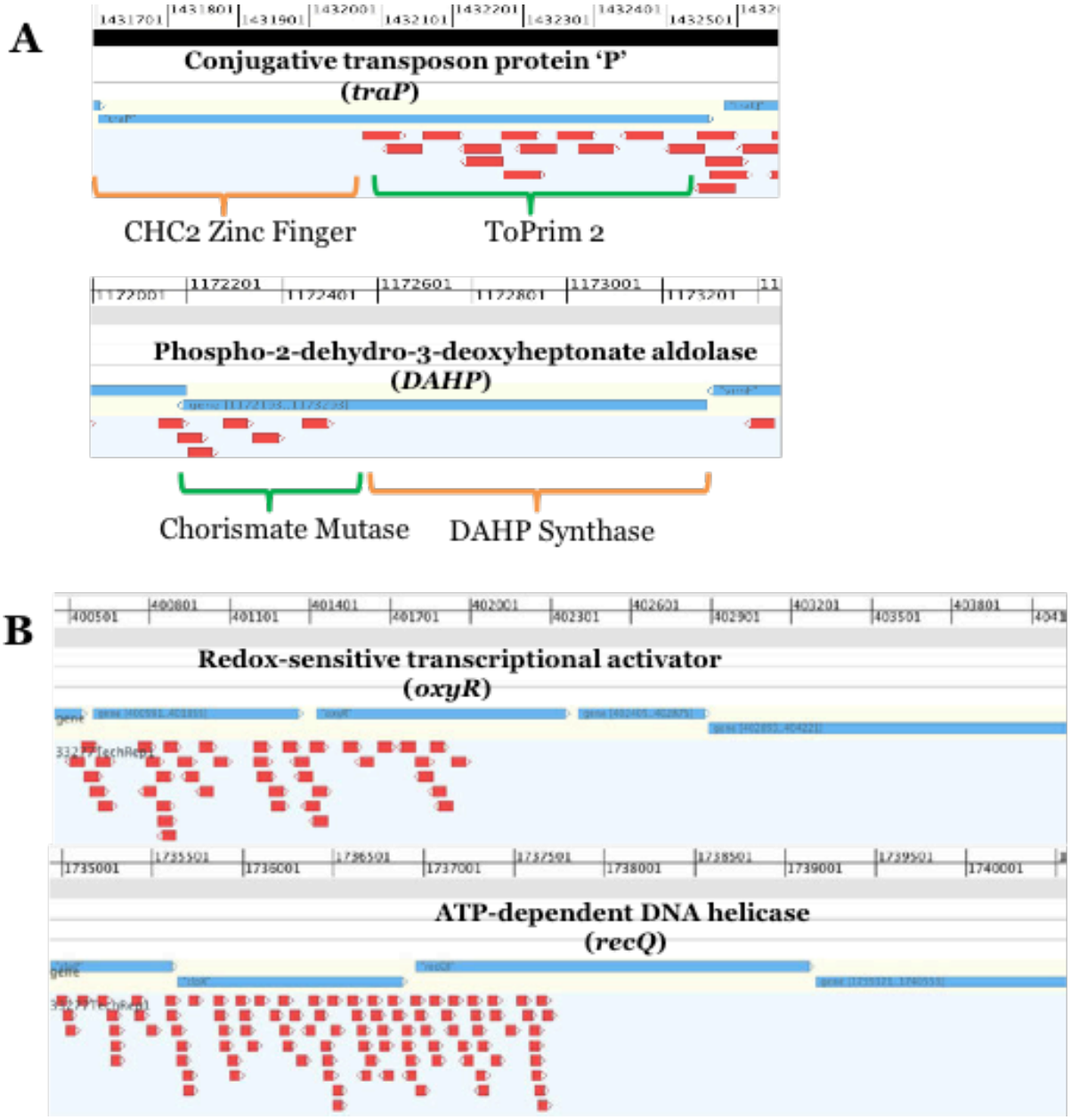


Figure 2-5. Examples of transposon insertion distribution into genes that express proteins with essential and non-essential domains. Panel A shows two examples of genes with multiple protein domains, one of which is essential and the other is not. In panel B, two non-essential genes, *oxyR* and *recQI*, are shown. Of note is the lack of transposon insertions into the latter regions of these genes. This would not be predicted based on the prevalence of TA sites in the non-inserted region of *oxyR* or *recQI*. GenomeView (http://genomeview.org/) was utilized for mapping and visualization of the transposon mutant insertion sequencing reads from Illumina sequencing to the *P. gingivalis* base genomes. BED file tracks created from Galaxy platform analyses were used for mapping insertions to a GenBank file base track.

2.3.4 Comparison of P. gingivalis essential genes to core genome and transcriptome

The core genome of *P. gingivalis* previously proposed by Brunner *et al*. was derived from hybridization analysis of 10 different strains to a DNA microarray of annotated genes from strain W83 (Brunner et al., 2010). Of note, any gene 'missing' from strain W83, even if present in all other strains of *P. gingivalis*, would be considered not part of the 'core'. Since both strains W83 and ATCC 33277 are now fully sequenced, it is known that 8 genes in the *P. gingivalis* ATCC 33277 essential list are missing in W83. Five of these genes have been identified in a third sequenced and annotated strain, TDC60. There was nearly complete overlap between putative essential genes determined by Tn-seq with the *P. gingivalis* core genome, 434 of 463 (93.7%) putatively essential genes overlapped (Additional file 1: Table S1) (Brunner et al., 2010). Also, several gene probes were left out of the core genome analysis due to low hybridization signals or redundancy; two of these are identified as essential in our study. Nearly half (12 out of 31) of the essential genes not found in the *P. gingivalis* core genome had BLAST matches in the Database of Essential Genes (DEG) (Zhang et al., 2004)(Zhang & Lin, 2009). The remaining small difference may be explained by the hypothesis that certain essential genes are strain- and not species-specific, and thus may not be identified in a core genome analysis. In the circular genome representation of the base genome *P. gingivalis* ATCC 33277 (Fig.2-3) essential genes are depicted in arrows denoting directionality (blue) and homologous coding sequences are shown as tick marks in strains W83 (red), W50 (green) and TDC60 (blue). The map also shows that areas of genetic aberrance between *P. gingivalis* strains are areas devoid of essential genes (Fig.2-3). This would be hypothesized as essential genes should be conserved throughout a species unless

duplication or gain-of-function mutation occur that can rescue the essential role of a give gene. As more *P. gingivalis* strain genomes are sequenced, bioinformatic analyses that provide mapped read-outs will delineate putative essential, core and accessory genetic regions, thus giving insight into strain-based differences within the species. Such differences may be useful to identify strain phylogeny and aid in clinical treatment regimens based on knowledge of genotype-to-phenotype virulence attributes (eg. antimicrobial resistance and gene transfer).

Chen *et al*. performed RNA-seq analysis of mRNA expression by *P. gingivalis* strain W83 from which 455 of the possible 463 ATCC 33277 essential genes were assayed (Høvik, Wen-Han, Olsen, & Chen, 2012). This analysis demonstrated that 452 of 455 *P. gingivalis* ATCC 33277 essential genes were expressed during growth on blood agar medium (Additional file 1: Table S1). The 3 genes not expressed on blood agar plates as determined by RNA-seq are annotated as 'hypothetical' proteins. Transcriptome analyses were also performed on *P. gingivalis* grown on minimal (MIN), tryptic soy (TSB) and blood agar (BA) media, however, no essential genes were expressed solely on BA and not TSB or MIN despite some differences in levels of expression between the three media.

2.3.5 Comparison of P. gingivalis essential genes with other essential gene analyses

Of the 463 putative essential genes in *P. gingivalis*, 364 (78.6%) have known essential gene homologues determined by BLASTP interrogation of the DEG (http://tubic.tju.edu.cn/deg/), version 6.8, updated on November 4, 2011 (Additional file 1: Table S1) (Zhang et al., 2004)(Zhang & Lin, 2009). The DEG curates a searchable list of "Essential genes [that] are those indispensable for the survival of an organism, and therefore are considered a

foundation of life". *P. gingivalis* essential genes were determined to have DEG homologues based strictly on BLASTP similarity. BLASTP similarities that resulted in e-values of $1x10^{-8}$ or less were considered matches. Homologies were found in at least one of the following species which had previously undergone essential gene studies: *Bacillus subtilis*, *B. thetaiotaomicron*, *E. coli*, *Francisella novicida*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycobacterium tuberculosis, Mycoplasma genitalium*, *Mycoplasma pulminous*, *Saccharomyces cerevisiae*, *Salmonella typhimurium, Staphylococcus aureus*, *Streptococcus pneumoniae* and *Vibrio cholerae* (Baba et al., 2006) (Christen et al., 2011) (Gawronski et al., 2009) (Goodman et al., 2009) (Chaudhuri et al., 2009) (van Opijnen et al., 2009) (Gerdes et al., 2003) (Langridge et al., 2009) (Khatiwara et al., 2012) (Glass et al., 2006) (Kobayashi et al., 2003) (De Berardinis et al., 2008) (Knuth et al., 2004) (Metris et al., 2011) (Gallagher et al., 2007) (Scholle & Gerdes, 2007) (Molzen et al., 2011) (French et al., 2008) (Sassetti, Boyd, & Rubin, 2003) (Salama et al., 2004) (Song et al., 2005) (Xu et al., 2011) (Wong & Akerley, 2007) (Cameron, Urbach, & Mekalanos, 2008). For more than half of the 364 BLAST-matching essential genes there was homology within two or more species. In cases where only one other species contained a BLASTP match to a *P. gingivalis* essential gene it was most frequently to a gene in *B. thetaiotaomicron*, *H. influenzae* or *H. pylori*, which are the most closely related species to *P. gingivalis* both based on phylogeny and ecology.

The remaining 21.4% of putative essential genes that have no known homologue in the DEG may be essential in a species-specific or niche-specific manner. These 99 genes, many of which are functionally classified as containing known Pfam protein motifs, 'conserved domains' or 'hypothetical' proteins, may reveal important aspects related to metabolism and physiology of *Porphyromonas* species and closely related organisms. Of the 46 annotated as

hypothetical proteins, 42 are among the 99 *P. gingivalis* essential genes not previously known to be essential from other studies.

Of the organisms for which an essential gene set has been identified, *H. influenzae*, *F. tularensis*, *Acinetobacter*, *M. tuberculosis*, *Salmonella typhimurium*, *S. aureus* and *B. thetaiotaomicron* are the most relevant based on genome size, ecological niche and genetic relatedness to *P. gingivalis*. The determined essential genes of the above species were 1,657 genes with 462 essential (28%); 1,719 genes with 390 essential (23%); 3,307 genes with 499 essential (15%); 3,988 genes with 614 essential (16%); 4,314 genes with 353 essential (8%); 2,892 genes with 351 (12%); and 4,902 genes with 325 (6.6%), respectively (Akerley et al., 2002) (Goodman et al., 2009) (Chaudhuri et al., 2009) (De Berardinis et al., 2008) (Gallagher et al., 2007) (Khatiwara et al., 2012) (Sassetti et al., 2003).

P. gingivalis is a member of the *Bacteroidetes*, and before reclassification was known as *B. gingivalis*. There are no *Bacteroidetes* species or other anaerobes represented in the DEG, however, a putative list of *B. thetaiotaomicron* strain VPI-5482 essential genes is available from the supplemental material of Goodman *et al*. 2009(Goodman et al., 2009). *B. thetaiotaomicron* strain *VPI*-5482 was originally isolated from human feces. The strain contains a 6.26 Mb chromosome and 0.03 Mb plasmid (NC_004663.1/NC_004703.1) with 4,864 genes (chromosome) and 38 (plasmid), 4,778 protein coding sequences (chromosome) and 38 (plasmid), 71 tRNA and 15 rRNA genes (Xu et al., 2003) [GenBank: AE015928.1 and AY171301.1]. In comparison, *P. gingivalis* ATCC 33277 has 43% (numerically) of protein coding sequences in a genome 37% of the size of that of *B. thetaiotaomicron* VPI-5482. It was estimated that *B. thetaiotaomicron* VPI-5482 contains 325 "candidate essential genes" (Goodman et al., 2009). Maintaining a larger genome and gene set provides more

opportunities for functional redundancy and alternative pathways which can lead to a relatively smaller number of essential genes. Thus, 268 of 325 (82.5%) of *B. thetaiotaomicron* 'essentials' have BLAST homologues in *P. gingivalis* strain ATCC 33277 and of these, 78% (209 of the shared 268) are also essential in both organisms (Additional file 1: Table S1). Fifty-nine *B. thetaiotaomicron* BLAST matches are not essential in *P. gingivalis* and 57 have no BLAST match at all in the organism (Additional file 2: Table S2). A significant number of the shared essential genes (25 of the 209) are not characterized in the DEG (Additional file 3: Table S3) and of these 25 *Bacteroidetes*-specific essentials, three are annotated as permeases and two appear to be regulatory. Three essentials, PGN_1026, PGN_1481 and PGN_0249, are likely associated with capsular polysaccharide biosynthesis based on PGN_1026 and PGN_0249 being involved in the dolichol pathway and PGN_1481 functionally annotated as polysaccharide biosynthesis related. Parsing out essential genes of specific groups of species, in this case *Bacteroidetes* and/or anaerobes, can allow for specific drug targeting or directed nutrient supplementation.

In agreement with multiple previous studies on essential genes of bacteria, in *P. gingivalis* a significantly greater number of essential genes (248 or 53.6%) are found on the negative DNA strand, and 215 (46.4%) are found on the positive DNA strand (Additional file 1: Table S1) (Lin, Gao, & Zhang, 2010). Similarly, there is a greater than expected proportion of enzymes, especially those within multiple functions or involved in multiple pathways, within the essential gene groups (Gao & Zhang, 2011).

Using the Cluster of Orthologous Groups (COG) functional class designations (NCBI), we identified significant enrichment of essential genes within groups 'D' (cell cycle control/cell division), 'I' (lipid transport and metabolism) and 'J' (Translation/Ribosome);

and a lack of enrichment was seen in ‘S’ (function unknown), ‘P’ (inorganic ion transport and metabolism) and ‘N’ (motility) (Fig.2-6A/2-6B) (Tatusov, Galperin, Natale, & Koonin, 2000; Tatusov et al., 2001; Tatusov et al., 2003). Enrichment (or lack thereof) of essential genes in these categories has been reported previously, however, essential gene enrichment in specific COG categories appears to be a species-specific characteristic.

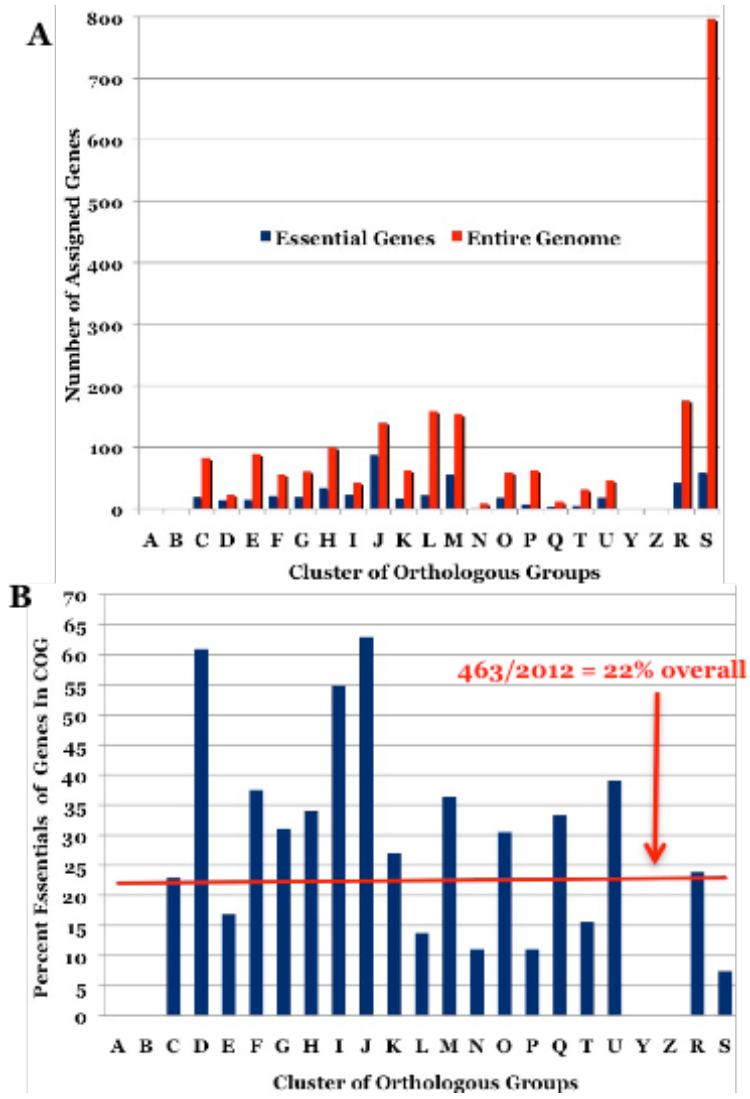


Figure 2-6. Distribution of *P. gingivalis* ATCC 33277 essential genes by Cluster of Orthologous Groups (COG) classifications. A) Number of genes within a COG category; essential genes in blue and entire genome in red. B) Percent of essential genes within a COG category from total number in genome; red line represents the 22% that 463/2102 of the percent essential over the entire genome. [A = RNA processing and modification, B = Chromatin structure and dynamics, C = Energy production and conversion, D = Cell cycle control, E = Amino acid metabolism and transport, F = Nucleotide metabolism and transport, G = Carbohydrate metabolism and transport, H = Coenzyme metabolism, I = Lipid metabolism, J = Translation, K = Transcription, L = Replication and repair, M = Cell wall/membrane/envelop biogenesis, N = Cell motility, O = Post-translational modification, protein turnover, chaperone functions, P = Inorganic ion transport and metabolism, Q = Secondary structure, T = Signal transduction, U = Intracellular trafficking and secretion, Y = Nuclear structure, Z = Cytoskeleton, R = General functional prediction only, S = Function unknown].

Based on operon prediction and known essentials contained in the DEG it was determined that 25 of the 463 putative essential genes of *P. gingivalis* identified by Tn-seq may be the result of polar effects of the transposon insertion on downstream essential genes (Additional file 1: Table S1). Specifically, these 25 genes were identified as being upstream and potentially in an operon with one or more known essential genes, and additionally do not have BLAST matches in the DEG. Further study of each of these genes would be required to confirm their essentiality.

Bringing the DEG and *P. gingivalis* core genome together in relation to *P. gingivalis* gene essentiality, we have determined that 369 genes within the core genome, ones not identified as essential in our study, have BLAST matches to genes within the DEG (Fig.2-7)(Additional file 4: Table S4). Within our mutant libraries we were able to identify transposon insertions into these genes such that they do not qualify as essential in *P. gingivalis*. Reasons for these genes being identified as essential in other species could be due to multiple variables such as *in vitro* selection media, species-specific essentiality, transposon type, library complexity, sequencing method, and criteria for essentiality. Such information gives importance to the distinction between a core gene set and an essential gene set as well as possible limitations of essential gene analyses based solely on *in silico* methodology.

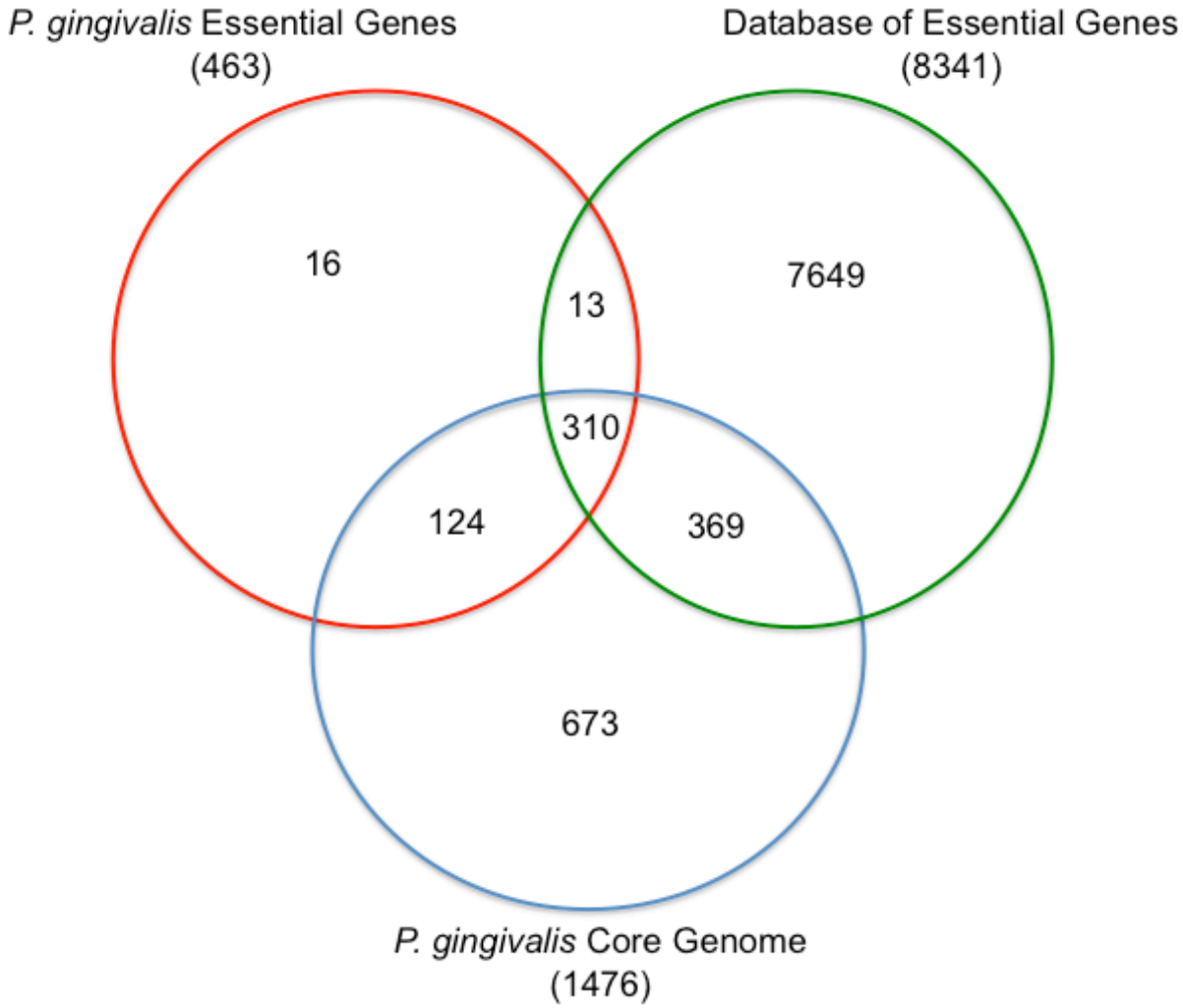


Figure 2-7. Comparison of *P. gingivalis* essential genes, core genome and the entire DEG. The red circle represents the distribution of the 463 *P. gingivalis* essential genes (determined using strain ATCC 33277), blue represents the distribution of the 1,476 gene *P. gingivalis* core genome and green represents the entirety of the 8,341 DEG curated genes. The innermost overlap shows genes found within all three datasets, while those found outside of all overlaps show genes specific to that individual dataset. Corresponding gene information can be found in Tables 1S and 3S.

2.3.6 Characterization of P. gingivalis essential genes

Metabolic pathways that lack redundancy or have critical functions have been identified previously through essential gene studies. In our analysis of *P. gingivalis* we noted the presence of entire pathways as well as specific parts of pathways that are essential to *P. gingivalis* and to all other bacterial species. A subset of *P. gingivalis*-specific essential genes, possibly related to the ecological niche of the species, have also been identified (Additional file 1: Table S1). Of pathways involved in ribosome function we identified the *rpsA*, *rpmA*, *rplB* and *rimP* systems, which encode for 30S, 50S and maturation of ribosomes, respectively. The three major protein translation regulatory pathways of *infB*, *tsf* and *prfA,* as well as translational machinery pathway *rpoA* were found to be essential in our study. DNA replication, recombination and repair pathways of *dnaA* and *ruvA* as well as cell division pathways *mreB*, *parA* and *ftsA* were also found to be essential. Multiple pathways involved in LPS, CPS, fatty acid and murein biosynthesis, including *lpxA*, *rmlA*, *manA*, *fabD* and *murA* were also judged to be essential, as well as genes involved in secretion and chaperone pathways such as *secD*, *groES/EL* and *surA*. Pathways involving *nrfA*, *etfA*, *sufB*, *nadD* and *ribE* associated with oxidation-reduction reactions, were found to be essential in *P. gingivalis*, which is not surprising for an anaerobic bacterium. Major metabolic pathways *purA*, *pyrB*, *coaA*, *accB*, *pdxA*, *ispA*, *thiF*, *serA* and *dapA*, which encode nucleotide, amino acid and co-factor building blocks, respectively, were determined essential under our *in vitro* conditions. All of the aforementioned systems and pathways have previously been identified as essential for *in vitro* growth of other bacterial species, which is not unexpected given that replication, transcription, translation, cell division, membrane stability and central metabolism are key to survival (Zhang & Lin, 2009).

Hypothetical genes are an often-overlooked group within any bacterial genome, including those of 'model' organisms. In our study we determined that approximately one-tenth of the essential genome of *P. gingivalis* encoded 'hypothetical' proteins, a few of which were homologous to other hypotheticals contained within the DEG. The majority of essential hypothetical genes are large and not within operons, suggesting that they encode functional proteins and are not essential due to a polar effect on a downstream essential gene. The finding that certain hypothetical proteins are essential will stimulate the search for protein motifs, structural bioinformatic and spatial organizational data and studies to define their function.

Although the notion that an essential gene within a given strain is likely to be essential in the species as a whole, intraspecies differences are known and often result in different phenotypes. For example, in strain ATCC 33277 we found no insertions into *ragA* and thus this gene was considered essential. Previous investigators also had difficulty making directed knockouts of *ragA* in strain W50; however, these investigators were successful in deleting *ragA* from strain WPH35. It is possible that *ragA* is only essential within specific strains and those strains in which it is non-essential compensate for loss of its function through the presence of other genes.

2.3.7 Limitations of essential gene analysis

Limitations to essential gene studies should be addressed regardless of systems and methods utilized for their identification. First, several studies have relied exclusively on *in silico* bioinformatic analyses to determine essentiality. These analyses were based on programs designed to combine information from previous *in vitro* and *in vivo* mutagenesis studies with genome annotation and composition scripts without having carried out actual

mutagenesis studies. Thus, any limitations of these experimental studies will be carried over into the new analyses and magnified by any inaccuracies of the program design itself. Second, in insertional mutagenesis methods to determine gene essentiality, genes may be misidentified as essential due to transposon insertion 'cold-spots'. There is no ideal transposon identified as yet that completely lacks any nucleotide specificity and which can create completely random and saturating mutant libraries. Thus, no matter what type of transposon is used, *Tn5*, *Tn7*, *Tn10*, a cryptic construct or Mariner, all studies will have regions of the genome where fewer insertions occur. Third, genes that are actually non-essential but when mutated cause severe growth defects may be scored as essential due to practical limits to the depth of sequencing of transposon insertion junctions. These 'sick' mutants could potentially be represented at levels below 1000-fold a neutral mutant due to the number of replications it could go through prior to being pooled from mutagenesis plates into the library. Fourth, non-essential genes immediately upstream of and co-transcribed with essentials may be incorrectly scored as essential due to polarity of the transposon insertion. Last, practical limits to library complexity can result in some genes that fail to get disrupted by the transposon and so are misidentified as essential. This is particularly a problem for small genes or genes that are within cold spots for the transposon. Several studies, based mostly on the genome size of the species under investigation and the type of transposon, have attained different levels of saturation prior to analyses for essential genes. The possible limitation of our library when combining the type of transposon and library complexity relates to genes that contain less than 10 TA sites in their coding-sequences. Of the 204 genes with fewer than 10 TA sites, 60 could potentially be scored as essential based on having zero insertions, but do not qualify, given our stringent criteria (Additional file 5: Table S5).

Adding confidence to the notion that many of these are non-essential is that 24 of the 60 genes encode proteins of less than 35 amino acids in length. Since these are all characterized as 'hypothetical' and are rather short to encode functional proteins, we believe that some of these may simply be artifacts of annotation programs and thus not true protein-coding genes.

Even complete gene deletion, non-transposon based studies of essential genes have limitations. The Keio collection of single and double gene deletions in *Escherichia coli* is considered the most comprehensive essential gene study to date (Baba et al., 2006) (Yamamoto et al., 2009). Genes that could not be deleted were scored as essential, however, failure to delete a gene is not a guarantee of essentiality and there a are few genes identified as essential in the Keio collection that were successfully deleted by other laboratories. Furthermore, a handful of genes labeled non-essential were actually essential. The Keio deletions of those genes have second site suppressor mutations that compensate for the loss of the essential gene.

The best understanding of essential genes is likely to come from combining different modalities to confirm their essential nature and comparison of these databases both within and between species.

2.4 Conclusions

We have described a method for performing Tn-seq with *P. gingivalis* using a Mariner mini-transposon and Illumina platform next-generation sequencing. We have also invented a new method for creating sequencing libraries from bacterial transposon libraries that has many advantages compared to previous methodologies. Using that method we identified specific mutants quantitatively in a highly reproducible manner using massively-parallel sequencing techniques. We used a near saturating insertion library generated in *P.*

gingivalis strain ATCC 33277 to define the set of genes essential for growth on blood agar. Both the availability of a *P. gingivalis* mutant library and the ability to screen quantitatively are a marked advance in genetic and molecular tools for future studies of *P. gingivalis* biology and pathogenesis. By applying different selective pressures to the library, it is now possible to identify genes critical for survival and growth under different conditions. Due to the quantitative nature of results provided by Tn-seq, both positive and negative gene effects, including partial phenotypes, can now be readily identified.

2.5 Methods

2.5.1 Bacterial strains and plasmids

P. gingivalis ATCC 33277 (RefSeq NC_010729.1) was obtained from the ATCC. *E. coli* S17-1 λ*pir* and plasmids pSAM, containing *bla*, and pSAM_Bt, containing *bla*, *ermG* and *himar1c9a* genes were obtained from Dr. Andrew Goodman.

2.5.2 Media and culture conditions

P. gingivalis strain ATCC 33277 was grown and maintained at 37°C under anaerobic conditions using the GasPak™ EZ Anaerobe Pouch System (BD Biosciences). Blood agar plates (BAPHK) containing trypticase soy agar supplemented with defibrinated sheep's blood (5% vol/vol), hemin (5 μg/ml), and menadione (0.5 μg/ml) as well as brain-heart infusion broth (BHIHK$S_{bc}S_{tg}$C) containing brain-heart infusion, yeast extract (1 mg/ml), hemin (5 μg/ml), and menadione (0.5 μg/ml), sodium bicarbonate (1 mg/ml), sodium thioglycolate (0.25 mg/ml), and cysteine (0.5 mg/ml) were used for solid and liquid culture of *P. gingivalis,* respectively. Gentamicin (25–50 mg/ml) and erythromycin (2–10 mg/ml) were used when appropriate for prevention of contamination as well as isolation and

maintenance of *P. gingivalis* mutants.

Escherichia coli strain S17-1 λpir/pSAM_Bt was grown at 37°C under aerobic conditions in Luria broth base (LB) and Luria agar (BD Biosciences). Carbenicillin (50 μg/ml) was added for plasmid maintenance and prevention of contamination. *E. coli* S17-1 λpir contains the *pir* gene and has chromosomally integrated conjugational transfer functions (RP4/RK6) such that bi-parental mating can take place in lieu of tri-parental mating using helper strains.

2.5.3 Transposon mutagenesis

P. gingivalis Mariner-based transposon mutagenesis was carried out as follows. Wild-type *P. gingivalis (*strain ATCC 33277) was inoculated into brain-heart infusion broth (BHIHK$S_{bc}S_{tg}$C) without antibiotics. Broth cultures were grown to optical densities (OD_{600}) between 0.50 and 1.00. *Escherichia coli* strain S17-1 λ*pir* containing the pSAM_Bt plasmid was grown to optical densities OD 0.50 - 1.00. Broth cultures were set up such that between a 5:1 and 10:1 ratio of *P. gingivalis* (recipient) to *E. coli* (donor) was achieved. Although *P. gingivalis* is categorized as an obligate anaerobe it is able to survive without significant CFU loss (less than a $\log_{10}$) for up to 6 hours under aerobic conditions when incubated alone on BAPHK at 37°C.

The *E. coli* donor strain carrying the Mariner transposon on a suicide plasmid vector was conjugated with wild-type *P. gingivalis* using a bi-parental procedure where the *E. coli* donor strain and *P. gingivalis* recipient strain are cultured together on an agar plate to allow for plasmid transfer. Conjugation was carried out aerobically at 37°C for 5 hr. As *P.*

gingivalis is naturally resistant to gentamicin, this antibiotic was used for selection against the donor *E. coli* following the conjugation. The transposon contains an erythromycin resistance gene (*ermG*) used to select for *P. gingivalis* transposon insertion mutants.

Individual *P. gingivalis* colonies were tested using PCR to detect the presence of the transposon as well as vector backbone components (*bla* and *himar1C9a* transposase). These tests indicated that few to no mutants contained incorrect/unwanted transpositions; vector insertions containing additional portions of the transposon-containing plasmid. Nested semi-random PCR and direct sequencing of the transposon-chromosome junction of individual mutants determined that the transpositions appear randomly distributed throughout the *P. gingivalis* chromosome.

2.5.4 PCR

Mutant colonies were initially isolated directly from transposition plates (post-conjugation) and sub-cultured under antibiotic selection on blood agar plates anaerobically three times to ensure purity. Genomic DNA was prepared using a DNeasy (Qiagen) kit as per the manufacturer's instructions. PCR was performed using primers to detect the presence of *ermG*, *bla* (amp^R) and *himar1C9a* genes as detailed in Additional file 6: Table S6.

PCR using semi-random priming was performed in order to determine the chromosomal location of the transposon insertion. Nested PCR primers directed to the pSAM_Bt vector (the remaining flanking regions of the transposon) were used in sequential PCR reactions and the final sequencing reaction. Primers are detailed in Additional file 6: Table S6.

Sequencing of the PCR amplicons was carried out at Tufts University Core Facilities

using ABI 3130XL DNA sequencers.

2.5.5 Construction and sequencing of libraries

Genomic DNA eluted in 100 µL elution buffer (Qiagen) was placed in a 2 mL microfuge tube and sheared for 2 minutes (10 sec on and 5 sec off duty cycle, 100% intensity) using a high intensity cup horn that was cooled by a circulating bath (4°C) and was attached to a Branson 450 sonifier. C-tails were then added to 1 µg of sheared DNA in a 20 µL reaction that contained 0.5 µL TdT enzyme (Promega), 4 µL 5x TdT reaction buffer (Promega), 475 µM dCTP and 25 µM dideoxy CTP. The dideoxy CTP functions as a chain terminator to limit the length of the poly-C tails. Following a 1-hour incubation at 37°C and a 20 minute heat-inactivation step at 75°C, dideoxy CTP and other small molecules were removed using a Performa gel filtration cartridge (Edge Biosystems). Transposon containing fragments were then amplified in a 50 µL PCR reaction that contained 5 µL C-tailed template, 600 nM C tail-specific primer (olj376 5′ GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGGGGGGGGGGGGG 3′), 600 nM transposon-specificprimer (pSAM1 5′ CCTGACGGATGGCCTTTTTGCGTTTCTACC 3′), 400 µM dNTPs, 5 µL 10x buffer, and 1 µL Easy-A DNA polymerase mix (Agilent). Sandwiched by an initial incubation at 95°C for 120 sec and a final extension of 120 sec at 72°C, 24 cycles were completed using 30 sec denaturation steps at 95°C, 30 sec annealing steps at 60°C, and 120 sec extension steps at 72°C. A second PCR reaction was then used to amplify the exact transposon-genomic DNA junction and add additional sequences needed for Illumina sequencing and indexing. This 50 µL reaction contained 1 µL of template from PCR #1, 600 nM transposon end-specific

primer (pSAM2 5′ AATGATACGGCGACCACCGAGATCTACACTCTTTGACCGGGGACTTATCATCCA ACCTGTTA 3′), 600 nM indexing primer (5′ CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCT CTTCCGATCT 3′, where NNNNNN represents the reverse complement of the index and varied with each sample), 400 μM dNTPs, 5 μL 10x buffer, and 1 μL easy-A DNA polymerase mix (Agilent). Sandwiched by an initial incubation at 95°C for 120 sec and a final extension of 120 sec at 72°C, 12 cycles were completed using 30 sec denaturation steps at 95°C, 30 sec annealing steps at 60°C, and 120 sec extension steps at 72°C. Libraries were then pooled and run for 51 cycles in a single end sequencing reaction on a single lane of an Illumina Genome Analyzer II using the custom sequencing primer pSAM3 (5′ ACACTCTTTGACCGGGGACTTATCATCCAACCTGTTA 3′) and the standard Illumina index sequencing primer.

2.5.6 Data analysis

All read mapping and primary data analysis was done on the Tufts University Galaxy server. Approximately seven-percent of all sequencing reads contained multiple 'C' nucleotides at their 3' end as a consequence of the C-tailing reaction. These C-tails were removed using the "clip adapter sequences script" with the 3' adapter set to CCCCCCCCCCCCCCCCCCCCCCCCCC and the minimum read length set to 26. The resulting clipped reads were aligned to the *P. gingivalis* strain ATCC 33277 and W83 reference genomes, accession numbers AP009380.1 and AE015924.1 respectively, using Bowtie with its default settings. The resulting bowtie output file was then used as input for a

custom script, “hopcount”. Hopcount tabulates the number of times individual insertion sites in the genome were re-sequenced. An Excel spreadsheet file is generated that indicates, for each insertion site, its position in the genome, gene locus to which that position maps, the strand (positive vs. negative) associated with the site as well as the frequency of its reads. Hopcount output was used to estimate the complexity of transposon libraries and to compare the fate of specific insertions sites in input and output samples. It was also used as input for a second custom script, “aggregate hop table”. The output of this script is an excel file in which all transposon insertion sites are tabulated by their collective frequency in each annotated gene of the genome. For each gene, the number of unique insertions sites observed, absolute count of sites in the positive strand, in negative strand and in both strands is recorded. Also recorded is the normalized value dvalgenome, which is an indication of whether the number of insertions observed in that gene is above or below the expected frequency. Dvalgenome equals the observed number of insertions in a gene / predicted number of insertions for that gene and the predicted number of insertions (size of that gene in base pairs divided by size of genome in base pairs) multiplied by (total number of insertions counted).

2.5.7 Bioinformatics resource for oral pathogens

Microbial Transcriptome Database, MTD (http://bioinformatics.forsyth.org/mtd/) maintained by the Forsyth Institute in Cambridge, MA, USA, was utilized for comparing the putative essential genes of *P. gingivalis* ATCC 33277 to that of the RNA-seq transcriptome information detailing gene expression when grown on blood agar medium (from strain W83) (Chen, Abbey, Deng, & Cheng, 2005).

P. gingivalis essential genes were determined to have DEG homologues based strictly

on BLASTP similarity. BLAST similarities at protein-protein level that resulted in e-values of $1x10^{-8}$ or less were considered matches. Pfam (http://pfam.sanger.ac.uk/), Welcome Trust Sanger Institute, Prosite (http://prosite.expasy.org/), Swiss Institute of Bioinformatics and Interproscan (http://www.ebi.ac.uk/Tools/pfa/iprscan/), European Bioinformatics Institute, protein information databases/platforms were utilized to query all *P. gingivalis* essential genes for functional motifs (including signal sequences) and post-translational/co-translational modification sites (Punta et al., 2012) (Quevillon et al., 2005) (Artimo et al., 2012). In cases where genes were previously described as un-annotated hypothetical proteins but were now found to have functional motifs they have added them to the analyses and lists. National Center for Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov/, National Institutes of Health, USA was used for gathering genome information on *Porphyromonas gingivalis* stains ATCC 33277 and W83. The Kyoto Encyclopedia of Genes and Genomes, KEGG (http://www.genome.jp/kegg/), database contains genetic information on all three of the sequenced and annotated strains of *P. gingivalis* (ATCC 33277, W83 and TDC60). The entry number for *P. gingivalis* 33277 is T00714, which is the reference genome used to this study. All *P. gingivalis* 'essential' genes were examined for KEGG-described functional characterizations through the T00714 KEGG gene list. In cases where genes were previously described as un-annotated hypothetical proteins but were now found to have functional motifs they have added them to the analyses and lists. Bioinformatics Resource for Oral Pathogens, BROP (http://www.brop.org/) maintained by the Forsyth Institute in Cambridge, MA, USA, was utilized for *P. gingivalis* genome annotation, comparison between annotations by NCBI, BROP, TIGR and Los Alamos National Laboratories, operon structure analysis and BLAST (Basic Local Alignment Search Tool) of nucleotide and

protein sequences between oral bacteria (Chen et al., 2005).

Additional file 1. Table S1. *P. gingivalis* strain ATCC 33277 essential gene list; functional characterization and bioinformatic analyses. Green highlight denotes being found in the DEG as essential in other genomes, blue highlight denotes *Bacteroides thetaiotaomicron* (Goodman *et al*. Cell Host & Microbe 2009) but NOT DEG essential gene match and orange highlight denotes being part of *P. gingivalis* core genome (Brunner *et al*. BMC Microbiology 2010).

Locus (ATCC 33277)	Locus (W83)	Polar	ProteinID	Abbreviation
PGN_0001	PG0001		Chromosome replication initiator DnaA	*dnaA*
PGN_0023	PG0027	Possibly	Por secretion system protein PorV	*lptO/porV*
PGN_0024	PG0028		2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	*ispF*
PGN_0033	PG0034		Thioredoxin	*txn/trx*
PGN_0034	PG0035		DNA polymerase III alpha subunit	*dnaE*
PGN_0042	PG0046		Phosphatidate cytidylyltransferase	*cdsA*
PGN_0043	PG0047	Possibly	Cell division protein FtsH	*ftsH*
PGN_0096	PG2189		Aspartate kinase	*ask/lysC*
PGN_0097	PG2190	Possibly	Cell-division ATP-binding protein	*ftsE*
PGN_0137	PG2085		Tryptophanyl-tRNA synthetase	*trpS*
PGN_0138	PG2086		Hypothetical protein	
PGN_0143	PG2091		Dihydroneopterin aldolase	*folB*
PGN_0149	PG2097		Ribose-phosphate pyrophosphokinase	*prsA*
PGN_0158	PG2108		Thiazole synthase	*thiG*
PGN_0172	PG2123	Possibly	Hypothetical protein	
PGN_0173	PG2124		Glyceraldehyde 3-phosphate dehydrogenase type I	*gapA*
PGN_0189	PG2141		3-oxoacyl-ACP synthase	*fabH*
PGN_0190	PG2142		GTP-binding protein Era	*era*
PGN_0191	PG2143		GTP-binding protein EngA	*engA*
PGN_0192	PG2144		Hypothetical protein	
PGN_0198	PG2150		LysM domain-containing protein	
PGN_0202	PG2157		Leucine aminopeptidase precursor	*pepA*
PGN_0203	PG2158		Sulfur Acceptor protein SufE	*sufE*
PGN_0206	PG2162		Lipid A disaccharide synthase	*lpxB*
PGN_0207	PG2163	Possibly	Stationary phase survival protein SurE	*surE*
PGN_0208	PG2164	Possibly	FKBP-type peptidylprolyl isomerase	
PGN_0209	PG2165		Glycyl-tRNA synthetase	*glyS*
PGN_0242	PG0129		Glycosyl transferase family 1	
PGN_0243	PG0130		Phosphoglyceromutase	*gpmA*
PGN_0247	PG0134		Magnesium transporter	
PGN_0248	PG0135		Dimethyladenosine transferase	*ksgA*
PGN_0249	PG0136		Hypothetical protein / membrane protein	
PGN_0250	PG0137	Possibly	Aminoacyl-histidine dipeptidase	*pepD*
PGN_0251	PG0138		Malonyl CoA-ACP transacylase	*fabD*
PGN_0253	PG0140		Conserved hypothetical protein	
PGN_0254	PG0141		Chromosome partitioning protein ParB	*parB*
PGN_0255	PG0142		Chromosome partitioning protein ParA	*parA*
PGN_0256	PG0143	Possibly	Hydrolase / amidase	
PGN_0258	NA		Preprotein translocase subunit SecG	*secG*
PGN_0260	PG0147		Lipoprotein	
PGN_0261	PG0148		Sigma-54-dependent transcriptional regulator	
PGN_0264	PG0151		Signal recognition particle-docking protein	*ftsY*

PGN_0266	PG0153		Aspartyl-tRNA synthetase	*aspS*
PGN_0267	PG0155		Riboflavin biosynthesis protein	*ribD*
PGN_0272	PG0160		Hypothetical protein	
PGN_0278	PG0166		Peptidyl-tRNA hydrolase	*spoVC*
PGN_0279	PG0167		50S ribosomal protein L25	*rplY*
PGN_0281	PG0170		Methionyl-tRNA synthetase	*metS*
PGN_0293	PG0185		Receptor antigen A	*ragA*
PGN_0297	PG0189		Membrane protein	
PGN_0298	PG0190		Undecaprenyl pyrophosphate synthase	*ispU*
PGN_0299	PG0191		Outer membrane protein	
PGN_0300	PG0192		Cationic Outer membrane protein OmpH	*ompH*
PGN_0301	PG0193		Cationic Outer membrane protein OmpH	*ompH*
PGN_0309	PG0201		Ribonuclease P	*rnpA*
PGN_0310	PG0202		Uroporphyrinogen-III synthase	*hemD*
PGN_0311	PG0203	Possibly	Hypothetical protein	
PGN_0353	PG0253		Ribosome maturation factor	*rimP*
PGN_0354	PG0254		Transcription elongation factor NusA	*nusA*
PGN_0355	PG0255		Translation initiation factor IF-2	*infB*
PGN_0357	PG0257		Cysteine desulfurase	*sufB*
PGN_0358	PG0258		ABC transporter ATP-binding protein	*sufC*
PGN_0359	PG0259		ABC transporter permease protein	*sufD*
PGN_0360	PG0263		Tyrosyl-tRNA synthetase	*tyrS*
PGN_0365	PG0267		Arginyl-tRNA synthetase	*argS*
PGN_0366	PG0268		tRNA-specific 2-thiouridylase MnmA	*mnmA*
PGN_0369	PG0271	Possibly	Single-stranded binding protein	*ssb*
PGN_0370	PG0272	Possibly	Putative transporter CBS domain-containing protein	
PGN_0371	PG0273		4'-phosphopantetheinyl transferase superfamily	*acpS*
PGN_0374	PG0276		Hypothetical protein	
PGN_0376	PG1743		2-dehydro-3-deoxyphosphooctonate aldolase	*kdsA*
PGN_0393	PG1724		DNA-binding/iron metalloprotein/AP endonuclease or O-sialoglycoprotein endopeptidase	
PGN_0398	PG1719		ABC transporter ATP-binding protein MsbA family	
PGN_0403	PG1714		Pyridoxamine-phosphate oxidase	*pdxH*
PGN_0408	PG1707		Septum formation initiator family	
PGN_0413	PG1702		DNA gyrase subunit B	*gyrB*
PGN_0418	PG1694	Possibly	YceG family protein / aminodeoxychorismate lyase	*pabC*
PGN_0419	PG1693	Possibly	Dinucleotide-utilizing enzyme involved in molybdopterin and thiamine biosynthesis	*thiF*
PGN_0420	PG1692		Lipoprotein releasing system ATP-binding protein	*lolD*
PGN_0433	PG1677		Phosphoglycerate kinase	*pgk*
PGN_0449	PG1662		Hypothetical protein	
PGN_0472	PG1622		DNA topoisomerase IV subunit A	*parC*
PGN_0473	PG1623		Membrane bound regulatory protein	
PGN_0487	PG1636		DNA translocase FtsK	*ftsK*
PGN_0499	PG1613		Glyoxalase / methylmalonyl-CoA epimerase	
PGN_0500	PG1612		Methylmalonyl-CoA decarboxylase alpha subunit	*accD*
PGN_0501	PG1611		Hypothetical protein	
PGN_0502	PG1610		Hypothetical protein	
PGN_0503	PG1609		Biotin carboxyl carrier protein	*accB*
PGN_0510	PG1603		Deoxyribonucleoside-triphosphatase	
PGN_0511	PG1602		Hypothetical protein	*mmcQ*
PGN_0512	PG1601		Biotin--acetyl-CoA-carboxylase ligase	*birA*
PGN_0515	PG1598		Lipoprotein signal peptidase	*lspA*

PGN_0516	PG1597		DnaK suppressor protein	
PGN_0517	PG1596		Isoleucyl-tRNA synthetase	*ileS*
PGN_0518	PG1595		Ribulose-phosphate 3-epimerase	*rpe*
PGN_0522	PG1589		Dihydropteroate synthase	*folP*
PGN_0523	PG1588		Conserved hypothetical protein	
PGN_0540	PG1570		Rhodanese-like domain-containing protein / phage shock protein E	*pspE*
PGN_0543	PG1966		Glutamyl-tRNA synthetase	*gltS*
PGN_0544	PG1565		3-deoxy-D-manno-octulosonic-acid transferase	*waaA*
PGN_0546	PG1563		Glucose-1-phosphate thymidylyltransferase	*rmlA*
PGN_0547	PG1562		dTDP-4-dehydrorhamnose 3,5-epimerase	*rmlC*
PGN_0548	PG1561		dTDP-4-dehydrorhamnose reductase	*rmlD*
PGN_0549	PG1560		dTDP-glucose 4,6-dehydratase	*rmlB*
PGN_0568	PG1541		2-amino-4-hydroxy-6- hydroxymethyldihydropteridine pyrophosphokinase	*folK*
PGN_0569	PG1540		S-adenosylmethionine:tRNA ribosyltransferase-isomerase	*queA*
PGN_0572	PG1537		Membrane protein	
PGN_0573	PG1536		Cell division protein FtsX	*ftsX*
PGN_0610	PG1280		HTH domain of SpoOJ/ParA/ParB/RepB family protein	
PGN_0611	PG1279		D-3-phosphoglycerate dehydrogenase	*serA*
PGN_0612	PG1278		Phosphoserine aminotransferase	*serC*
PGN_0618	PG0571		Aspartate-semialdehyde dehydrogenase	*asd*
PGN_0620	PG0573		S-adenosyl-methyltransferase MraW	*mraW*
PGN_0621	PG0574	Possibly	Hypothetical protein	
PGN_0622	PG0575		Penicillin-binding protein	*pbp*
PGN_0623	PG0576		UDP-N-acetylmuramoylalanyl-D-glutamate--2, 6-diaminopimelate ligase	*murE*
PGN_0624	PG0577		Phospho-N-acetylmuramoyl-pentapeptide- transferase	*murX*
PGN_0625	PG0578		UDP-N-acetylmuramoylalanine--D-glutamate ligase	*murD*
PGN_0626	PG0579		Rod shape-determining protein RodA	*ftsW*
PGN_0627	PG0580		Undecaprenyldiphospho-muramoylpentapeptide beta-N- acetylglucosaminyltransferase	*murG*
PGN_0628	PG0581		UDP-N-acetylmuramate--L-alanine ligase	*murC*
PGN_0629	PG0582		Cell division protein FtsQ	*ftsQ*
PGN_0630	PG0583		Cell division protein FtsA	*ftsA*
PGN_0635	PG0589		GMP synthase	*guaA*
PGN_0638	PG0594		RNA polymerase sigma factor RpoD	*rpoD*
PGN_0639	PG0595		30S ribosomal protein S6	*rpsF*
PGN_0642	PG0598		Permease	
PGN_0643	PG0599		3,4-dihydroxy-2-butanone 4-phosphate synthase	*ribA*
PGN_0645	PG0602		Por Secretion system protein PorQ	*porQ*
PGN_0646	PG0603		cytidylate kinase	*cmk*
PGN_0647	PG0604		4-hydroxy-3-methylbut-2-enyl diphosphate reductase	*ispH*
PGN_0662	PG0620		ATP-dependent protease La	*La*
PGN_0663	PG0621		Hypothetical protein	
PGN_0664	PG0622		Hypothetical protein	
PGN_0665	PG0623		Triosephosphate isomerase	*tpiA*
PGN_0666	PG0624		Sporulation domain-containing protein	
PGN_0667	PG0625		GTP cyclohydrolase I	*folE*
PGN_0669	PG0628		ABC transporter ATP-binding protein	*lptB*
PGN_0670	PG0629		Inorganic polyphosphate/ATP-NAD kinase	*nadK*
PGN_0671	PG0630		Pyridoxine 5'-phosphate synthase	*pdxJ*
PGN_0679	PG0638		Tetraacyldisaccharide 4'-kinase	*lpxK*
PGN_0680	PG0639		Signal peptide peptidase SppA 67K type	*sppA*
PGN_0713	PG0677		Saccharopine dehydrogenase	*sdh*

PGN_0717	PG0681		Hypothetical protein	
PGN_0723	PG0687		Succinate-semialdehyde dehydrogenase	*sdhA*
PGN_0726	PG0691		NifU-like protein	
PGN_0734	PG0700		Hypothetical protein	
PGN_0759	PG0728		Hypothetical protein	
PGN_0760	PG0729		D-alanyl-alanine synthetase A	*ddl*
PGN_0761	PG0730		Ribosomal large subunit pseudouridine synthase	*rluD*
PGN_0762	PG0731		PASTA domain protein	
PGN_0764	PG0733		Riboflavin synthase subunit RibE	*ribE*
PGN_0765	PG0734		Nitroreductase	
PGN_0766	PG0735		Cysteine desulfurase/Selenocysteine lyase	*nifS*
PGN_0777	PG0750		Glycosyl transferase (group 2)	
PGN_0781	PG0754		DNA topoisomerase I	*topA*
PGN_0799	PG0775		Acyl-CoA dehydrogenase	*fadE*
PGN_0800	PG0776		Electron transfer flavoprotein alpha subunit	*etfA*
PGN_0801	PG0777		Electron transfer flavoprotein beta subunit	*etfB*
PGN_0803	PG0779		ExbD	*exbD*
PGN_0804	PG0780		ExbD	*exbD*
PGN_0805	PG0781		Mebrane protein	
PGN_0806	PG0782		MotA/TolQ/ExbB proton channel protein	*exbB*
PGN_0809	PG0785		TonB protein	
PGN_0813	PG0790		GTPase ObgE	*obgE*
PGN_0814	PG0791	Possibly	Adenylate kinase	*adk*
PGN_0815	PG0792	Possibly	Hypoxanthine phosphoribosyltransferase	*hpt*
PGN_0816	PG0793		Fructose-1,6-bisphosphatase	*fbp*
PGN_0817	PG0794		Penicillin-binding protein 1A	*pbp*
PGN_0819	PG0796		Leucyl-tRNA synthetase	*leuS*
PGN_0820	NA		Flavoprotein	
PGN_0824	PG0801		tRNA nucleotidyltransferase	*rph*
PGN_0827	PG0803		Glucosamine-6-phosphate deaminase	*nagB*
PGN_0828	PG0804		Flavodoxin	
PGN_0829	PG0805	Possibly	Prolipoprotein diacylglyceryl transferase	*lgt*
PGN_0830	PG0806	Possibly	Meso-diaminopimelate D-dehydrogenase	
PGN_0831	PG0807		Nitrogen utilization substance protein NusB-family	*nusB*
PGN_0833	PG0811		Holliday junction DNA helicase RuvA	*ruvA*
PGN_0841	PG0028		2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	*ispD*
PGN_0851	NA		PcfK-like protein	
PGN_0865	PG1397		Bifunctional phosphoribosylaminoimidazolecarboxamidesformyltransferase	*purH*
PGN_0866	PG1396		Rod shape-determining protein MreB	*mreB*
PGN_0867	PG1395		Rod shape-determining protein MreC	*mreC*
PGN_0868	PG1394		Rod shape-determining protein MreD	*mreD*
PGN_0870	PG1392		Rod shape-determining protein RodA	*rodA*
PGN_0875	PG1386		DNA gyrase A subunit	*gyrA*
PGN_0883	PG1404		Rhomboid family protein	
PGN_0884	PG1405		Organic solvent tolerance protein OstA	*ostA*
PGN_0894	PG1418		DNA polymerase III, gamma and tau subunits	*dnaX*
PGN_0901	PG1428		6,7-dimethyl-8-ribityllumazine synthase	*ribH*
PGN_0902	PG1430		TPR domain-containing protein	
PGN_0907	PG1063		Transcriptional regulator	
PGN_0916	PG1208		Molecular chaperone DnaK	*dnaK*
PGN_0919	NA		Hypothetical protein	

PGN_0962	PG0992		Threonyl-tRNA synthetase	*thrS*
PGN_0963	PG0991		Translation initiation factor IF-3	*infC*
PGN_0964	PG0990		50S ribosomal protein L35	*rpmI*
PGN_0965	PG0989		50S ribosomal protein L20	*rplT*
PGN_0970	PG0985		RNA polymerase sigma-70 factor ECF subfamily	*rpoD*
PGN_0984	PG0965		Phosphatidylserine decarboxylase	*psd*
PGN_0985	PG0964		Phosphatidylserine synthase	*pssA*
PGN_0987	PG0962		Prolyl-tRNA synthetase	*proS*
PGN_0988	PG0961		Hypothetical protein	
PGN_0992	PG0957		Riboflavin biosynthesis protein	*ribF*
PGN_0993	PG0956		Peptidase M23/M37 family	
PGN_0994	PG0955		Hypothetical protein	
PGN_0997	PG0953		Deoxyuridine 5'-triphosphate nucleotidohydrolase	*dut*
PGN_0998	PG0952		4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	*lspG*
PGN_0999	PG0951		Phosphoribosylaminoimidazole carboxylase	*purE*
PGN_1000	PG0950	Possibly	Glycine cleavage system protein H	*gcvH*
PGN_1001	PG0932		DNA polymerase III, delta subunit	*holA*
PGN_1002	PG0948		AMP nucleosidase	*amn*
PGN_1003	PG0947		Type I restriction enzyme R protein	*hsdR*
PGN_1005	PG0945		ABC transporter permease protein	
PGN_1010	PG0937		Hypothetical protein	
PGN_1011	PG0936		Xanthine/uracil/vitamin C permease	
PGN_1012	PG0935		4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	*ispE*
PGN_1019	PG0928		Response regulator	
PGN_1020	PG0927		ATP/GTP-binding transmembrane protein	
PGN_1022	PG0925		Thymidine kinase	*tdk*
PGN_1023	PG0924		Acid phosphatase OlpA / lipoprotein / 5'-nucleotidase	*olpA*
PGN_1024	PG0923		Ribosome-binding factor A	*rbfA*
PGN_1025	PG0922		Membrane protein	
PGN_1026	PG0920		Glycosyl transferase family 2	
PGN_1033	PG0912		Polysaccharide transport protein	
PGN_1038	PG0903		Arginine decarboxylase, pyruvoyl-dependent	*adiA*
PGN_1058	PG0880		Bacterioferritin comigratory protein	*bcp*
PGN_1078	PG1288		GDP-mannose 4,6-dehydratase	*gmd*
PGN_1079	PG1289		GDP-fucose synthetase	*wcaG*
PGN_1129	PG1341		Lipoprotein	
PGN_1130	PG1342		UDP-N-acetylenolpyruvoylglucosamine reductase	*murB*
PGN_1134	PG1345		Glycoside hydrolase family protein	
PGN_1151	PG1364		1-deoxy-D-xylulose 5-phosphate reductoisomerase	*ispC*
PGN_1152	PG1365		16S rRNA-processing protein RimM	*rimM*
PGN_1153	PG1366		UDP-N-acetylglucosamine 1-carboxyvinyltransferase	*murA*
PGN_1154	PG1367	Possibly	Methionyl-tRNA formyltransferase	*fmt*
PGN_1155	PG1368		Glucose-6-phosphate isomerase	*pgi*
PGN_1156	PG1369		Glycerol-3-phosphate dehydrogenase	*gspA*
PGN_1157	PG1370		Lysyl-tRNA synthetase	*lysS*
PGN_1173	PG1077		Electron transfer flavoprotein beta subunit	*etfB*
PGN_1178	PG1081		Acetate kinase	*ackA*
PGN_1179	PG1082		Phosphotransacetylase	*eutD*
PGN_1187	PG1091		DHH subfamily 1 protein	
PGN_1188	PG1093		Hypothetical protein	
PGN_1189	PG1094		Phosphoglucomutase/phosphomannomutase	*manB*

PGN_1194	PG1097		Bifunctional UDP-N-acetylmuramoyl-tripeptide:D-alanyl-D-alanine ligase	*alr*
PGN_1202	PG1105		RNA polymerase sigma-54 factor	*rpoN*
PGN_1203	PG1106		UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D- alanine ligase	*murF*
PGN_1204	PG1114		Aspartate alpha-decarboxylase	*panD*
PGN_1205	PG1115		Signal recognition particle protein	*ffh*
PGN_1215	NA		Hypothetical protein	
PGN_1218	PG1121		Asparaginyl-tRNA synthetase	*asnS*
PGN_1219	PG1122		Ribosomal large subunit pseudouridine synthase B	*rluB*
PGN_1220	PG1123		Adenylosuccinate lyase	*purB*
PGN_1222	PG1125		Hypothetical protein	
PGN_1223	PG2148		Uracil permease	*uraA*
PGN_1229	PG1132		Valyl-tRNA synthetase	*valS*
PGN_1232	PG1134		Thioredoxin reductase	*trxB*
PGN_1240	PG1141		Glycosyl transferase, group 1 family protein	
PGN_1241	NA		Hypothetical protein	
PGN_1242	PG1142		Exopolysaccharide synthesis-like protein	
PGN_1244	PG1144		Peptide chain release factor 2	*prfB*
PGN_1250	NA		Hypothetical protein	
PGN_1251	PG1149		Glycosyl transferase	
PGN_1301	PG1052		Transcriptional regulator	
PGN_1312	PG1040		Transcriptional regulator	
PGN_1315	PG1037		Hypothetical protein	
PGN_1356	PG1217		RNA polymerase Rpb6	*rpoZ*
PGN_1357	PG1218		Membrane protein	
PGN_1358	PG1219		Transcriptional accessory protein	
PGN_1359	PG1220		Erythronate-4-phosphate dehydrogenase	*pdxB*
PGN_1367	PG1232		Glutamate dehydrogenase	*gdhA*
PGN_1375	PG1239		Beta-ketoacyl-acyl carrier protein reductase	*fabG*
PGN_1376	PG1240		Transcriptional regulator	*tetR*
PGN_1377	PG1241		GTP-binding protein LepA	*lepA*
PGN_1378	PG1242		Replicative DNA helicase	*dnaB*
PGN_1381	PG1246		Alanyl-tRNA synthetase	*alaS*
PGN_1387	PG1252		ABC transporter permease protein	
PGN_1388	PG1253		DNA Ligase	*ligA*
PGN_1389	PG1254		Acetyltransferase	
PGN_1390	PG1255	Possibly	Recombination protein RecR	*recR*
PGN_1391	PG1256		Ribonuclease E	*rne*
PGN_1393	PG1258		DNA-binding protein HU	*hns/hupB*
PGN_1441	PG0531		Glutamine-dependent NAD+ synthetase	*nadE*
PGN_1446	PG0526		Inner membrane protein translocase component YidC	*yidC*
PGN_1447	PG0525		CTP synthetase	*pyrG*
PGN_1449	PG0523		Inosine 5-monophosphate dehydrogenase	*guaB*
PGN_1451	PG0520		Co-chaperonin GroES	*groES*
PGN_1452	PG0521		Molecular chaperone GroEL	*groEL*
PGN_1457	PG0515		Alkaline phosphatase	*phoA*
PGN_1458	PG0514		Preprotein translocase subunit SecA	*secA*
PGN_1459	PG0513		Hypothetical protein	
PGN_1460	PG0512		Guanylate kinase	*gmk*
PGN_1461	PG0511		Spore maturation protein A/B	*spmB*
PGN_1462	PG0510		Metalloprotease	
PGN_1464	PG0508		HAD-superfamily subfamily IB hydrolase / phosphoserine phosphatase	

PGN_1470	PG0502		SsrA-binding protein	*smpB*
PGN_1471	PG0501		Membrane protein	
PGN_1472	PG0500		Queuine tRNA-ribosyltransferase	*TGT*
PGN_1481	PG0489		Polysaccharide biosynthesis protein	
PGN_1482	PG0488		Holliday junction DNA helicase RuvB	*ruvB*
PGN_1484	PG0486		Methylated-DNA-protein-cysteine methyltransferase	*MGMT*
PGN_1485	PG0485		Preprotein translocase subunit YajC	*yajC*
PGN_1486	PG0484	Possibly	YbbR-like protein	
PGN_1487	PG0483		Dephospho-CoA kinase	*coaE*
PGN_1500	PG0469		Hypothetical protein	
PGN_1501	PG0468		Mannose-6-phosphate isomerase	*manA*
PGN_1503	PG0465		Ferric uptake transcriptional regulator	
PGN_1504	PG0464		Adenylosuccinate synthetase	*purA*
PGN_1505	PG0463		Folylpolyglutamate synthase	*folC*
PGN_1510	PG0452		Peptidyl-prolyl cis-trans isomerase	*surA*
PGN_1512	PG0450		Lipoprotein	
PGN_1513	PG0449		TPR domain-containing protein	
PGN_1514	PG0448		Outer membrane protein	
PGN_1515	PG0447		Pantothenate kinase	*coaA*
PGN_1516	PG0446		Molybdopterin biosynthesis MoeB protein	*thiF*
PGN_1518	PG0444		Oligopeptide transporter, OPT family	
PGN_1552	PG0415		Peptidyl-prolyl cis-trans isomerase	*sylD*
PGN_1554	PG0413		Membrane protein	
PGN_1555	PG0412		DNA mismatch repair protein	*mutL*
PGN_1562	PG0403		Hypothetical protein	
PGN_1564	PG0401		Phosphodiesterase	
PGN_1565	PG0400		PSP1 domain protein	
PGN_1566	PG0399		Gliding motility-associated lipoprotein GldH	*gdlH*
PGN_1568	PG0397		Hypothetical protein	
PGN_1570	PG0395		DNA-directed RNA polymerase subunit beta	*rpoC*
PGN_1571	PG0394		DNA-directed RNA polymerase subunit beta	*rpoB*
PGN_1572	PG0393		50S ribosomal protein L7/L12	*rplL*
PGN_1573	PG0392		50S ribosomal protein L10	*rplJ*
PGN_1574	PG0391		50S ribosomal protein L1	*rplA*
PGN_1575	PG0390		50S ribosomal protein L11	*rplK*
PGN_1576	PG0389		Transcription antitermination protein	*nusB*
PGN_1577	NA		Preprotein translocase SecE subunit	*secE*
PGN_1578	PG0387		Elongation factor tu	*tu*
PGN_1587	PG0378		Elongation factor Ts	*tsf*
PGN_1588	PG0377		30S ribosomal protein S2	*rpsB*
PGN_1589	PG0376		30S ribosomal protein S9	*rpsI*
PGN_1590	PG0375		50S ribosomal protein L13	*rplM*
PGN_1593	PG0369		Phosphopantetheine adenylyltransferase	*coaD*
PGN_1594	PG0368		DNA topoisomerase IV subunit B	*parE*
PGN_1599	PG0362		Hypothetical protein	
PGN_1600	PG0361		Hypothetical protein	
PGN_1601	PG0360	Possibly	Conserved hypothetical protein with lemA family domain	
PGN_1602	PG0359		Flavoredoxin	
PGN_1615	PG0346		Ribosome biogenesis GTP-binding protein YsxC	*engB*
PGN_1630	PG0332		Transcription termination factor Rho	*rho*
PGN_1646	PG0316		Seryl-tRNA synthetase	*serS*

PGN_1647	PG0315		50S ribosomal protein L27	*rpmA*
PGN_1648	PG0314		50S ribosomal protein L21	*rplU*
PGN_1650	PG0312		Hypothetical protein	
PGN_1651	PG0311		Glycoslytransferase	
PGN_1652	PG0310		Nitroreductase	
PGN_1653	PG0309		Thiamine biosynthesis lipoprotein ApbE	*apbE*
PGN_1654	PG0308		Electron transport complex RnfABCDGE type A subunit	*nrfA*
PGN_1655	PG0307		Electron transport complex RsxE subunit	*nrfE*
PGN_1656	PG0306		Electron transport complex, RnfABCDGE type, G subunit	*nrfG*
PGN_1657	PG0305		Electron transport complex RnfABCDGE type D subunit	*nrfD*
PGN_1658	PG0304		Electron transport complex, RnfABCDGE type, C subunit	*nrfC*
PGN_1659	PG0303		Ferredoxin / electron transport complex type B subunit	*nrfB*
PGN_1660	PG0302	Possibly	Positive regulator of sigma(E), RseC/MucC superfamily	
PGN_1678	PG0286		Hypothetical protein	
PGN_1688	PG1747		Ribose 5-phosphate isomerase B	*rpiB*
PGN_1689	PG1748		Transketolase	*tktA*
PGN_1695	PG1755		Fructose-1,6-bisphosphate aldolase	*fbaA*
PGN_1698	PG1758		30S ribosomal protein S15	*rpsO*
PGN_1702	PG1762		Bifunctional preprotein translocase subunit SecD/SecF	*secD*
PGN_1704	PG1764		Beta-ketoacyl-acyl-carrier-protein synthase II	*fabF*
PGN_1705	PG1765		Acyl carrier protein	
PGN_1711	PG1771		Phenylalanyl-tRNA synthetase subunit alpha	*pheS*
PGN_1716	PG1776		Molecular chaperone DnaJ	*dnaJ*
PGN_1717	PG1777		Mrp protein homolog	*apbC*
PGN_1718	PG1778		UDP-2,3-diacylglucosamine hydrolase	*lpxH*
PGN_1736	PG1824		Glycogen synthase	*glgA*
PGN_1749	PG1816			*mdaB/kefG*
PGN_1750	PG1815		3-deoxy-manno-octulosonate cytidylyltransferase	*kdsB*
PGN_1751	PG1814		DNA primase	*dnaG*
PGN_1757	PG1808		GTP pyrophosphokinase (stringent factor)	*relA*
PGN_1766	PG1799			
PGN_1773	PG1792		Sodium/hydrogen antiporter	*kefC*
PGN_1784	PG1851		DNA/pantothenate metabolism flavoprotein	
PGN_1785	PG1852		DNA polymerase III epsilon chain	*dnaQ*
PGN_1786	PG1853		DNA polymerase III beta chain	*dnaN*
PGN_1789	PG1856		Deoxycytidylate deaminase	*tadA*
PGN_1794	PG1861		Hypothetical protein	
PGN_1805	PG1878		Cysteinyl-tRNA synthetase	*cysS*
PGN_1806	PG1879		Patatin	
PGN_1807	PG1880		Glycosyltransferase	
PGN_1827	PG1896		S-adenosylmethionine synthetase	*metK*
PGN_1829	PG1898		Nicotinamide mononucleotide transporter	
PGN_1832	PG1901		Ribosome recycling factor	*frr*
PGN_1833	PG1902		Uridylate kinase	*pyrH*
PGN_1834	PG1903		Hypothetical protein	
PGN_1841	PG1911		DNA-directed RNA polymerase subunit alpha	*rpoA*
PGN_1842	PG1912		30S ribosomal protein S4	*rpsD*
PGN_1843	PG1913		30S ribosomal protein S11	*rpsK*
PGN_1844	PG1914		30S ribosomal protein S13	*rpsM*
PGN_1845	PG1915		50S ribosomal protein L36	*rpmJ*
PGN_1846	PG1916		Translation initiation factor IF-1	*infA*

PGN_1847	PG1917		Methionine aminopeptidase type I	
PGN_1848	PG1918		Preprotein translocase subunit SecY	*secY*
PGN_1849	PG1919		50S ribosomal protein L15	*rplO*
PGN_1850	PG1920		50S ribosomal protein L30	*rpmD*
PGN_1851	PG1921		30S ribosomal protein S5	*rpsE*
PGN_1852	PG1922		50S ribosomal protein L18	*rplR*
PGN_1853	PG1923		50S ribosomal protein L6	*rplF*
PGN_1854	PG1924		30S ribosomal protein S8	*rpsH*
PGN_1855	PG1925		30S ribosomal protein S14	*rpsN*
PGN_1856	PG1926		50S ribosomal protein L5	*rplE*
PGN_1857	PG1927		50S ribosomal protein L24	*rplX*
PGN_1858	PG1928		50S ribosomal protein L14	*rplN*
PGN_1859	PG1929		30S ribosomal protein S17	*rpsQ*
PGN_1861	PG1931		50S ribosomal protein L16	*rplP*
PGN_1862	PG1932		30S ribosomal protein S3	*rpsC*
PGN_1863	PG1933		50S ribosomal protein L22	*rplV*
PGN_1864	PG1934		30S ribosomal protein S19	*rpsS*
PGN_1865	PG1935		50S ribosomal protein L2	*rplB*
PGN_1866	PG1936		50S ribosomal protein L23	*rplW*
PGN_1867	PG1937		50S ribosomal protein L4	*rplD*
PGN_1868	PG1938		50S ribosomal protein L3	*rplC*
PGN_1869	PG1939		30S ribosomal protein S10	*rpsJ*
PGN_1870	PG1940		Elongation factor G	*fusA*
PGN_1871	PG1941		30S ribosomal protein S7	*rpsG*
PGN_1872	PG1942		30S ribosomal protein S12	*rpsL*
PGN_1883	PG1951		Glutaminyl-tRNA synthetase	*glnS*
PGN_1884	PG1952		Alkaline phosphatase	*phoA*
PGN_1885	PG1953		Membrane protein	
PGN_1886	PG1954		NAD dependent epimerase	
PGN_1892	PG1961		Hypothetical protein	
PGN_1895	PG1963		Sua5/YciO/YrdC/YwlC family protein	
PGN_1934	PG1989		Hypothetical protein	
PGN_1946	PG2001		Signal peptidase I	*lepB*
PGN_1947	PG2002		Dihydrodipicolinate reductase	*dapB*
PGN_1955	PG2010		Phosphomannomutase	*cpsG*
PGN_1968	PG2022		Hypothetical protein	
PGN_1969	PG2023		Methionyl-tRNA formyltransferase	*fmt*
PGN_1990	PG2044		Zinc ribbon domain-containing protein	
PGN_1991	PG2046		Cell-cycle protein	
PGN_1996	PG2052		Dihydrodipicolinate synthase	*dapA*
PGN_2005	PG0056		Hypothetical protein	
PGN_2006	PG0057		Nicotinate phosphoribosyltransferase	*pncB*
PGN_2007	PG0058		Nicotinic acid mononucleotide adenylyltransferase	*nadD*
PGN_2018	PG0070		UDP-N-acetylglucosamine acyltransferase	*lpxA*
PGN_2019	PG0071		UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	*lpxC*
PGN_2020	PG0072		UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	*lpxD*
PGN_2022	PG0074		Peptide chain release factor 1	*prfA*
PGN_2023	PG0075		Phosphoribosylformylglycinamidine cyclo-ligase	*purM*
PGN_2024	PG0076		N-acetylmuramoyl-L-alanine amidase	*amiB*
PGN_2034	PG0087		Sugar isomerase	
PGN_2045	PG0099		Phenylalanyl-tRNA synthetase subunit beta	*pheT*

PGN_2051	PG2071		Hypothetical protein	
PGN_2053	PG2069		Oxidoreductase	*fabG*
PGN_2054	PG2068		Glycerol-3-phosphate cytidylyltransferase	*tagD*
PGN_2055	PG2067		4-hydroxythreonine-4-phosphate dehydrogenase	*pdxA*
PGN_2056	PG2066		Lipoprotein	
PGN_2060	PG2062		Histidyl-tRNA synthetase	*hisS*
PGN_2061	PG2061	Possibly	Dihydrofolate reductase	*folA*
PGN_2062	PG2060	Possibly	Thymidylate synthase	*thyA*
PGN_2079	PG2215		Mannose-1-phosphate guanylyltransferase	*manC*
PGN_2081	PG2217		1-deoxy-D-xylulose-5-phosphate synthase	*ispA*
PGN_2085	PG2221		Fe-S oxidoreductase / MiaB-like tRNA modifying enzyme	
PGN_2086	PG2222		Acetyltransferase	*lpxM*
PGN_2087	PG2223		Glycoslytransferase	*wbbL*

Additional file 2. Table S2. Shared genes between *P. gingivalis* strain ATCC 33277 and *B. thetaiotaomicron* that are only essential in *B. thetaiotaomicron*.
[File can be accessed online through BMC Genomics]

Additional file 3. Table S3. Details of *P. gingivalis* strain ATCC 33277 essential genes shared only with *B. thetaiotaomicron*.
[File can be accessed online through BMC Genomics]

Additional file 4. Table S4. *P. gingivalis* core genome in relation to gene essentiality. The 1476 genes that comprise the *P. gingivalis* core genome are listed in order of their TIGR gene identification number (Brunner *et al*. BMC Microbiology 2010). Genes with their TIGR functional characterizations highlighted in green are *P. gingivalis* essential gene homologues in strain W83, while those highlighted in blue are non-essential *P. gingivalis* core genes that have BLAST matches within the DEG. BLAST matches were determined as having protein-protein similarity of e-values $1x10^{-8}$ or less. Black lettering within brackets describing "only in" denote what species in the DEG a core gene had similarity to if there was only one. Red lettering within brackets denotes BROP annotation to genes when differing from TIGR annotations.
[File can be accessed online through BMC Genomics]

Additional file 5. Table S5. *P. gingivalis* strain ATCC 33277 genes without insertions that are excluded from consideration as essential genes based on total TA site number.
[File can be accessed online through BMC Genomics]

Additional file 6. Table S6. Primer sequences for vector-part PCR, nested semi-random PCR sequencing and Illumina sequencing.

Primer	Sequence	Comments
ermG +	TAGGTGCAGGGAAAGGTCAT	Vector
ermG -	CCATTTTTGCTGGCTTTCTT	Vector
bla +	TTGCCGGGAAGCTAGAGTAA	Vector
bla -	GCTATGTGGCGGTATTAT	Vector
himar1c9a +	GACGGAAAAACTCGGGTGTA	Vector
himar1c9a -	TTCAAGCGTGGTGAAATGAG	Vector
SAMseq1	ACGTACTCATGGTTCATCCCGATA	Semi-random
SAMseq2	GCGTATCGGTCTGTATATCAGCAA	Semi-random
SAMseq3	TCTATTCTCATCTTTCTGAGTCCAC	Semi-random
ARB1	GGCCACGCGTGCACTAGTACN10TACNG	Semi-random
ARB2	GGCCACGCGTGCACTAGTAC	Semi-random
pSAM1	CCTGACGGATGGCCTTTTTGCGTTTCTACC	Illumina
pSAM2	AATGATACGGCGACCACCGAGATCTACACTCTTTGACCGGGGACTTATCATCCAACCTGTTA	Illumina
pSAM3	ACACTCTTTGACCGGGGACTTATCATCCAACCTGTTA	Illumina
Olj527	CAAGCAGAAGACGGCATACGAGATAAAAAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj528	CAAGCAGAAGACGGCATACGAGATACACACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj529	CAAGCAGAAGACGGCATACGAGATAGAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj530	CAAGCAGAAGACGGCATACGAGATATATATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj531	CAAGCAGAAGACGGCATACGAGATCACACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj532	CAAGCAGAAGACGGCATACGAGATCCCCCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj533	CAAGCAGAAGACGGCATACGAGATCGCGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj534	CAAGCAGAAGACGGCATACGAGATCTCTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj535	CAAGCAGAAGACGGCATACGAGATGAGAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj536	CAAGCAGAAGACGGCATACGAGATGCGCGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj537	CAAGCAGAAGACGGCATACGAGATGGGGGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj538	CAAGCAGAAGACGGCATACGAGATGTGTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj539	CAAGCAGAAGACGGCATACGAGATTATATAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj540	CAAGCAGAAGACGGCATACGAGATTCTCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj541	CAAGCAGAAGACGGCATACGAGATTGTGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina
Olj542	CAAGCAGAAGACGGCATACGAGATTTTTTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina

Chapter 3: Identification and characterization of a Miniature Inverted-repeat Transposable Element

[The content of the following chapter has been submitted for publication and is under review]

Identification and characterization of a minisatellite contained within a novel miniature inverted-repeat transposable element (MITE) of Porphyromonas gingivalis
Brian Andrew Klein; Tsute Chen; Jodie C Scott; Andrea L Koenigsberg; Margaret J Duncan; Linden T Hu

BAK conceived of the study, participated in its design and coordination, carried out molecular genetics, carried out bioinformatic analyses and drafted the manuscript. TC participated in study design and coordination, carried out bioinformatic analyses and drafted the manuscript. AK participated in study design and coordination and carried out molecular genetics. JCS participated in study design and coordination and carried out molecular genetics. MJD participated in study design and coordination and drafted the manuscript. LTH conceived of the study, participated in its design and coordination and drafted the manuscript.

3.1 Abstract

Background

Repetitive regions of DNA and transposable elements have been found to constitute large percentages of eukaryotic and prokaryotic genomes. Such elements are known to be involved in transcriptional regulation, host-pathogen interactions and genome evolution.

Results

We identified a minisatellite contained within a miniature inverted-repeat transposable element (MITE) in *Porphyromonas gingivalis*. The *P. gingivalis* minisatellite and associated MITE, named 'BrickBuilt', comprises a tandemly repeating twenty-three nucleotide DNA sequence lacking spacer regions between repeats, and with flanking 'leader' and 'tail' subunits that include small inverted-repeat ends. Forms of the BrickBuilt MITE are found 19 times in the genome of *P. gingivalis* strain ATCC 33277, and also multiple times within the strains W83, TDC60, HG66 and JCVI SC001. BrickBuilt is always located intergenically ranging between 49 and 591 nucleotides from the nearest upstream and downstream coding sequences. Segments of BrickBuilt contain promoter elements with bidirectional transcription capabilities.

Conclusions

We performed a bioinformatic analysis of BrickBuilt utilizing existing whole genome sequencing, microarray and RNAseq data, as well as performing *in vitro* promoter probe assays to determine potential roles, mechanisms and regulation of the expression of these elements and their affect on surrounding loci. The multiplicity, localization and limited host range nature of MITEs and MITE-like elements in *P. gingivalis* suggest that these elements

may play an important role in facilitating genome evolution as well as modulating the transcriptional regulatory system.

3.2 Background

Porphyromonas gingivalis, a Gram-negative, anaerobic, asaccharolytic, black-pigmenting bacterium, is a keystone pathogen in the development and progression of periodontal disease (Hajishengallis et al., 2011)(Curtis et al., 2011). Multiple repetitive and transposable elements were previously identified in the *P. gingivalis* genomes (Maley & Roberts, 1994) (Wang, Bond, & Genco, 1997) (Lewis & Macrina, 1998) (Sawada et al., 1999) (Califano et al., 2000) (Califano, Arimoto, & Kitten, 2003) (Chen et al., 2004)(Naito et al., 2008) (Naito et al., 2011) (Bainbridge, Hirano, Grieshaber, & Davey, 2015). Genome sequences are now available for multiple strains of *P. gingivalis* which has greatly facilitated genetic and genomic analyses of the species (Nelson et al., 2003) (Chen et al., 2004)(Naito et al., 2008) (Watanabe et al., 2011) (McLean et al., 2013)(Siddiqui et al., 2014). Each of the sequenced *P. gingivalis* genomes has contained multiple repetitive and transposable elements, an aspect that makes sequencing and alignment difficult.

Repetitive Elements (REs) are DNA sequences present in multiple copies throughout a genome, chromosome or vector. They are broadly classified into ‘terminal’, ‘tandem’ and ‘interspersed’ repeats, however, each of these classifications encompasses several sub-types of REs. Tandem repeats are classified as either identical or non-identical based on the level of nucleic acid matching. They are then further classified as either micro, mini or macro satellites based on size of the repeat. REs can either be localized at a single site where a motif is recurrent sequentially adjacent to each other or at many loci as reiteration (Treangen,

Abraham, Touchon, & Rocha, 2009)(Zhou, Aertsen, & Michiels, 2014) (Padeken, Zeller, & Gasser, 2015).

Transposable Elements (TEs) are 'mobile' DNA sequences that can change locus or multiply and insert into new loci within a genome or between genomes via excision/replication and insertion. They can insert into chromosomes, plasmids and bacteriophages. Class I TEs are retrotransposons, which require reverse-transcriptase activity to transpose. Class II TEs are DNA transposons, which unlike reverse transcriptase-utilizing Class I elements, require a transposase or a replicase to transpose (Siguier, Gourbeyre, & Chandler, 2014)(Piégu, Bire, Arensburger, & Bigot, 2015) (Padeken et al., 2015). Class II elements can either be autonomous or non-autonomous, the latter [canonically] having undergone mutations involving the transposase such that they can no longer duplicate or excise without the assistance of a parent element that utilizes a similar transposase. Within the non-autonomous element sub-class are miniature inverted-repeat transposable elements, or MITEs (Gonzalez & Petrov, 2009)(Ilyina, 2010)(Fattash et al., 2013)(Darmon & Leach, 2014).

MITEs have a distinct structure relative to other TEs. They are between 50-1000 bp in length and are often present in high copy numbers per genome. MITEs are typically AT-nucleotide (nt) rich and frequently contain terminal inverted repeats (TIRs) and transposon site duplications (TSDs), but they lack the capacity to code for functional transposases (Gonzalez & Petrov, 2009) (Ilyina, 2010) (Fattash et al., 2013) (Darmon & Leach, 2014). Transposable elements, in particular MITEs, can be found in all taxa, varying in number and type between species and can account for greater than half of a genome. Bacteria typically carry between 10-20 copies of a MITE per genome, while plants may have up to 20,000

copies of a given MITE. MITE copy numbers are suggested to depend on non-coding region availability, polyploidy, the presence of a fully-functional autonomous version of a transposase, evolutionary 'burst' opportunities and regulatory potential of the given element (Feschotte, Jiang, & Wessler, 2002) (Naito et al., 2009) (Lu et al., 2012)(Wang et al., 2013). Eukaryotic MITEs are frequently found in or closely associated with the coding region while prokaryotic MITEs are almost exclusively found intergenically (Bureau & Wessler, 1992) (Bureau & Wessler, 1994) (Feschotte et al., 2002) (Chen & Shapiro, 2003) (Han & Korban, 2007) (Nelson, Bhaya, & Heidelberg, 2012) (Deng et al., 2013) (Sampath et al., 2014). Intergenically located MITEs in prokaryotes have been shown to be able to affect gene expression by the organism (Ilyina, 2010) (Darmon & Leach, 2014).

Several studies have demonstrated potential interactions of repetitive elements with transposable elements, which are generally thought to work independently and be mutually exclusive. In the wedge clam (*Donax trunculu*) genome as well as the butterfly and moth (*Lepidoptera*) genomes, 'hitchhiking' microsatellites were found within transposable elements (Coates, Kroemer, Sumerford, & Hellmich, 2011) (Satovic & Plohl, 2013). Microsatellites and simple sequence repeats have also been found closely associated with transposable elements in *Neisseria meningitidis* (Parkhill et al., 2000).

Here we describe 'BrickBuilt', a miniature inverted-repeat transposable element containing a minisatellite, in *P. gingivalis*. The sequences, location, copy number, prevalence throughout the species, as well as implications on genome (in)stability and transcriptional regulation are described. Similarities to other autonomous and non-autonomous *P. gingivalis* transposable elements are addressed with the goal of defining a potential network for the biogenesis of these elements in *P. gingivalis* and their effects on the *P. gingivalis* genome.

3.3 Results and Discussion

3.3.1 Identification of a MITE in Porphyromonas gingivalis

We identified a DNA element, 'BrickBuilt', in the genome of *P. gingivalis* strain ATCC 33277. The element was initially identified as a tandemly-repeated sequence of 23 nt located intergenically at a single site (Supplementary Fig.3-1). A more thorough investigation of the genome revealed 19 independent, non-identical segments of the element scattered throughout the genome of strain ATCC 33277 (Table 3-1). The smallest number of 23 nt direct repeats is 1 (BrickBuilt_1) and the largest 22.8 (BrickBuilt_12). The 23 nt direct repeats are imperfect within a given element, imperfect bases vary from one element to another and imperfections do not correlate with length or total number of repeats within a given element (Fig.3-1). The percent of mismatches within a given element varies from 0 to 11, and the percent of insertions and deletions within an element varies from 0 to 6. Within the 23 nt repeats there are conserved and non-conserved nucleotide sites, with the latter half of the element containing the majority of non-conserved sites (Fig.3-1). No spacers are found between the tandem repeats in BrickBuilt.

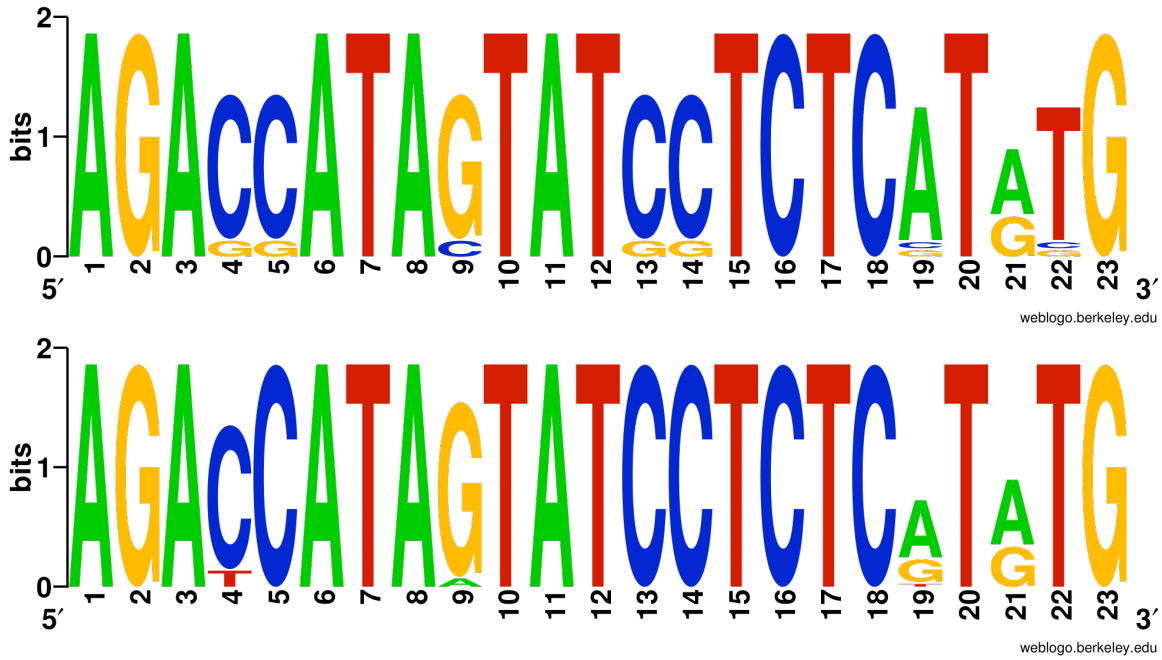


Figure 3-1: Sequence logo representation of the 23 nucleotide repeat region from *P. gingivalis* strain ATCC 33277 multiple alignments. Generated with weblogo software. Total height of a nucleic acid stack represents sequence conservation at a given position. Height of symbols at a stack represents relative frequency of a given nucleic acid at that position. Top sequence logo corresponds to compilation of the consensus of the 19 BrickBuilt elements; consensus for each made prior to combining for sequence logo. Bottom sequence logo corresponds to BrickBuilt_5 alone, constructed from its 18 full repeats.

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
250--683	23	18.9	23	95	0	724	27	25	15	31	1.96

```
                    *
250 AGACCATAGTATCCTCTCATGTG
  1 AGACCATAGTATCCTCTCATATG

                   *
273 AGACCATAGTATCCTCTCGTATG
  1 AGACCATAGTATCCTCTCATATG

                    *
296 AGACCATAGTATCCTCTCATGTG
  1 AGACCATAGTATCCTCTCATATG

      *             *
319 AGATCATAGTATCCTCTCATGTG
  1 AGACCATAGTATCCTCTCATATG

      *           * *
342 AGATCATAGTATCCTCTCTTGTG
  1 AGACCATAGTATCCTCTCATATG

                    *
365 AGACCATAGTATCCTCTCATGTG
  1 AGACCATAGTATCCTCTCATATG

                    *
388 AGACCATAGTATCCTCTCATGTG
  1 AGACCATAGTATCCTCTCATATG

                   *
411 AGACCATAGTATCCTCTCGTATG
  1 AGACCATAGTATCCTCTCATATG


434 AGACCATAGTATCCTCTCATATG
  1 AGACCATAGTATCCTCTCATATG

                    *
457 AGACCATAGTATCCTCTCATGTG
  1 AGACCATAGTATCCTCTCATATG


480 AGACCATAGTATCCTCTCATATG
  1 AGACCATAGTATCCTCTCATATG


503 AGACCATAGTATCCTCTCATATG
  1 AGACCATAGTATCCTCTCATATG


526 AGACCATAGTATCCTCTCATATG
  1 AGACCATAGTATCCTCTCATATG

                   *
549 AGACCATAGTATCCTCTCGTATG
  1 AGACCATAGTATCCTCTCATATG


572 AGACCATAGTATCCTCTCATATG
  1 AGACCATAGTATCCTCTCATATG

          *
595 AGACCATAAATATCCTCTCATATG
  1 AGACCATAGTATCCTCTCATATG

                   *
618 AGACCATAGTATCCTCTCGTATG
  1 AGACCATAGTATCCTCTCATATG

                   *
641 AGACCATAGTATCCTCTCGTATG
  1 AGACCATAGTATCCTCTCATATG


664 AGACCATAGTATCCTCTCAT
  1 AGACCATAGTATCCTCTCAT
```

Supplemental Figure 3-1: Tandem Repeat Finder analysis of *P. gingivalis* strain ATCC 33277 BrickBuilt_5. Overall statistics of the repeats/repeat region found within BrickBuilt_5; 23 nt repeat indicies of the element relative to the entire element, period size, copy number, consensus size, percent matches, percent indels, alignment score, percent composition for each nucleotide and entropy measure based on percent composition. The individual locations of mismatches and InDels within BrickBuilt_5 are shown as positions marked by stars (∗).

After determining the length and locations of each independent direct repeat element we performed alignments of the regions flanking the direct repeats to determine whether specific DNA sequences or overall regions were necessary for the presence of the element. Alignments of the regions flanking the direct repeats revealed well-conserved regions of high homology, different for the two flanks of the repeat, which were determined to be 'leader' and 'tail' regions that encompassed the direct repeats (Fig.3-2). Of the 19 elements, 11 are flanked by portions of both a leader and a tail, 3 by just leader, 2 by just tail, and 3 by neither. When considered as a single whole element, all BrickBuilt elements are intergenic, although some are within regions where annotation pipelines predicted hypothetical genes that do not appear to be expressed based on proteomic data (Xia et al., 2007)(Kuboniwa et al., 2009) (Maeda, Nagata, Ojima, & Amano, 2014). Total length of the complete elements ranges from 991 (BrickBuilt_5) to 84 (BrickBuilt_14), which is determined by number of internal 23nt repeats as well as the specific element may contain full, partial or no leader and tail segments. The longest leader segment is 285 nucleotides (BrickBuilt_17) and the longest tail segment is 318 nucleotides (BrickBuilt_4). In no instance are individual BrickBuilt elements separated by full-length autonomous transposable elements or complete, non-repetitive genes. Additionally, no leader-to-tail versions are present without a full 23 nt repeat and no TIR-containing individual leader or tail segments are present without the 23 nt repeats. A single site adjacent to the Hmu operon has a partial tail-only version that lacks the terminal 20 nt that would include the TIR; no partial leader-only versions of the element are present.

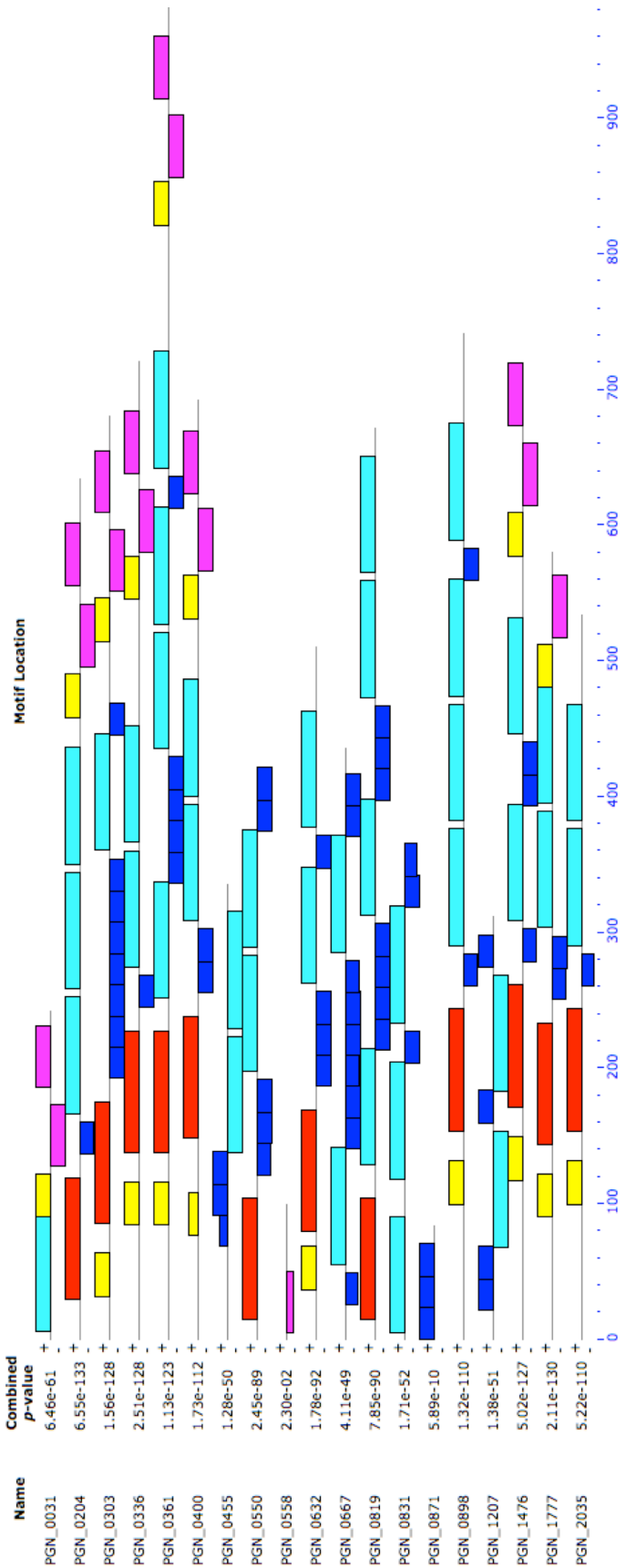

Name
Combined p-value
Motif Location
PGN_0031
6.46e-61
PGN_0204
6.55e-133
PGN_0303
1.56e-128
PGN_0336
2.51e-128
PGN_0361
1.13e-123
PGN_0400
1.73e-112
PGN_0455
1.28e-50
PGN_0550
2.45e-89
PGN_0558
2.30e-02
PGN_0632
1.78e-92
PGN_0667
4.11e-49
PGN_0819
7.85e-90
PGN_0831
1.71e-52
PGN_0871
5.89e-10
PGN_0898
1.32e-110
PGN_1207
1.38e-51
PGN_1476
5.02e-127
PGN_1777
2.11e-130
PGN_2035
5.22e-110
0
100
200
300
400
500
600
700
800
900

Figure 3-2: MEME motif analysis block output of the 19 BrickBuilt elements in *P. gingivalis* strain ATCC 33277. Entire element FASTA sequences were used as the input. Settings for analysis were: Distribution of motif occurance as 'any number of repititions', number of different motifs as '5', minimum motif with as '23', and maximum motif width as '200'. Dark and light blue blocks correspond to 23 nucleotide repeat regions, red blocks to leader regions and purple blocks to tail regions. Yellow blocks, potentially representing a 5^{th} motif, were only found on the positive strand and had the lowest e-values associated with their significance scores.

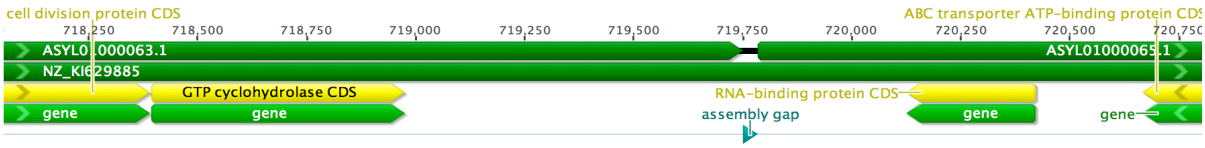
The genome sequence of strain ATCC 33277 contains 2,345,886 bases. When compiled together all 19 BrickBuilt elements in strain ATCC 33277 make up 10,276 bases, or 0.44 percent of the overall genome content; the equivalent of 9 protein coding sequences in this strain on average.

3.3.2 Conservation of BrickBuilt elements in other strains of P. gingivalis

Of the 19 versions of BrickBuilt found within strain ATCC 33277, 16 are conserved between the analogous coding sequences within strains W83, TDC60 and HG66 (Table 3-1 and Supplementary Table 3-1). Strains HG66 and TDC60 contain 19 versions of BrickBuilt, equivalent to the number in strain ATCC 33277. However, strain W83 only contains 18 versions of the element. Strains ATCC 33277 and HG66 share the exact same 19 loci for BrickBuilt elements. One locus in strain TDC60 (BrickBuilt_4) is deviant and is encompassed by two *ISPg*1 elements. Strain W83 has three sites that differ from the other strains, all which are located adjacent to other types of IS or repetitive elements. In this strain BrickBuilt_4 is completely lacking, while BrickBuilt_7 and BrickBuilt_18 are aberrant with respect to size having only maintained 23 nt repeats.

BrickBuilt elements can be identified in the genome of *P. gingivalis* strain JCVI SC001, which is not yet included in the default NCBI BLAST nucleotide database settings. Genome searches of the FASTA files from the JCVI SC001 revealed that most BrickBuilt loci in the JCVI SC001 genome contained strings of undetermined bases, which can be attributed to the manner of isolation, sequencing and assembly. Eight other *P. gingivalis* genomes have since been sequenced and deposited in NCBI, yet they are not completely assembled. Assembly gaps are at sites where the corresponding surrounding CDS from

ATCC 33277 contains BrickBuilt elements, suggesting BrickBuilt is present in those genomes as well and potentially capable of causing assembly difficulties (Supplementary Fig.3-2). Additionally, a degenerate version of the 23 nt repeat consensus sequence (AGAYCATARTATCCTCTCRTRTG) was searched against all 13 (8 unfinished) *P. gingivalis* genomes, each giving positive hits (data not shown). Because of assembly gaps and undetermined bases around BrickBuilt sites in the unfinished sequencing projects they were not included in multiple alignments.



Supplemental Figure 3-2: *P. gingivalis* strain SJD2 assembly showing an 'assembly gap' at the site of BrickBuilt elements from strains ATCC 33277, W83, TDC60 and HG66. The same genes flanking the assembly gap are found flanking BrickBuilt_11 in strains ATCC 33277, W83, TDC60 and HG66. The top dark green track depicts individual contigs (a total of 140 for this strain), the second dark green track depicts individual scaffolds (a total of 117 for this strain), the yellow track shows predicted coding sequences, the light green track shows predicted genes, and the bottom track shows the assembly gap as a rightward-pointing triangle.

From initial characterizations of the size and locations of BrickBuilt elements, we believe they can be classified within the large group of non-autonomous transposable elements, potentially best fitting within the MITE subcategory. A caveat must be placed, however, given that MITE elements are typically described as being comprised of two homologous flanking regions, and we have determined that BrickBuilt elements contain distinct 'leader' and 'tail' segments. Since all accessible genome sequences of *P. gingivalis* strains contain BrickBuilt elements, the parent element or first version of BrickBuilt probably occurred early within the phylogeny of the *P. gingivalis* species. Insertion of the 23 nt repeat(s) into the original parent element may be the event that catalyzed the inactivation of an autonomous transposable element. Alternatively, a version of BrickBuilt already containing the 23 nt repeats could have been laterally-transferred via plasmid or horizontally-transferred via phage. Given that no full-length (TIR-containing) leader or tail regions are present without a 23 nt repeat it may be deleterious to maintain a full leader or tail region on the chromosome, or the 23 nt repeat is required by the autonomous element.

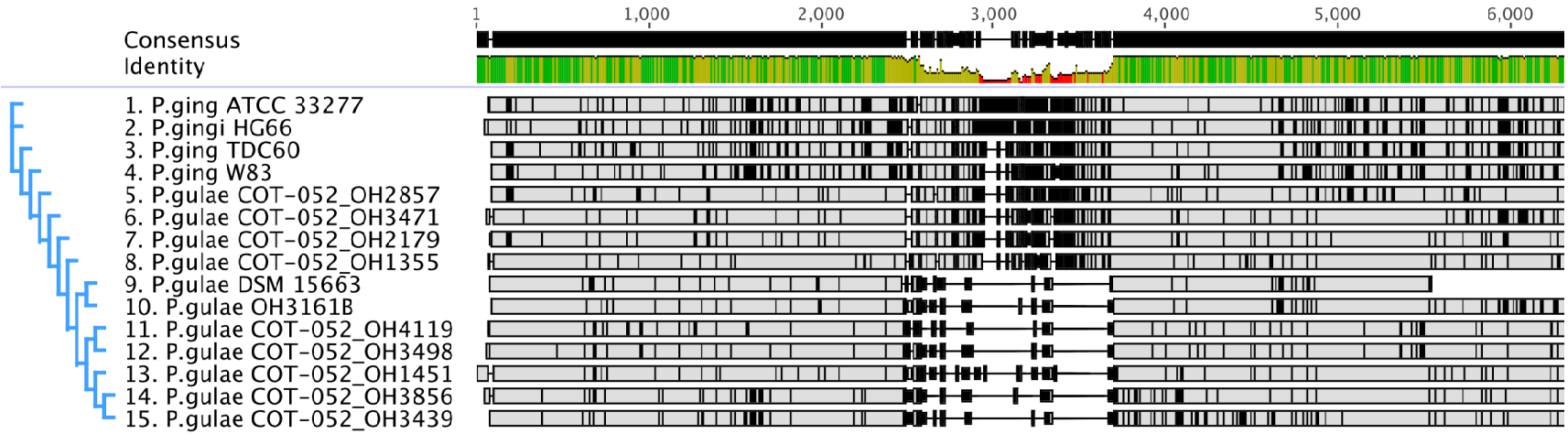
The repetitive nature of the BrickBuilt elements, both the internal repeats and that they are found multiple times through the genome, can lead to sequencing, assembly and annotation issues. Because the strains W83, ATCC 33277, TDC60, HG66 and JCVI SC001 are unique strains, were sequenced independently, and were *de novo* assembled, placement of BrickBuilt elements at the same locus across genomes is unlikely to be due to use of a shared scaffold. However, care should be taken when aligning newly-sequenced *P. gingivalis* genomes to a scaffold.

3.3.3 Homology to Other MITEs and Repetitive Elements

The 23 nt repeats and the leader and tail segments of the element were analyzed using BLAST (NCBI server) to determine whether the element is present in genomes other than the species *P. gingivalis* (Altschul, Gish, Miller, Myers, & Lipman, 1990). With default BLAST nucleotide settings, a full-length BrickBuilt and each of the three distinct parts of the element match solely to *P. gingivalis* strains (as of 10/2013). All four sequenced and annotated strains available for BLAST searching harbored high-homology hits for BrickBuilt. No hits, full or partial, were found in other *Porphyromonas* species or related *Bacteroides* species. Through querying discontiguous megablast as well as using less stringent search constraints within megablast with 'max target sequences', 'expect threshold', 'word size' and 'filter low complexity regions', low-homology hits were obtained with the terminal inverted repeat regions. However, these matches did not show homology to the rest of BrickBuilt. Additionally, when initially queried in October of 2013, whole genome shotgun sequencing projects did not contain BrickBuilt matches.

Of note, when BLASTx searches (protein database search using a translated nucleotide query) were performed with the leader and tail sequences under default settings several *Bacteroidetes* species contained tail hits and one species contained a leader hit. *Porphyromonas gulae* contained strong hits with both leader and tail, while *Prevotella tannerae* and *P. dentalis* contained weak tail hits only. All of the BLASTx hits were either part of a predicted transposase/partial transposase or a hypothetical protein. If BrickBuilt were indeed a non-autonomous transposable element, homology to sections of related transposons through BLASTx would not be unexpected. As such, low BLASTx homology does not point to these hits being potential parent or identical elements of BrickBuilt.

As we had initially determined BrickBuilt to be a species-specific element, which it was during our early investigations, our intentions were to focus on *P. gingivalis*. The first *P. gulae* genome was initially published on 04/22/2013, then edited and accessible through NCBI as of 12/12/2013 (Tatusova, Ciufo, Fedorov, O'Neill, & Tolstoy, 2015). Additional *P. gulae* genomes became available as of 10/10/2014 (Coil et al., 2015). Following the upload of the additional *P. gulae* genomes to NCBI, we utilized BLAST to search the related WGS shotgun sequencing data and found that some of the *P. gulae* strains do in fact carry BrickBuilt homologues (Supplementary Fig.3-3). Some of the *P. gingivalis* BrickBuilt element locations are conserved within *P. gulae* strains. However, greater strain variation seems evident in *P. gulae* at certain BrickBuilt loci than between *P. gingivalis* strains (Supplementary Fig.3-3). Of note, the original *P. gulae* genome was obtained from a wolf and the subsequent strains were obtained from domesticated dogs. The original strain (DSM 15663) only contains 4 BrickBuilt homologues within the genome, and importantly lacks the BrickBuilt_5 homologue that was used for the majority of BLAST database queries.


1
1,000
2,000
3,000
4,000
5,000
6,000
Consensus
Identity
1. P.ging ATCC 33277
2. P.gingi HG66
3. P.ging TDC60
4. P.ging W83
5. P.gulae COT-052_OH2857
6. P.gulae COT-052_OH3471
7. P.gulae COT-052_OH2179
8. P.gulae COT-052_OH1355
9. P.gulae DSM 15663
10. P.gulae OH3161B
11. P.gulae COT-052_OH4119
12. P.gulae COT-052_OH3498
13. P.gulae COT-052_OH1451
14. P.gulae COT-052_OH3856
15. P.gulae COT-052_OH3439

Supplemental Figure 3-3: BrickBuilt_5 region MAFFT alignment and PHYML tree of *P. gulae* and *P. gingivalis* strains. All COT_052 *P. gulae* strains were sequenced/deposited during the preparation of the manuscript. Additionally, all *P. gulae* strains are currently scaffold or contig assemblies; none are completed chromosomes and thus are also not available for default BLASTn query on NCBI. The aligned bases between 2,500-3,800 contain the BrickBuilt_5 MITE; flanking regions contain the same 2 upstream and downstream genes in all strains. Of the 12 nodes in the tree, 8 have a bootstrap value of greater than 85 (100 bootstrap iterations). In the 'consensus identity' track, green indicates sites of complete conservation, yellow of partial conservation and red of little conservation. Within each of the 15 strain tracks the black lines or blocks indicate sites that deviate from the consensus at that given site.

Within *P. gingivalis* there have been three previously identified groups of MITEs or non-autonomous transposable elements; named the 239, 464 (PgRS) and 700 groups (Nelson et al., 2003)(Naito et al., 2008). During our examination of BrickBuilt we analyzed whether any sequence overlap was apparent between the elements and found that the terminal inverted repeats are similar, yet the rest of the elements do not bear similarity. 464/PgRS elements were previously identified as containing 41 nucleotide tandem direct repeats (Nelson et al., 2003) (Naito et al., 2008). The 23 and 41 nucleotide internal tandem direct repeats of the elements do not share homology with each other and neither have non-*P. gingivalis* BLAST matches within the NCBI database. The segments of the 464/PgRS elements flanking the 41 nt tandem direct repeats are themselves repetitive, which is unlike the non-repetitive leader and tail segments of BrickBuilt. Although not related by sequence, similarities in copy number between 464/PgRS and BrickBuilt elements are evident. With *P. gingivalis* harboring four types of MITE-like elements it is interesting that two types, BrickBuilt and PgRS (464), contain microsatellite repeats. Although several 236 and 700-type elements are located near repeats or other repetitive elements, they seem not to have encompassed any mini- or microsatellites from analyses of the currently available genomes.

In addition to Tn and IS elements, multiple groups have described repetitive sequences within *P. gingivalis* genomes ranging from single nucleotide tracts to mini- and microsatellites (Nelson et al., 2003)(Coenye & Vandamme, 2005). Several 41, 23 and 22 nucleotide tandem direct repeats were described in *P. gingivalis* strain W83, yet the exact locations of such repeats were not identified, nor were comparative genomics an option at the time of the report (Nelson et al., 2003). Some of the 23 nt tandem direct repeats noted are presumably the direct repeat portions of BrickBuilt.

Within *P. gingivalis* there are 11 recognized IS elements and 2 different composite transposon (Ctn) elements (Nelson et al., 2003)(Duncan, 2003) (Tribble, Kerr, & Wang, 2013). Ten of the inverted repeats for the 13 IS and Ctn are characterized. The MITE-like elements in *P. gingivalis* share either identical or within one nucleotide TIRs with those of full-length IS elements within the *P. gingivalis* genomes; *ISPg1*, *ISPg3*, *ISPg4* and *ISPg9* (Table 3-2). The matching full-length ISPg elements are all categorized within the IS5 family. BrickBuilt's TIRs are most similar to those of *ISPg1* and *ISPg9* (which share identical TIRs); *ISPg4* is the next closest match with 2 nucleotides different (Table 3-2). Although the TIRs are similar, no remnant of a transposase from any *P. gingivalis* IS or Tn element remains within any of the BrickBuilt copies.

After determining the IS5 family-like TIRs of BrickBuilt other IS5 elements were scanned for potential similarity. Identical TIRs to that of BrickBuilt were found in the *Neisseria meningitidis ISNmeI* and its derivatives (Table 3-2). *ISNmeI* is the proposed (based on TIRs) parent element for the type II MITE ATR (AT-rich Repeat) in *Neisseria meningitidis* genomes (Parkhill et al., 2000). ATR elements are found 19 times within *N. meningitidis* genomes, which is similar to that of BrickBuilt's distribution. Additionally, ATR elements are frequently associated with direct repeat elements of *N. meningitidis* known as REP2.

Contributing to the 'mobility' of transposable elements, the elements usually encode terminal inverted repeats, which are generally the motif that is recognized and processed by the transposase(s) that mobilize them. No reports of repeated non-autonomous transposable elements within a prokaryotic genome that does not carry an autonomous 'parent' element from which the defective elements have been borne out of have been made. However, it is

possible that if a large enough portion of the genome was rearranged via horizontal gene transfer then multiple versions of a non-autonomous element could come to exist in a genome without an autonomous parent. It is also possible that TIRs of a host transposon were procured or happened by chance, thus not necessitating the element be a defective version of a full-length transposable element.

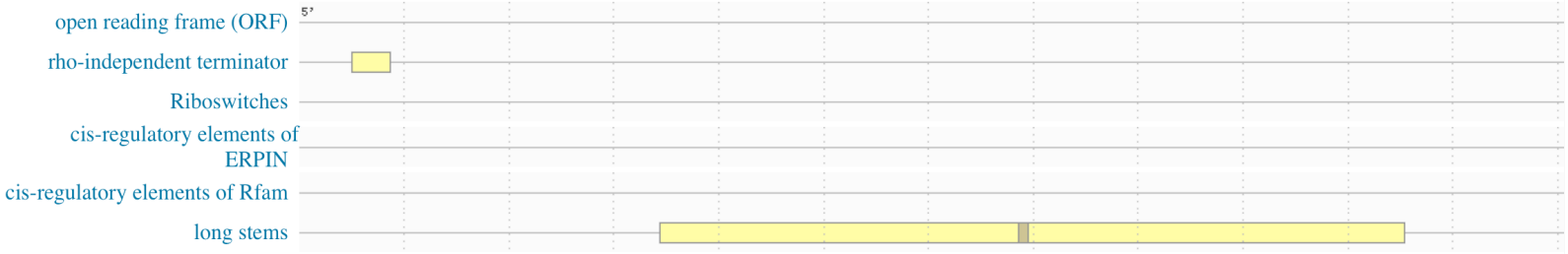
The BLAST-based limited host range nature of BrickBuilt is intriguing yet not uncommon for non-autonomous transposable elements (Hikosaka & Kawahara, 2004) (Yang & Barbash, 2008)(Koressaar & Remm, 2012)(Halász, Kodad, & Hegedus, 2014). Once a non-autonomous element occurs within a genome, usually by deletions of an autonomous transposable element but potentially via conjugation based horizontal transfer of a plasmid or transduction via a bacteriophage, movement between species will become less likely. Additionally, few bacterial species have multiple genome assemblies available for intraspecies comparisons, which could lead to missed elements due to strain variation. Furthermore, it is possible that repetitive sequences could be mis-sequenced or left out of genome assemblies due to repeat region sequencing difficulties or unassigned bases.

3.3.4 Predicted Secondary Structure of BrickBuilt

The direct repeats within BrickBuilt are predicted to form long stem loop structures (Fig.3-3 and 3-4). Three DNA/RNA structure prediction programs, Mfold, RNAstructure and RegRNA2.0, independently predicted long stem loops to form from/within the element (Zuker, 2003) (Bellaousov, Reuter, Seetin, & Mathews, 2013) (Chang et al., 2013). The length of the version of BrickBuilt affects the size of the predicted stem loop structure and the associated entropy. BrickBuilt_1, BrickBuilt_9 and BrickBuilt_14 are not predicted to

form long stem loops by the RegRNA2.0 program due to the length of the internal 23 nt repeats, however, shorter stem loops due to dyad symmetry may occur. BrickBuilt elements are predicted to be surrounded/flanked by Rho-independent terminators and/or polyadenylation sites in 10 of 19 instances. No portion of BrickBuilt matched to any structures in Rfam (Nawrocki et al., 2015).

▶ Map View



▶ Table View

Motif Type	Motif Name	Position	Length	Sequence
open reading frame (ORF)				
rho-independent terminator	Rho-independent Terminator	51 ~ 86	36	aggccgtgtgtgatcgtattcggatcgccatggccg
Riboswitches				
cis-regulatory elements of ERPIN				
cis-regulatory elements of Rfam				
long stems	Stem_1_5prime	427 ~ 638	212	atagaccatagtatcctctcatgtgagaccatagtatcctctcgtatgag accatagtatcctctcatgtgagatcatagtatcctctcatgtgagatca tagtatcctctcttgtgagaccatagtatcctctcatgtgagaccatagt atcctctcatgtgagaccatagtatcctctcgtatgagaccatagtatcc tctcatatgaga
	Stem_1_3prime	650 ~ 862	213	tctcatgtgagaccatagtatcctctcatatgagaccatagtatcctctc atatgagaccatagtatcctctcatatgagaccatagtatcctctcgtat gagaccatagtatcctctcatatgagaccataatatcctctcatatgaga ccatagtatcctctcgtatgagaccatagtatcctctcgtatgagaccat agtatcctctcat
	Stem_2_5prime	345 ~ 684	340	atctatcggtttaacagctggtacggaacctcccactccccgaatctgcc atccggcagagagcagatagcatcttccatcaatagaccatagtatcctc tcatgtgagaccatagtatcctctcgtatgagaccatagtatcctctcat gtgagatcatagtatcctctcatgtgagatcatagtatcctctcttgtga gaccatagtatcctctcatgtgagaccatagtatcctctcatgtgagacc atagtatcctctcgtatgagaccatagtatcctctcatatgagaccatag tatcctctcatgtgagaccatagtatcctctcatatgaga
	Stem_2_3prime	696 ~ 1053	358	tctcatatgagaccatagtatcctctcatatgagaccatagtatcctctc gtatgagaccatagtatcctctcatatgagaccataatatcctctcatat gagaccatagtatcctctcgtatgagaccatagtatcctctcgtatgaga ccatagtatcctctcatgtaagttcattctttgtatctcagggaaaaacc aatgagcaggtgtatggctgccaagtaaacagaaagcagtgtccgaatac gtcgccaatcgataagcatatcgacttggttcgaaaatacattccgtgct ttcctgtatcgattcgttgacatgacgtccctgaatgagacctcttcgtc ggatgaat

Figure 3-3: RegRNA2.0 analysis output of *P. gingivalis* strain ATCC 33277 BrickBuilt_5 and surrounding CDS-to-CDS area. The immediate 5' CDS is PGN_0361 and the immediate 3' CDS is PGN_0364 (*argS*). Options of open reading frames, Rho-independent terminators, transcriptional regulatory motifs, riboswitches, cis-regulatory elements, ERPINs, Rfam database matches, long stems and functional RNA sequences were queried. The Rho-independent terminate identified occurs prior to the BrickBuilt element.

dG = -128.02 BrickBuilt_5 [Pg ATCC 33277]

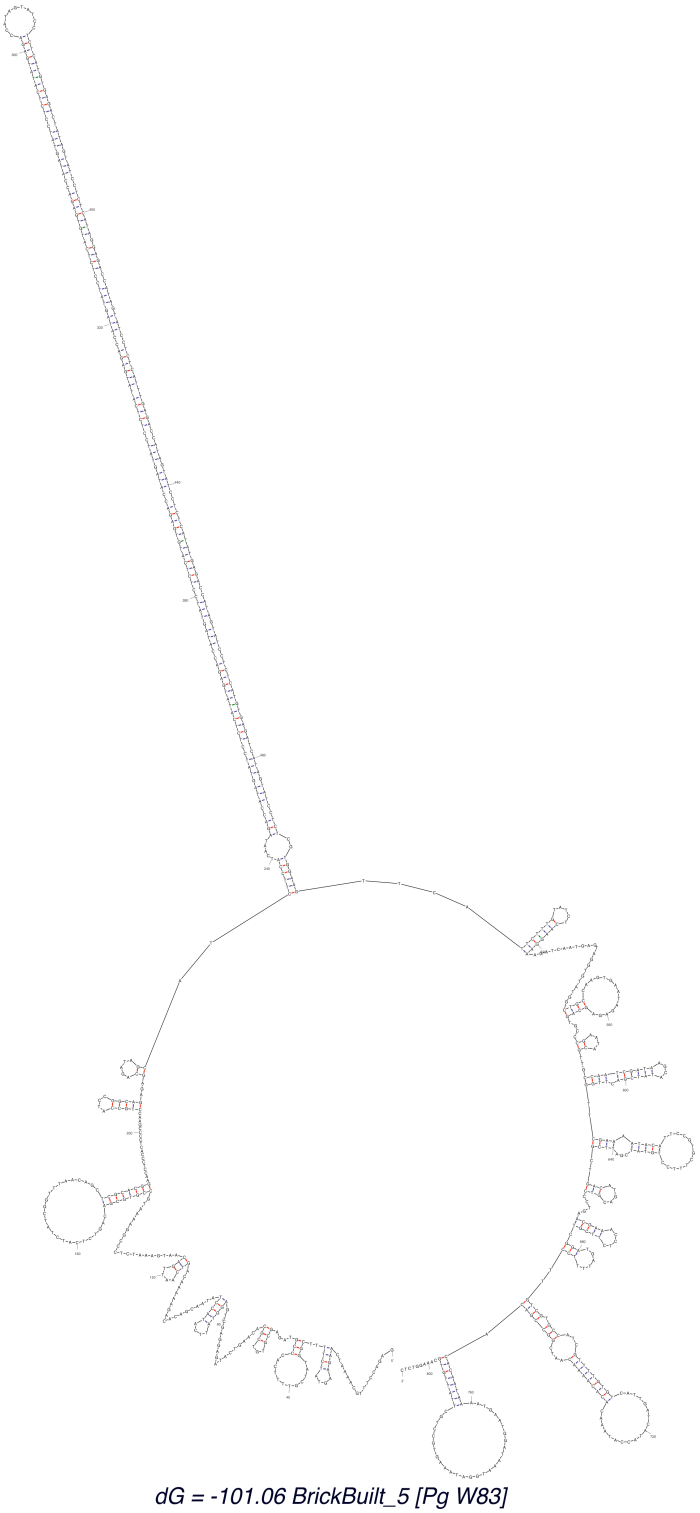


Figure 3-4: Mfold analysis output of BrickBuilt_5 for *P. gingivalis* strain ATCC 33277 (top) and W83 (bottom). Calculated entropy for the ATCC 33277 structure is ΔG = -128.02 and for W83 is ΔG = -101.06. Input for strain ATCC 33277 was 991 nt and 807 nt for strain W83. The 184 nt length difference is due to 23 nt repeat numbers. The 23 nt repeats comprise the long stem loops.

Predicted structures of BrickBuilt vary slightly between strains at a given conserved locus. The BrickBuilt_5 Mfold entropy predictions for strains ATCC 33277 and W83 are -128.02 and -101.06, respectively (Fig.3-4). BrickBuilt_5 in ATCC 33277 is 991 nucleotides long and the analogous W83 version is 807 nucleotides. In this case the length difference is due to W83 BrickBuilt_5 having fewer 23 nt internal direct repeats; the leader and tails are of the same length. Within a multiple alignment of the four *P. gingivalis* strains at the BrickBuilt_5 locus there are 18 single nucleotide polymorphisms (SNPs) that separate the lineages (Fig.3-4). Substituting SNPs between strains at the BrickBuilt_5 locus into the ATCC 33277 model changes the predicted entropy of the element by -7, or 5%, to -135. Thus, the majority of the entropy differences can be attributed to the 23 bp repeat numbers. However, a slightly more stable structure may form with the SNPs found naturally in strain W83.

Although no BLASTn matches for BrickBuilt were found outside of the *P. gingivalis* species, MITEs from other species have been shown or are predicted to be of similar modular makeup and form long stem loops (Zhou, Tran, & Xu, 2008) (Ilyina, 2010) (Delihas, 2011)(Deng et al., 2013) (Satovic & Plohl, 2013)(Zhang, Xu, Shen, Han, & Zhang, 2013). In addition, repetitive sequences that are not MITE-associated also frequently form stem loop structures (De Gregorio, Silvestro, Petrillo, Carlomagno, & Di Nocera, 2005) (Petrillo, Silvestro, Di Nocera, Boccia, & Paolella, 2006) (Bertels & Rainey, 2011) (Delihas, 2011). Stem loop structures, especially long stem loops, are capable of modulating transcript half-life, modulating translational efficiency as well as serving as docking/receptor sites for proteins (Petrillo et al., 2006) (Delihas, 2011) (Bainbridge et al., 2015). The Rho-independent terminator upstream of BrickBuilt_5 is located 111-148nt from the 5' CDS, with the leader

region of BrickBuilt_5 located 182 nt from the 5' CDS (Fig.3-3). Thus, in this case, the BrickBuilt element has not disrupted the 'natural' terminator for the 5' CDS. However, given the promoter capabilities and proximity to the Rho-independent terminator, this BrickBuilt element may be able to modulate the stability or accessibility of the terminator. The long stem loop structures of BrickBuilt_5 start 257 nt from the Rho-independent terminator and end 965 nt away.

3.3.5 Genome Locations and Surroundings

All BrickBuilt elements are located intergenically; no direct overlap or interruptions of genuine protein coding sequences are apparent in the complete genomes available to date (Table 3-1 and Supplemental Table 3-1). Several 'hypothetical proteins' are annotated to be within BrickBuilt elements, however expression of these proteins has not been confirmed experimentally (Xia et al., 2007) (Kuboniwa et al., 2009) (Maeda et al., 2014). Several of the predicted hypothetical proteins are part of repeated/overlapping probes on *P. gingivalis* microarrays. Additionally, the 23 nt repeats within BrickBuilt elements are predicted to cause frequent translational stops (data not shown). Lack of experimental confirmation of protein products, nonunique microarray probes and abundant translational stops suggests that translation of these regions is unlikely, and even if translation were to occur it would probably be truncated versions of a repetitive or mobile element.

Of the 38 genes surrounding the BrickBuilt elements in the ATCC 33277 genome there are several functional clusters (Table 3-1). Six genes encode proteases of the C1, (2) C10, (2) M16 and S41 families. Five genes are predicted to encode DNA/RNA-binding proteins, and another four are involved in tRNA metabolism. Noticeably, of the 19

BrickBuilt elements in strain ATCC 33277, 5 are located adjacent to a gene/protein containing a Por Secretion System C-Terminal Domain (PorSS CTD) (Sato, 2011) (Table 3-1). Likewise, 5 of the previously identified *P. gingivalis* MITEs within the W83 genome are also located next to PorSS CTDs. A total of 34 PorSS CTDs have been predicted within the W83 genome (only 22 annotated on NCBI with TIGR); 29 percent of PorSS CTDs are associated with MITEs (Seers et al., 2006). The PorSS is connected to pigmentation and haem acquisition in *P. gingivalis*. Apart from those associated with PorSS, other genes surrounding BrickBuilt elements are *hemG*, *dps, trx, and hmuY,* which are involved in haem biosynthesis, detoxification and acquisition, respectively. Additionally, two separate DUF 805 motifs are found in genes surrounding BrickBuilt elements, which are associated with phage integrases.

Expansion and contraction of the 23 nt repeats between strains is evident at the conserved BrickBuilt loci, albeit a static snapshot of a potentially highly dynamic region. Entire 23 nt repeat segments have been removed or added. Full and/or partial deletions of the leader and tail regions are also apparent. Deletions of the leader and tail regions occur from the distal ends of each segment with respect to the 23 nt internal repeats.

Pairwise and multiple alignments of a respective BrickBuilt locus across the four strains of *P. gingivalis* revealed SNPs that potentially suggest lineages or selected and compensatory mutations (Fig.3-5 and 3-6). Whether the SNPs are generated *de novo* at each site, occur in stages and are distributed, or occur through site-to-site recombination cannot be determined definitively from currently published genome assemblies alone. However, multiple alignments of the conserved BrickBuilt loci within a given strain show patterns of non-random mutation (Fig.3-6). For sites at which SNPs have occurred, the SNP is

frequently distributed at several positions within the element, yet this occurs at intervals. Additionally, SNPs appear localized around a 2-4 nt site when compared in multiple alignments (Fig.3-6). Long Term Evolution Experiments (LTEE) and plasmid-based recombination systems could be employed to determine mutation rates within BrickBuilt in comparison to the rest of the genome, whether 23 nt repeats expand and contract at a given locus, and how recombinogenic the elements are.

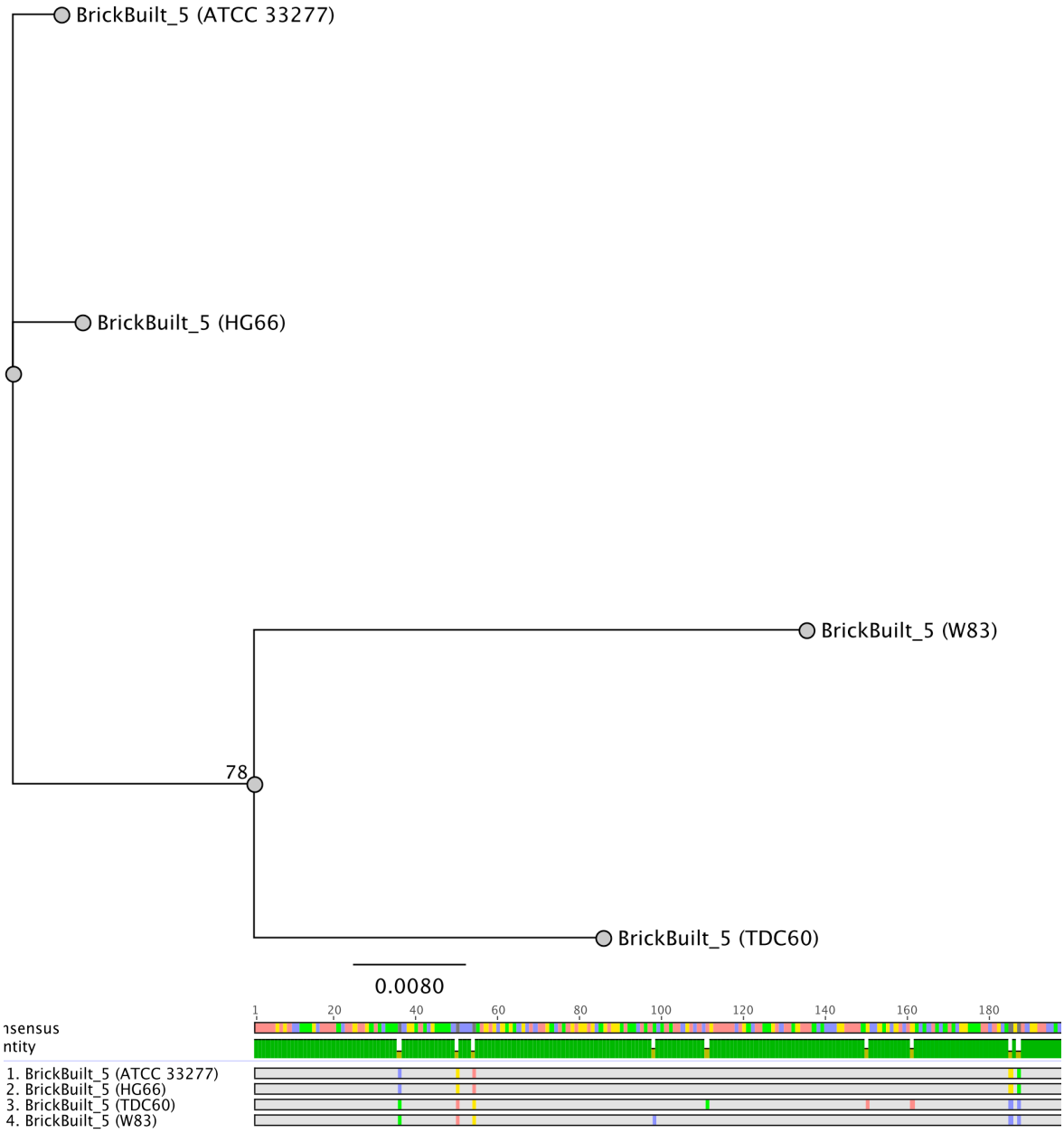


Figure 3-5: BrickBuilt_5 phylogenetic tree (top) and multiple alignment (bottom) generated using PHYML and MAFFT on the Geneious R8 platform with strain ATCC 33277, W83, TDC60 and HG66 inputs. The bootstrap value of 78 for the node separating strains W83 and TDC60 was generated through 100 total iterations. The multiple alignment is focused on the first 200 nt of the leader region. Grey within the tracks denotes complete conservation at a site. Color within the tracks denotes variable sites. At all five sites where more than one strain is variable strains ATCC 33277 and HG66 cluster together, as do strains TDC60 and W83. Within this region only strains TDC60 and W83 have sites where they alone differ from the other three strains; this is consistent throughout for this specific element.

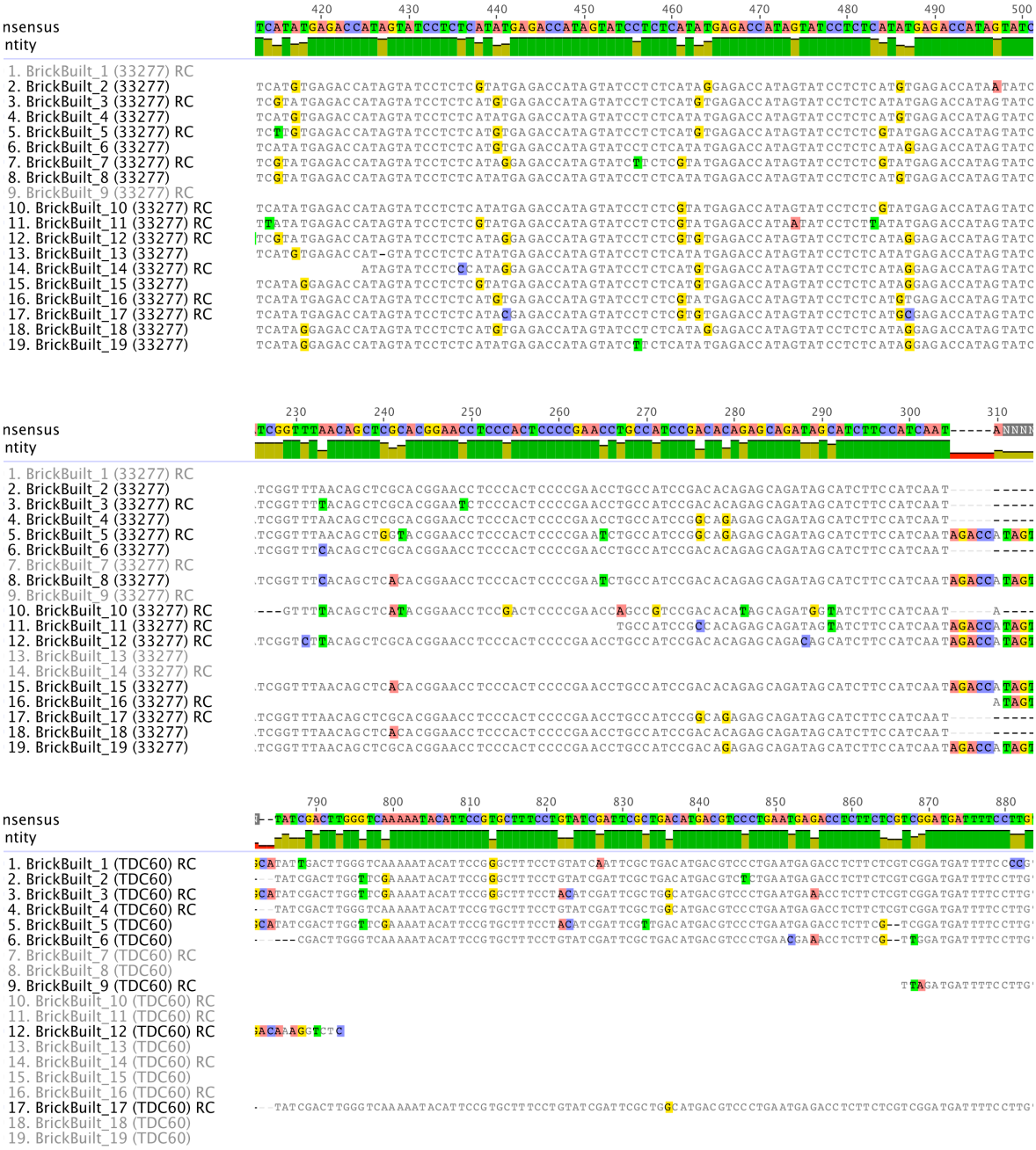


Figure 3-6: BrickBuilt element multiple alignments using Geneious R8 platform. The top panel depicts a segment from the 19-element multiple alignment of strain ATCC 33277 focused on the 23 nucleotide repeat region. The middle panel depicts a segment from the 19-element multiple alignment of strain ATCC 33277 focused on the leader region. The bottom panel depicts a segment from the 19-element multiple alignment of strain TDC60 focused on the tail region. Highlighted regions denote regions with SNPs relative to the consensus sequence.

Multiple alignments of BrickBuilt elements using the strains ATCC 33277, W83, TDC60 and HG66 allowed for the observation that strain ATCC 33277 shares greater local identity at the nucleic acid level with HG66, and strain W83 with TDC60. These similarities are consistent in phylogenetic analyses (Fig.3-5). The matching of the strains in alignments is consistent throughout 18 of 19 elements. Branching of BrickBuilt elements is congruent with the dendrogram generated based on genomic BLAST for all *P. gingivalis* 13 genomes. Interestingly, strain HG66 was deposited as 'being closely related to strain W83', yet as the full genome BLAST from NCBI shows in the dendrogram as well as how the BrickBuilt elements align this seems to be incorrect. The BrickBuilt_5, BrickBuilt_8, BrickBuilt_10, BrickBuilt_11, BrickBuilt_12, BrickBuilt_13, BrickBuilt_14 and BrickBuilt_15 sites all lie between the same two genes; strains ATCC 33277 and HG66 usually having more 23 nucleotide repeats than W83 and TDC60. BrickBuilt_6 is the only site at which strains W83 and TDC60 have more 23 nt repeats than ATCC 33277 and HG66. Of note, based on type of alignment program and parameters slightly different alignments occur, which seems highly dependent on the internal 23 nt repeats and gap penalties. BrickBuilt_9, the shortest element which is also the only element without a 23 nt repeat has only one SNP across the 100 nucleotides. Unlike all the other BrickBuilts, that single SNP would align strain ATCC 33277 with W83 and strain TDC60 with HG66.

BrickBuilt_4 may best demonstrate the lingering 'mobility' of BrickBuilt elements. Strain HG66 shares the same locus with strain ATCC 33277. However, the TDC60 BrickBuilt_4 is not located between the same two genes (Fig.3-7). No other mis-located BrickBuilt elements occur in strain TDC60 and the element aligns very closely with the other

strain's versions. Thus, it is probable that the BrickBuilt_4 homologue has been induced to transpose by or transposed with the surrounding *ISPg1* elements. Additionally, no BrickBuilt_4 homologue is present in strain W83, adding to a mobility pattern of BrickBuilt_4 (Fig.3-7).

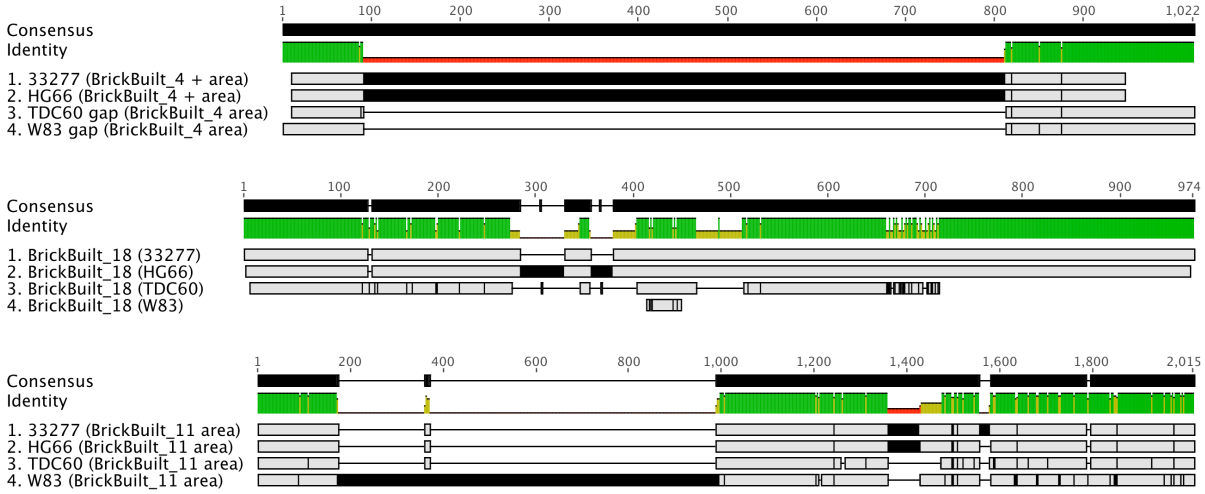


Figure 3-7: MAFFT-based alignments of aberrant BrickBuilt elements and areas across *P. gingivalis* strains. The top panel depicts the CDS-to-CDS region of BrickBuilt_4, using the surrounding CDS that would correctly correspond to the ATCC 33277 genome. Strain W83 has no BrickBuilt_4 and strain TDC60 has a BrickBuilt_4 that has moved or been moved to a different locus. The middle panel depicts BrickBuilt_18, in which strains ATCC 33277 and HG66 have a 236-type MITE within the BrickBuilt element. The bottom panel depicts the CDS-to-CDs region BrickBuilt_11 from strain ATCC 33277, in which strain W83 has acquired a gene immediately upstream of the BrickBuilt element. Light grey boxing indicates completely identical sequence regions. Black lines or boxing indicates areas of aberrance (e.g. SNPs or additional IS-like element).

BrickBuilt_18 in the strains ATCC 33277 and HG66 contain/encompass MITE239_11 between nucleotides 659-904 of the sequence. Strain TDC60 has a gap where MITE239_11 occurs in the other two strains, while the flanking portions of the BrickBuilt match (Fig.3-7). The strain W83 version at this site is diminutive, having been reduced to 1.5 copies of the 23 nt internal repeat. While strain W83 doesn't harbor a MITE-within-a-MITE configuration at any locus, BrickBuilt_11 in strain W83 contains an 'extra' gene adjacent to the 5' region of the element unlike any other strain (Fig.3-7).

3.3.6 Transcriptional Expression of BrickBuilt

Repetitive and transposable elements are capable of modulating the genome stability and evolution of species (Treangen et al., 2009)(Delihas, 2011) (Bennetzen & Wang, 2014)(Padeken et al., 2015). Interestingly, no endogenous plasmids have been found for *P. gingivalis* to date. The presence of many copies and types of repetitive and transposable elements could serve a quick way by which *P. gingivalis* could recombine/adapt to external stimuli beyond traditional host-directed transcriptional and translational controls (Slotkin & Martienssen, 2007) (Feschotte, 2008) (Treangen et al., 2009) (Lisch, 2013) (Bennetzen & Wang, 2014) (Zhou et al., 2014).

Analysis of previously published data

Previous microarray and RNAseq studies have shown transcripts originating from within BrickBuilt elements, yet none characterized these regions in detail (Chen et al., 2004)(Høvik et al., 2012). Several of the microarray probes are themselves repetitive and many of the oligos/~20 mers used for identifying transcripts could map to multiple sites within the genome. Although BrickBuilt elements are highly conserved and repetitive, small

variations due to SNPs, size of leader and tail regions, and the surrounding intergenic context make it possible to map at least some transcripts to the correct sites (Fig.3-8 and Supplemental Fig.3-4). For situations where completely identical regions could produce the same transcript, the mapping programs and settings used will determine whether the transcripts are placed at one of the matching loci exclusively, distributed amongst the loci evenly, or left out of the results entirely. Importantly, the placement of any transcript at one or distributed across all of a given repetitive oligo/~20 mer sites suggests that at least one of the sites contributes active transcription.

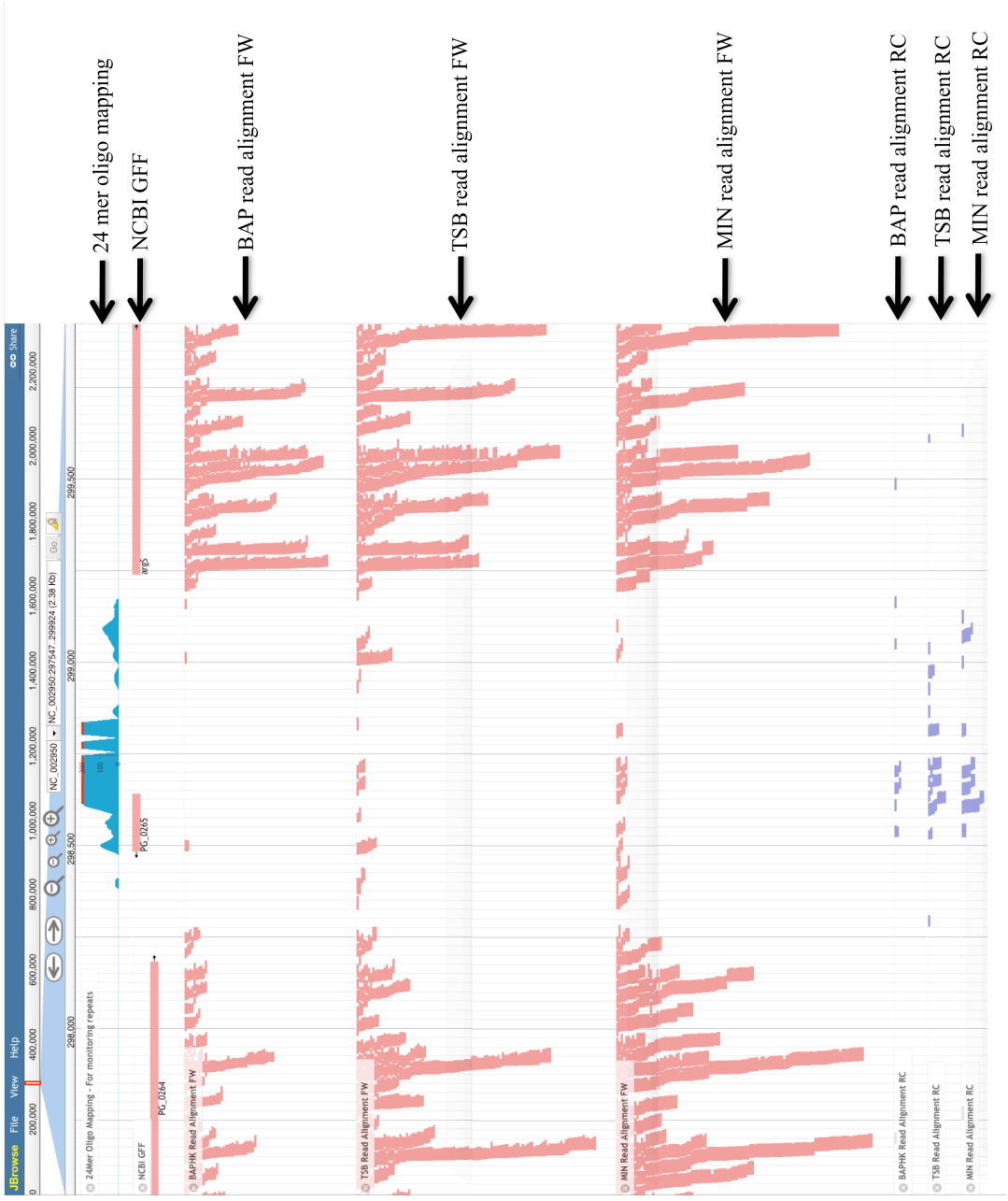
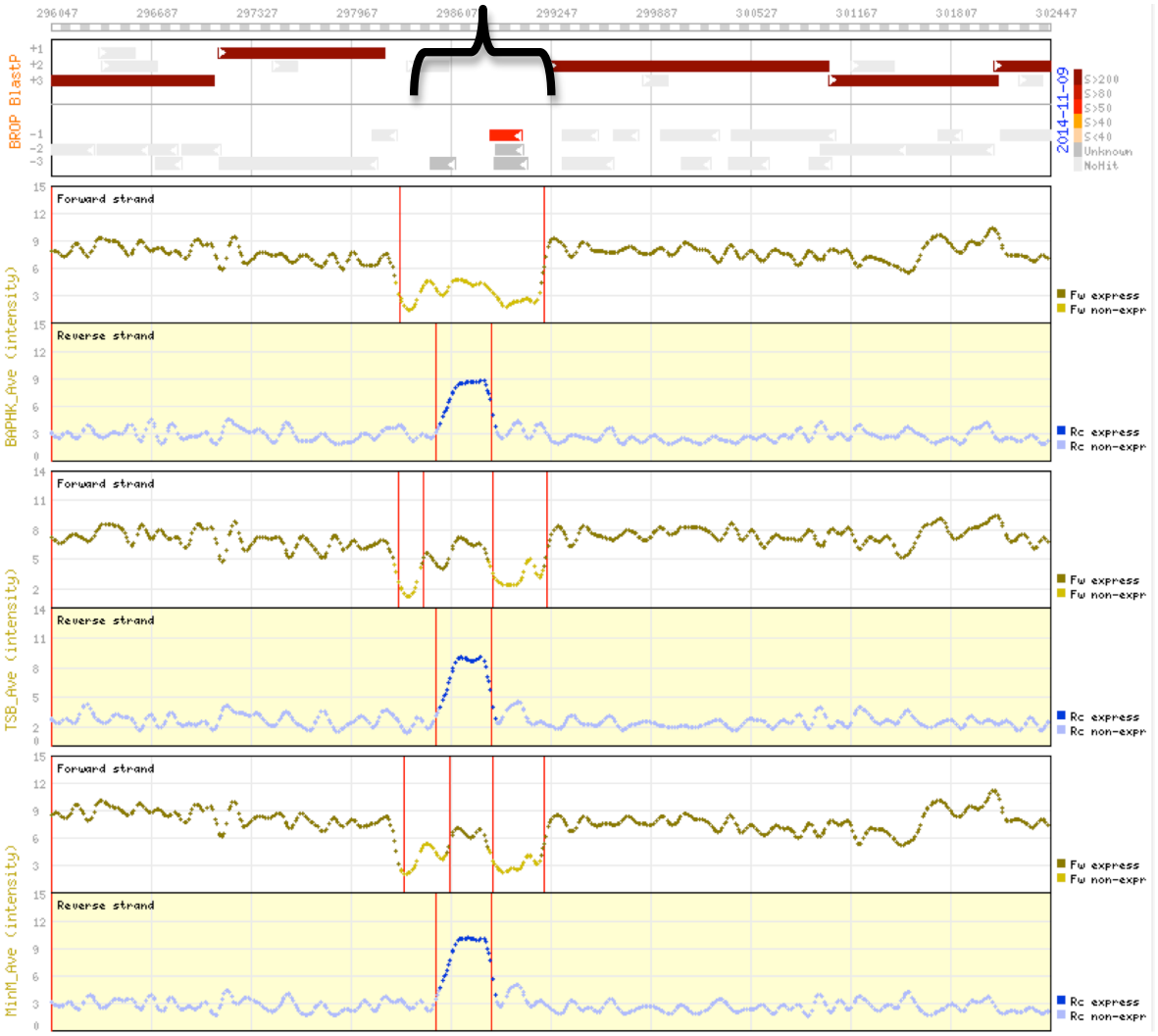


Figure 3-8: RNAseq display of transcripts of/from BrickBuilt_5 and surrounding area in strain W83 using JBrowse. Only uniquely mappable transcripts are displayed. Red horizontal lines correspond to forward strand-based transcripts from blood agar, tryptic soy and minimal media, respectively. Blue horizontal lines correspond to reverse strand-based transcripts from blood agar, tryptic soy and minimal media, respectively. (Full screen PDFs or screenshots are not currently possible with JBrowse, thus three different panels had to be compiled for the image. Direct link to data: http://bioinformatics.forsyth.org/jbrowse/index.html?data=PgRNAseq%2Fjson&loc=NC_002950%3A297655..299836&tracks=24Mer_Repeat%2Cncbigff%2Cbaphk_fw_bam%2Ctsb_fw_bam%2Cmin_fw_bam%2Cbaphk_rc_bam%2Ctsb_rc_bam%2Cmin_rc_bam&highlight.



Supplemental Figure 3-4: Microarray display of transcripts in BROP MTD database of BrickBuilt_5 and surrounding area in strain W83. Tracts represent positive and negative strand blood agar (top), tryptic soy broth (middle) and minimal media (bottom), respectively. BrickBuilt_5 noted with black bracket.

Within the transcriptome transcript levels of individual BrickBuilt elements vary markedly and also vary according to growth medium, e.g. in/on minimal, tryptic soy, and blood media (Chen et al., 2004)(Høvik et al., 2012). Generally, transcripts from BrickBuilt regions are lowest on the blood-containing media. Transcript levels and distribution of transcripts of BrickBuilt elements, using strain W83 RNAseq data, can be grouped generally into three categories. Group one, displaying relatively high transcript levels throughout the element on only one strand bridging the entire CDS-to-CDS gap, includes BrickBuilt_1, 3, and 13. BrickBuilt_13, comprised only of the internal 23 nt repeats. The element's expression correlates directly with that of the upstream gene, thus, the expression of BrickBuilt_13 could be completely due to transcriptional read-through from adjacent genes. Consistent with this, BrickBuilt_13-associated transcripts are all on the negative strand. Group two, displaying low to medium intermittent transcript in tryptic soy and minimal media but none on blood agar, includes BrickBuilt_2, 5, 6, 8, 10, 11, 12, 14, 15, 16, 17, and 19. Group three, displaying no transcript yet adjacent to upstream transcript that is well beyond an annotated CDS, includes only BrickBuilt_9.

Additional information about BrickBuilt elements and their surrounding regions can be garnered from the above microarrays and RNAseq studies as well as additional studies that have been carried out with *P. gingivalis* under defined conditions. High-density tiling microarray of *P. gingivalis* strain W83 by Chen et al. showed differential expression of BrickBuilt elements at several loci (Chen et al., 2004). Using a W83 strain based microarray, genes PG0626 and PG0089 were found to be aberrant in strain ATCC 33277, corresponding

to BrickBuilt_ 11 and BrickBuilt_ 19 loci. The area in and around BrickBuilt_10 was identified as a potential sRNA (sRNA35) by Philips et al. (Phillips, Progulske-Fox, Grieshaber, & Grieshaber, 2014). The highest expression of the putative sRNA35 occurred during mid-log cultures grown under hemin excess conditions after an initial period of hemin starvation. Under the experimental methods used by Philips et al., no other BrickBuilt loci were determined to be or be part of putative sRNAs expressed in response to hemin-variable growth conditions. BrickBuilt elements are not directly affected by *FimR* or *LuxS* regulation (Nishikawa, Yoshimura, & Duncan, 2004) (Hirano, Beck, Demuth, Hackett, & Lamont, 2012). However, genes surrounding BrickBuilt elements are regulated by *LuxS*. Lack of expression as well as partial expression of annotated genes surrounding BrickBuilt elements is evident from *P. gingivalis* strain W83 transcriptomic analyses by Hovik et al. (Høvik et al., 2012). Five of the conserved 30 genes flanking BrickBuilt elements in the W83 genome are predicted to not be expressed and 11 (including 3 of the 5 'not expressed') give partial or abortive transcripts in blood, tryptic soy or minimal media.

The genomic association with haem biosynthesis and pigmentation-associated genes in conjunction with transcriptional data from RNAseq and microarray studies may point to regulation of BrickBuilt regions by haem or iron. DNA tandem repeats have been shown previously to affect transcription of iron and haem-associated genes (Chen & Li, 2007)(Zhou et al., 2014).

E. coli expression vectors

Promoter probe vectors pCB182 and pCB192 were used to determine the potential for transcription and transcriptional regulation of the full BrickBuilt element and segments in a heterologous host, *E. coli*. Four potential promoter sites were hypothesized based on previous

RNAseq and microarray data (Fig.3-9). Four configurations of the leader, tail and element were constructed using BrickBuilt_5 as a template: full element in leader-to-tail orientation ('normal', with tail abutting *lacZ*); full element in tail-to-leader orientation ('reverse', with flipped leader abutting *lacZ*); tail-only in reverse orientation; and leader-only in forward orientation. The reverse orientation of the full element, with the beginning of the leader abutting the promoterless *lacZ*, displayed the greatest promoter activity of the four constructs (Fig.3-10). All four constructs displayed statistically significant expression. No expression from the vectors lacking inserts was seen on plates with X-gal, while each insert-containing construct showed blue colonies due to expression by 24 hr (Supplemental Fig.3-5). Expression from these constructs demonstrates bi-directional promoter ability (when tested in an *E. coli* system) within the tail segment as well as in the leader segment facing out of the element.

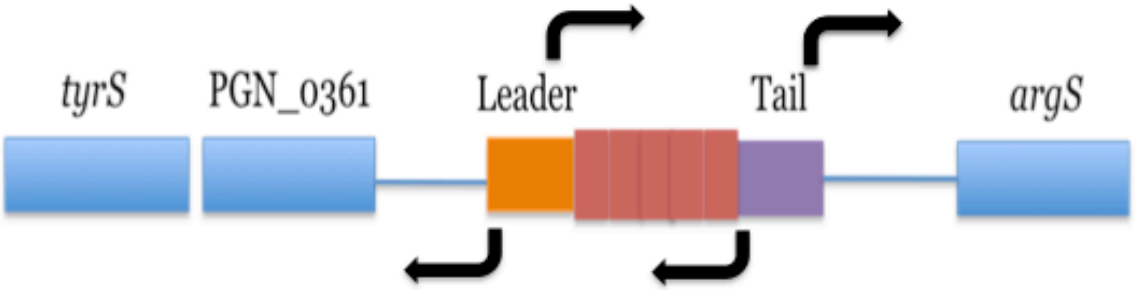


Figure 3-9: Model of promoter capabilities of BrickBuilt_5 based on *lacZ* promoter probe ability. Bi-directional promoters are present in both the leader and tail segments of BrickBuilt_5. As such, at this locus antisense transcripts may be produced toward PGN_0361, sense transcripts produced toward arginyl-tRNA synthetase (*argS*), and transcripts of the 23 nucleotide repeat regions within the element may be produced from both strands. The distance from the tail to *argS* is less than 100 nt and may be or contain the promoter for *argS*.

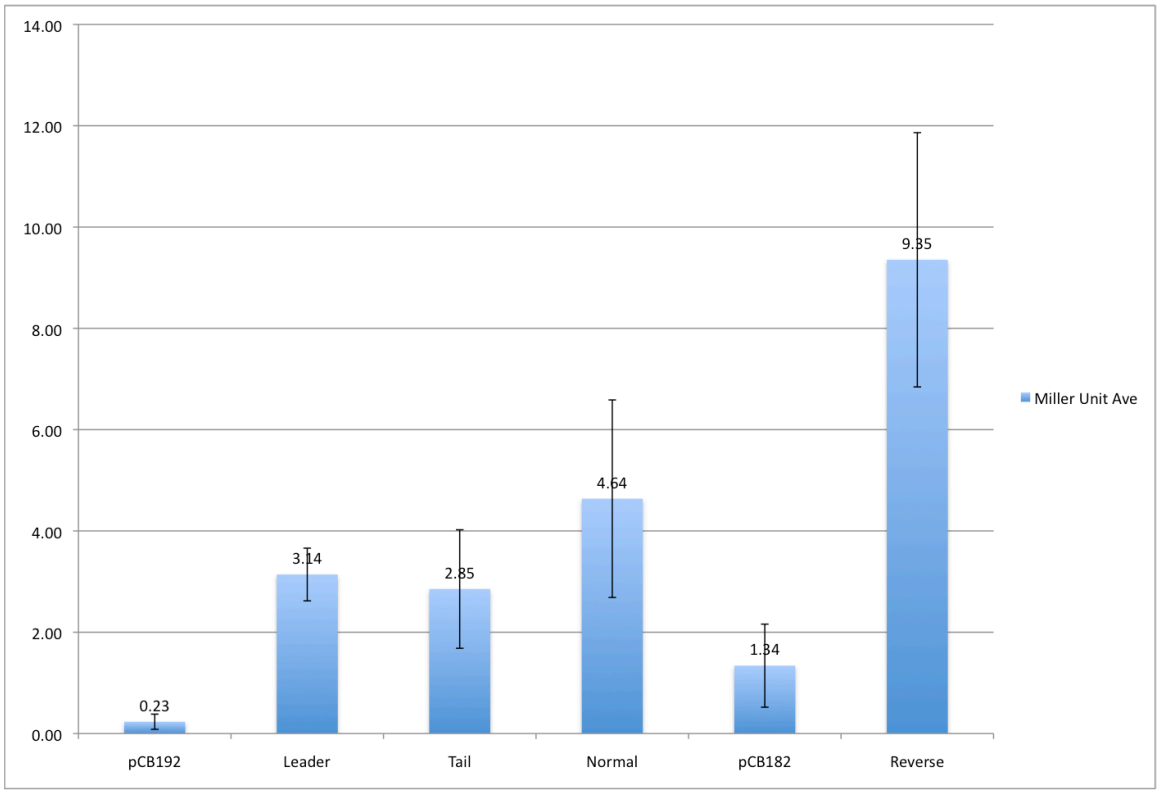
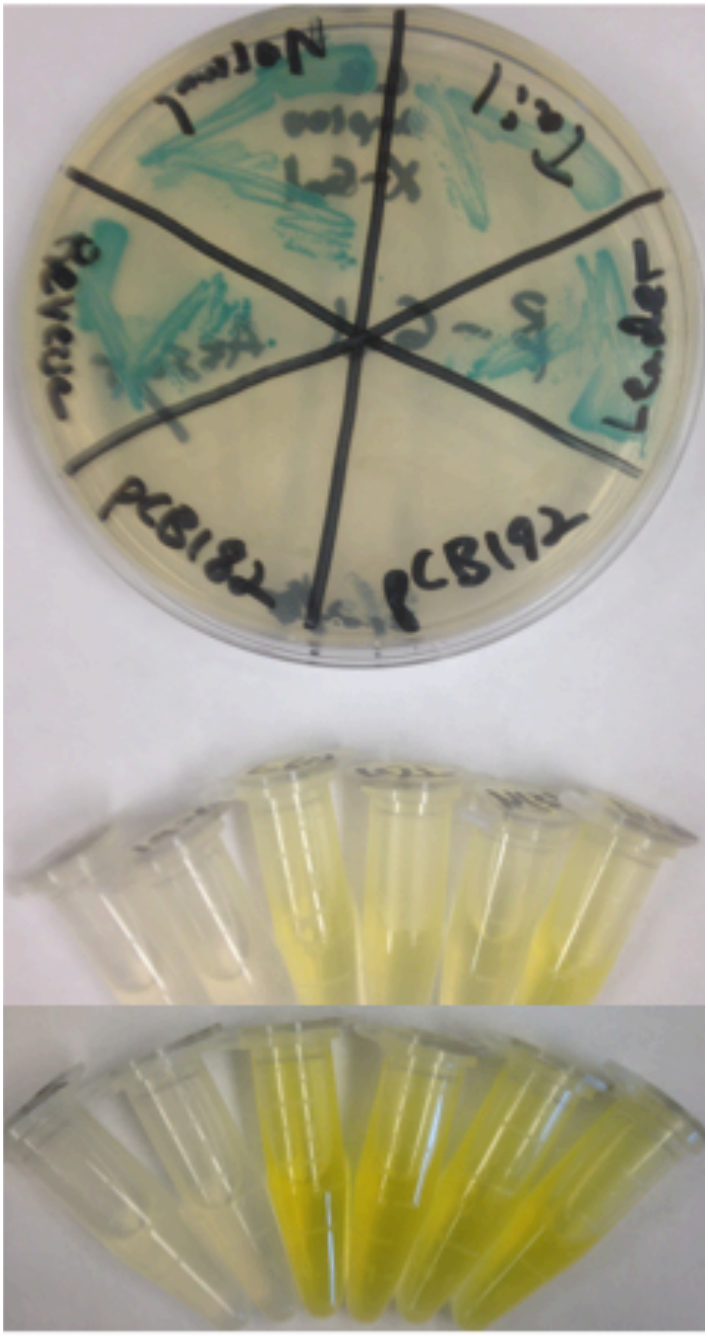


Figure 3-10: ONPG assays for promoter capabilities of BrickBuilt_5 based on *lacZ* promoter probe constructs. Promoter-less *lacZ* backbone vectors pCB182 and pCB192 give low apparent β –Galactosidase activity. β –Galactosidase activity measured through ONPG cleavage after 3 hr incubation at 28°C. Error bars represent standard deviation between biological replicates in triplicate. Statistical significance determined by t-test ($p < 0.05$).



Supplemental Figure 3-5: X-gal and ONPG assays of promoter capabilities of BrickBuilt_5 based on *lacZ* promoter probe constructs. Promoter-less *lacZ* vectors pCB182 and pCB192 give no apparent β –Galactosidase activity. BrickBuilt_5 leader and tail oligos (Eurofins Operon) were cloned into vector pCB192. Full-length BrickBuilt_5 was cloned into pCB192, 'Normal', with the tail segment of the element upstream of *lacZ* facing in the same orientation (tail abutting *lacZ*). Full-length BrickBuilt_5 was cloned into pCB182, 'Reverse', with the leader segment of the element upstream of *lacZ* facing in the same orientation (flipped leader abutting *lacZ*). X-gal activity visualized after 24 hr incubation. Top of two liquid assays of ONPG activity visualized after 3 hr incubation; bottom after 24 hr incubation.

3.4 Conclusions

We identified and provided preliminary characterization of a genetic element, 'BrickBuilt', in the genome of *Porphyromonas gingivalis*. BrickBuilt appears to be a MITE-like element that has trapped a 23 nt direct repeat; propagating itself and the direct repeat throughout the genome. From promoter-less *lacZ* assays and analyses of previous microarray and RNAseq data we determined BrickBuilt contains promoter elements capable of bi-directional transcription. Given the element's exclusively intergenic locations and surrounding gene directionality, these transcripts may serve to regulate expression of surrounding genes. Relative stability of locations, overall copy number and expression levels of the elements throughout the sequenced *P. gingivalis* genomes point to neutral or advantageous maintenance of BrickBuilt.

Further sequencing projects and phylogenomics will be necessary to determine which other species and strains contain the BrickBuilt element and at what evolutionary point these species and/or strains diverged. Additionally, strain-specific experimental evolution and plasmid-based recombination systems could be employed to determine mutation rates within BrickBuilt in comparison to the rest of the genome, whether and how 23 nt repeats expand and contract at a given locus, and how recombinogenic the elements are.

Adding BrickBuilt to the list of transposable and repetitive elements types in *P. gingivalis* brings the current total to 4 MITEs, 11 ISs, 2 Ctn and 1 Tn. The ORFs and total base pairs encompassed by these elements constitute an impressive proportion of the genome; with BrickBuilt alone accounting for 0.44 percent of bases. The ability of several of these elements being involved in genome evolution has been established (Duncan, 2003)

(Tribble et al., 2013). However, the full effects of these elements on genome stability and evolution as well as transcriptional, translational and post-translational response to stimuli remains to be experimentally determined.

3.5 Methods

3.5.1 Genomes and Strains

Genome sequence FASTA and GenBank files were downloaded from the NCBI database. At the time of this research, strains ATCC 33277, TDC60, W83, HG66 and JCVI SC001 were available as completed sequencing and assembly projects (ATCC 33277, TDC60 and W83 as 'gapless chromosome' status, HG66 as a single contig, and JCVI SC001 as a draft of many stitched contigs) (Nelson et al., 2003) (Naito et al., 2008) (Watanabe et al., 2011) (McLean et al., 2013) (Siddiqui et al., 2014). The five sequenced wild-type strains are disparate based on origin or lineage: W83 isolated in Germany (1950's) from an oral lesion; ATCC 33277 was isolated in the USA (1980's) from subgingival plaque; TDC60 was isolated in Japan (2011) from an oral lesion; HG66 isolated in the USA (1989) from a dental school patient; and JCVI SC001 was isolated in the USA (2013) from a hospital sink. The sequencing projects utilized different sequencing and assembly methods; each providing a *de novo* assembly. The JCVI SC001 genome sequence contained unidentified bases and residual gaps in the sequencing after the completed project.

3.5.2 Sequence Analysis, Clustering, Alignment, Phylogenetics/Phylogenomics

NCBI BLAST suites were utilized to determine locations, structure and potential protein-coding capacity of the MITEs (Altschul et al., 1990). Query inputs were FASTA

sequences taken directly from NCBI genome sequencing projects. For initial characterizations prior to determining species-specificity of the elements, the entire NCBI sequence database was queried. Following determination that the elements were only found (as of 11/2013) in the genomes of *P. gingivalis* strains, subsequence queries were focused to either the *P. gingivalis* species as a whole or specific *P. gingivalis* strains. Megablast, discontiguous megablast and BLASTn program selections for search optimization were all used in determining species-specificity as well as genome localizations.

MultAlin, Clustal Omega and the MEME suite were used to perform DNA-based and amino acid-based multiple alignments of the MITEs to determine conserved nucleotides and the start and stop points of the elements as well as proteins surrounding the MITEs. Amino acid-based alignments were used to determine whether the surrounding genes had structural domains at either the 5' or 3' ends that could potentially account for or facilitate MITE localization (Corpet, 1988) (Bailey et al., 2009) (Sievers et al., 2011).

The BioCyc sequence pattern search tool 'PatMatch' was used to determine the number and genomic location of *P. gingivalis* MITEs in strain ATCC 33277 (Karp et al., 2010). PatMatch identifies potential sites given variations in the consensus sequences of the MITE direct repeats, TIRs, 'leader' and 'tail' regions because different mismatch numbers are allowed. Query inputs were nucleotide consensus sequences determined for each of the given parts of the MITE. Both DNA strands, as well as intergenic and coding sequences, were queried separately. Mismatches of '0' through '3' were allowed, with the constraint of the 'mismatch type' being a substitution.

The Tandem Repeats Database software was used for determining all types of tandem repeat elements in the *P. gingivalis* genomes (strains ATCC 33277 and W83 hosted on the

server as of 12/2013), and to determine whether the tandem repeats or MITE as a whole was conserved in other sequenced species (Gelfand, Rodriguez, & Benson, 2007). BLAST query of the entire bacterial and viral tandem repeat database was carried out using the FASTA sequence downloaded from NCBI for *P. gingivalis* strain ATCC 33277 MITE. The Tandem Repeats Finder software was used to determine the composition of the *P. gingivalis* tandem repeat element (Benson, 1999). 'Basic' sequence analysis was selected for queries. Tandem Repeats Finder was also used to determine repeat conservation within and between loci as well as where a given element started and ended.

The Geneious software platform (version R8) was used to download, store, deposit, manipulate and query *P. gingivalis* genomes and BrickBuilt MITEs (Kearse et al., 2012).

3.5.3 MITE and Surrounding Coding Sequences' Nucleic Acid and Protein Motif Analysis

The Pfam and InterProScan databases and programs software were used to determine the presence and characteristics of nucleic acid and protein motifs (Finn et al., 2014) (Jones et al., 2014). Query inputs were FASTA sequences from NCBI download files. For Pfam, an E-value of 1.0 and checking Pfam-B motifs were selected options prior to submission.

ExPASy Translate Tool software was used to determine whether the MITEs potentially encoded proteins, and thus are not strictly nucleic acid elements (Artimo et al., 2012). Genetic code option 'standard' was used for all queries. All six possible frames of translation were considered.

Modeling and structure prediction programs Mfold, RNAstructure and RegRNA2.0 were used to predict potential 2-D structures formed by MITE DNA and RNA (Zuker, 2003)

(Bellaousov et al., 2013) (Chang et al., 2013). Default options for programs in regard to structure prediction were chosen.

3.5.4 Cloning and Reporter Strains, Media and Growth Conditions

Escherichia coli DH5α and TOP10 were used for cloning, plasmid maintenance and transcriptional assays. Ampicillin (100 μg/ml) was used when appropriate for prevention of contamination as well as isolation and maintenance of transformants containing plasmids. Strains were grown and maintained on LB [Lennox] agar or in LB [Lennox] broth (Invitrogen).

PCR primers containing BamHI and XmaI (NEB) restriction sites were designed immediately flanking the BrickBuilt_5 MITE associated with PGN_0361 (Supplemental Table 2). PCR products were generated using GoTaqLong Master Mix (Promega) and resultant bands were cloned into vector pCR TOPO-XL (Invitrogen), which was transformed into *E. coli* DH5α. Transformants were selected for kanamycin resistance and clones were confirmed by restriction digest and sequencing.

To generate constructs using the promoter probe vectors pCB182 and pCB192, pCR2.1 and pCR-TOPO-XL cloned BrickBuilt MITE constructs and pCB182/pCB192 were double-digested with BamHI (NEB) and XbaI (NEB) or BamHI and XmaI (NEB) and then transformed into *E. coli* TOP10. Transformants were selected for ampicillin resistance generated by an insertion event. Clones were confirmed by restriction digest and sequencing.

3.5.5 Transcriptional Analyses

BROP, specifically the 'Genomics Tools for Oral Pathogens' and 'Microbial Transcriptome Database' sections of the resource, were used to determine genome location, characteristics of coding sequencings surrounding MITEs, differences between strains as well as transcriptome data (Chen et al., 2005). Over the course of the research, two different variations of the RNAseq data for *P. gingivalis* strain W83 were supported, one directly on BROP and then a later form on JBrowse. The JBrowse form gives greater functionality in displaying data and visualization (Westesson, Skinner, & Holmes, 2013). Under the 'Genomics Tools for Oral Pathogens' subset, the 'GenomeViewer' function was used to compare genome arrangements of *P. gingivalis* strains ATCC 33277 and W83 with relation to MITEs, as well as display the previously performed microarray data (under the strain W83 section) for MITE-associated genome areas under the three different nutrient conditions (the same conditions performed in the RNAseq) (Chen et al., 2004)(Høvik et al., 2012).

β –Galactosidase Assays

Escherichia coli strains were grown and maintained in Luria-Bertani (LB) media supplemented with ampicillin (100 μg/ml) as required. PCR primers and synthesized oligos used for strain constructions are listed in Supplemental Table 2. The pCB182 and pCB192 vectors lack promoters but contain translational start codons (Schneider & Beck, 1986). As such, gene expression of *lacZ*, and in turn protein expression of LacZ read out through β –galactosidase activity, should be the result of promoter activity from fragments cloned into the vector. β -galactosidase assays were performed under plate-based (X-Gal) and broth-based (ONPG) setups. For plate-based assays, frozen stock cultures of the BrickBuilt MITE derivatives transcriptionally fused to *lacZ* in their respective *E. coli* strains were plated onto LB agar containing X-gal and ampicillin. For broth-based assays, cultures of the BrickBuilt

MITE derivatives transcriptionally fused to *lacZ* in their respective *E. coli* strains were grown in LB broth for 3 hr with shaking at 37°C. An aliquot of each culture (500 µl) was added to a lysis and assay solution mixture (500 µl), vortexed, and then incubated at 28°C for 3 hr. Color development was measured spectrophotometrically at OD_{420} nm and cell debris at OD_{550} nm. Respective Miller units were calculated as previously described (Wang, Lin, Kidder, Telford, & Hu, 2002).

Table 3-1: Genes/Coding Sequences located 5' and 3' to BrickBuilt elements. Gene numbers and characterizations correspond to strain ATCC 33277. Loci of BrickBuilt elements across four sequenced and annotated *P. gingivalis* strains. BrickBuilt elements are situated intergenically between the genes noted. Grayed-out boxes represent loci at which BrickBuilt is aberrant. 'NP' stands for 'Not Present'.

MITE	33277 Locus	W83 Locus	TDC60 Locus	HG66 Locus	Gene Characterization	Gene	Strand
BrickBuilt_1	PGN_0031	PG0033	PGTDC60_0032	EG14_02395	RmuC domain (DUF 805)		-
	PGN_0033	PG0034	PGTDC60_0034	EG14_02400	Thioredoxin	*trx*	-
BrickBuilt_2	PGN_0204	PG2159	PGTDC60_1262	EG14_03005	Protoporphyrinogen oxidase	*hemG*	+
	PGN_0205	PG2161	PGTDC60_1265	EG14_03010	AraC family transcriptional regulator		-
BrickBuilt_3	PGN_0303	PG0196	PGTDC60_0466	EG14_04795	Zinc protease (Peptidase M16)		+
	PGN_0306	PG0198	PGTDC60_0471	EG14_04805	PF05656 family protein (DUF 805)		+
BrickBuilt_4	PGN_0336	NP	PGTDC60_1661	EG14_04940	Immunoreactive antigen / PorSS CTD		-
	PGN_0340	NP	PGTDC60_1665	EG14_04960	Peptidase S41 / PorSS CTD		+
BrickBuilt_5	PGN_0361	PG0264	PGTDC60_0543	EG14_05065	Glycosyl transferase family 2		+
	PGN_0365	PG0267	PGTDC60_0547	EG14_05075	Arginyl-tRNA synthetase	*argS*	+
BrickBuilt_6	PGN_0400	PG1715	PGTDC60_0586	EG14_05255	TonB-dependent receptor Cna protein		+
	PGN_0403	PG1714	PGTDC60_0590	EG14_05265	Pyridoxamine-phosphate oxidase	*pdxH*	+
BrickBuilt_7	PGN_0455	PG0549	PGTDC60_0639	EG14_03135	Partial ISPg5		+
	PGN_0456	PG0553	PGTDC60_0641	EG14_03150	Methylmalonyl-CoA mutase	*scpA*	-
BrickBuilt_8	PGN_0550	PG1559	PGTDC60_0739	EG14_03610	Glycine cleavage system subunit T	*gcvT*	+
	PGN_0553	PG1556	PGTDC60_0743	EG14_03615	Conserved hypothetical (DUF2149)		-
BrickBuilt_9	PGN_0558	PG1548	PGTDC60_0748	EG14_03640	Haem-binding protein	*hmuY*	-
	PGN_0559	PG1550	PGTDC60_0751	EG14_03655	Serine protease (Peptidase C10)	*prtT*	-
BrickBuilt_10	PGN_0632	PG0585	PGTDC60_1709	EG14_09225	Aspartyl-tRNA amidotransferase B		+
	PGN_0633	PG0587	PGTDC60_1713	EG14_09235	Membrane protein putative ion channel	*btuF*	-
BrickBuilt_11	PGN_0667	PG0625	PGTDC60_1753	EG14_09060	GTP cyclohydrolase I / PorSS CTD	*folE*	+
	PGN_0668	PG0627	PGTDC60_1756	EG14_09045	RNA-binding protein / PorSS CTD		-
BrickBuilt_12	PGN_0819	PG0796	PGTDC60_1912	EG14_08255	Leucyl-tRNA synthetase	*leuS*	-
	PGN_0823	PG0800	PGTDC60_1917	EG14_08240	NAD-utilizing dehydrogenase		-
BrickBuilt_13	PGN_0831	PG0807	PGTDC60_1926	EG14_08205	N utilization substance / PorSS CTD		-
	PGN_0832	PG0809	PGTDC60_1927	EG14_08200	Gliding motility protein / PorSS CTD	*sprA*	-
BrickBuilt_14	PGN_0871	PG1389	PGTDC60_2074	EG14_07995	Membrane protein		-
	PGN_0872	PG1391	PGTDC60_2073	EG14_08000	DNA-binding protein (PF00216)		+
BrickBuilt_15	PGN_0898	PG1424	PGTDC60_2039	EG14_07870	Peptidylarginine deiminase / PorSS CTD	*PAD*	-
	PGN_0900	PG1427	PGTDC60_2036	EG14_07865	Peptidase C10 / PorSS CTD		-
BrickBuilt_16	PGN_1207	PG1117	PGTDC60_1098	EG14_06320	Transport multidrug efflux		+
	PGN_1208	PG1118	PGTDC60_1096	EG14_06310	ClpB chaperone and protease	*clpB*	-
BrickBuilt_17	PGN_1476	PG0494	PGTDC60_1611	EG14_09640	PorSS C-terminal sorting domain		-
	PGN_1479	PG0491	PGTDC60_1606	EG14_09650	Peptidase S10 / PorSS CTD	*dppVII*	-
BrickBuilt_18	PGN_1777	PG1784	PGTDC60_0106	EG14_00235	Cysteine protease (Peptidase C1)		-
	PGN_1780	PG1786	PGTDC60_0110	EG14_00245	Endoribonuclease L-PSP		-
BrickBuilt_19	PGN_2035	PG0088	PGTDC60_0367	EG14_01925	Peptidase M16		+
	PGN_2037	PG0090	PGTDC60_0370	EG14_01930	DNA-binding protein from starved cells	*dps*	+

Table 3-2: Terminal Inverted Repeats (TIRs) of BrickBuilt elements from strain ATCC 33277. Terminal Inverted Repeats and family of selected IS and MITE-like elements in *Porphyromonas gingivalis* as well as ISNme1 from *Neisseria meningitidis*. Blank spaces for TIRs indicate situations in which the region is degenerate or not present.

Locus (33277)	**TIR 5' (nt)**	**TIR 3' (nt)**	
BrickBuilt_1	GAGACCTTTCGGCAA	TTGATGGAAGATGCT	
BrickBuilt_2		TTGCCGAAAGGTCTC	
BrickBuilt_3	GAGACCTTTCGGCAA	TCGTGCAAAGGTCT	
BrickBuilt_4	GAGACCTTTGCAAAA	TTGTGCAAAGGTCTC	
BrickBuilt_5	GAGACCTTTGCAAAA	TTGCGCAAAGGTCTC	
BrickBuilt_6		TTGCACAAAGGTCTT	
BrickBuilt_7		GGATACTATGGTCTC	
BrickBuilt_8		TTGTGCAAAAGTCTC	
BrickBuilt_9			
BrickBuilt_10			
BrickBuilt_11			
BrickBuilt_12	GAGACCTTTGTCCGA		
BrickBuilt_13			
BrickBuilt_14			
BrickBuilt_15	GAGCCCTTTGCAAAA	TTGCGCAAAGGTCTC	
BrickBuilt_16			
BrickBuilt_17	GAGACCTTTGCACAA	TTTTGCGAAGGGCTC	
BrickBuilt_18	GAGCCCTTTGCAAAA	TTGTGCAAAGGCCTC	
BrickBuilt_19	GAGCCCTTTGCAAAA		
Element	**Left TIR**	**Right TIR**	**Family**
ISPg1	GAGACCATTGCA	TTCAAAGGTCTC	IS5
ISPg3	ACGTCAGTTCGA	TCGAACTGACGT	IS5
ISPg4	GAGACTGTTGCA	CGCAACAGTCTC	IS5
ISPg9	GAGACCATTGCA		IS5
MITE239	ACGTGAGTTCGATATAAAGGAA	TTCGCTTAAATCGAACTGGCGT	
MITEPgRS/MITE464	GAGACTGTTGCA	TGCAACGGTCTC	
MITE700	ACGTCATTCGA	TCGAACTCACGT	
BrickBuilt	GAGACCTTTGCAAAA	TTGCGCAAAGGTCTC	
Nsm ISNme1	GAGACCTTTGCAAAA	TTTTGCAAAGGTCTC	IS5

Supplemental Table 3-1: Loci and nucleotide sites of BrickBuilt elements across four sequenced and annotated *P. gingivalis* strains. BrickBuilt elements are situated intergenically between the genes noted. Grayed-out boxes represent loci at which BrickBuilt is aberrant.

33277 Locus	Site (nt)		W83 Locus	Site (nt)		TDC60 Locus	Site (nt)		HG66 Locus	Site (nt)
PGN_0031	40105		PG0033	41680		PGTDC60_0032	38907		EG14_02395	533165
PGN_0033	40339		PG0034	41911		PGTDC60_0034	39090		EG14_02400	533444
PGN_0204	218332		PG2159	2266071		PGTDC60_1262	1303575		EG14_03005	669014
PGN_0205	218966		PG2161	2266741		PGTDC60_1265	1304210		EG14_03010	669671
PGN_0303	335482		PG0196	235960		PGTDC60_0466	510624		EG14_04795	1054743
PGN_0306	336162		PG0198	236646		PGTDC60_0471	511423		EG14_04805	1055427
PGN_0336	370117		NP	NP		PGTDC60_1661	1737582		EG14_04940	1089370
PGN_0340	370831		NP	NP		PGTDC60_1665	1738184		EG14_04960	1090090
PGN_0361	396602		PG0264	298363		PGTDC60_0543	573845		EG14_05065	1115859
PGN_0365	397602		PG0267	299179		PGTDC60_0547	574707		EG14_05075	1116882
PGN_0400	435731		PG1715	1802652		PGTDC60_0586	616440		EG14_05255	1154758
PGN_0403	436440		PG1714	1803366		PGTDC60_0590	617172		EG14_05265	1155398
PGN_0455	494054		PG0549	608294		PGTDC60_0639	672362		EG14_03135	694269
PGN_0456	494332		PG0553	608341		PGTDC60_0641	672604		EG14_03150	694678
PGN_0550	601270		PG1559	1637236		PGTDC60_0739	782604		EG14_03610	800540
PGN_0553	601821		PG1556	1637764		PGTDC60_0743	783086		EG14_03615	801091
PGN_0558	610873		PG1549	1628094		PGTDC60_0748	792138		EG14_03640	810144
PGN_0559	610969		PG1550	1628190		PGTDC60_0751	792234		EG14_03655	810240
PGN_0632	692403		PG0585	644119		PGTDC60_1709	1783831		EG14_09225	2047372
PGN_0633	692912		PG0587	644534		PGTDC60_1713	1784225		EG14_09235	2046886
PGN_0667	727570		PG0625	680646		PGTDC60_1753	1820657		EG14_09060	2010462
PGN_0668	728015		PG0627	680999		PGTDC60_1756	1820964		EG14_09045	2010884
PGN_0819	904861		PG0796	853989		PGTDC60_1912	1990956		EG14_08255	1830381
PGN_0823	905627		PG0800	854410		PGTDC60_1917	1991492		EG14_08240	1829799
PGN_0831	916153		PG0807	864955		PGTDC60_1926	2002085		EG14_08205	1818878
PGN_0832	916514		PG0809	865025		PGTDC60_1927	2002352		EG14_08200	1819282
PGN_0871	961376		PG1389	1471517		PGTDC60_2074	2150897		EG14_07995	1773937
PGN_0872	961459		PG1391	1471761		PGTDC60_2073	2150602		EG14_08000	1774020
PGN_0898	999323		PG1424	1511107		PGTDC60_2039	2112709		EG14_07870	1735309
PGN_0900	1000064		PG1427	1511688		PGTDC60_2036	2112060		EG14_07865	1736073
PGN_1207	1344456		PG1117	1191795		PGTDC60_1098	1136698		EG14_06320	1394744
PGN_1208	1344767		PG1118	1192037		PGTDC60_1096	1136249		EG14_06310	1394640
PGN_1476	1653510		PG0494	534679		PGTDC60_1611	1675602		EG14_09640	2153511
PGN_1479	1654229		PG0491	535398		PGTDC60_1606	1674859		EG14_09650	2154299
PGN_1777	1999152		PG1784	1873638		PGTDC60_0106	128421		EG14_00235	69154
PGN_1780	1999735		PG1786	1873670		PGTDC60_0110	129005		EG14_00245	70118
PGN_2035	2280789		PG0088	104802		PGTDC60_0367	389219		EG14_01925	418933
PGN_2037	2281319		PG0090	105331		PGTDC60_0370	389749		EG14_01930	419486

Supplemental Table 3-2: Primer and oligonucleotide sequences for PCR and cloning.

Primer or Oligo	Sequence
BrickBuilt_5 MITE forward	CTACTTGGATCCGCGCAGGCAATCTGTATCAA
BrickBuilt_5 MITE reverse	AAGTAGCCCGGGCGTTTAATCTAAATATACACAAAGGTAGACT
Leader oligo	GGATCCAAAGAGACCTTTGCAAAATCAAGAAGTATCTTTCCG GTATCGTTTTCGCCCATAGAGCGTGTCGCACAAGTCATAGAG GAGCTGAGGTATTCATATCTATAACGACACAGAAAAACAGTCA ATTGACAATGAAATCTCCCGGAAAATGCCGTGCGACAGTCTC ATCTATCGGTTTAACAGCTGGTACGGAACCTCCCACTCCCCG AATCTGCCATCCGGCAGAGAGCAGATAGCATCTTCCATCAAT CCCGGG
Tail oligo	GGATCCAAGTTCATTCTTTGTATCTCAGGGAAAAACCAATGAG CAGGTGTATGGCTGCCAAGTAAACAGAAAGCAGTGTCCGAAT ACGTCGCCAATCGATAAGCATATCGACTTGGTTCGAAAATACA TTCCGTGCTTTCCTGTATCGATTCGTTGACATGACGTCCCTGA ATGAGACCTCTTCGTCGGATGAATTTCCTTGTCGTGCCATCG TTTTGTGTCATTGATCATACCATAAACACACAAAACAATCGGC CGACACACTGTGAAATGAATGGATAAATGGATAAAGTGCCTG CTTATTGCGCACCCGGG

Chapter 4: Mapping and Characterization of Colony Pigmentation-associated Loci

4.1 Abstract

Pigmentation of *Porphyromonas gingivalis* plays an important role in multiple bacterial functions including but not limited to: colonization, survival and defense against bacteria-bacteria and bacteria-host interactions. Pigmentation-associated loci identified to date have involved lipopolysaccharide, fimbriae and haem uptake, acquisition and processing. We generated transposon mutant libraries in distinct *P. gingivalis* strains and screened for pigmentation-defective clones to identify genes involved in pigmentation and map a global pigmentation network. 235 sites (67 genes and 15 intergenic regions) were detected as pigmentation-related via Tn-seq analysis of strain ATCC 33277 background mutants; 7 of which were known and 75 of which are novel. PGN_0361 (PG0264), a putative glycosyltransferase located between two tRNA-synthetases and adjacent to miniature inverted-repeat transposable element, was identified in the Tn-seq screen and then verified through targeted deletion and complementation. Loss of the PGN_0361 glycosyltransferase abolishes pigmentation, changes hemolysis, modulates gingipain protease activity and alters lipopolysaccharide. The utilization of a saturated mutant library coupled with a high-throughput sequencing approach has allowed for the rapid identification of multiple genes involved in a virulence-associated phenotype. Given the macroscopic presentation of *P. gingivalis* pigmentation, knowledge of the global pigmentation 'network' will help allow for the development of inhibitors targeted at specific virulence attributes that effectively attenuate the keystone pathogen.

4.2 Background

Microbial pigments, ubiquitous throughout the microbial world, serve diverse functional roles in the lifecycle and physiology of their producers and affect interactions within the microbial community (Liu & Nizet, 2009). The roles of microbial pigment in virulence, protection against antimicrobial compounds and as components of toxins garner significant attention. In addition, both pathogenic and commensal microbes elaborate pigments that are integrally involved in nutrient acquisition, energy production and harvesting, as well as protection from ultraviolet light, heat, cold and oxidation (Liu & Nizet, 2009). Pigments elaborated by microbial species include yellow, orange, red, blue, green, brown and black. Expression of pigments *in vitro* can be constitutive or regulated based on medium nutrients and concentrations.

The black pigment produced by *Porphyromonas gingivalis* has been reported as a μ-oxo bis-haem dimer complexed to the outer surface of the cell (Smalley et al., 1998). *P. gingivalis* pigment was initially used as a diagnostic phenotype and is considered a major virulence factor of the species (Holt et al., 1999) (Nakayama, 2003). Parts of the *P. gingivalis* colony pigmentation biosynthetic and regulatory interaction network have been identified through targeted deletions and *Tn4400* or *Tn4351*-based transposon mutagenesis. From these investigations, thirty pigmentation-associated loci have been identified in lipopolysaccharide (LPS) biosynthesis and modification, fimbriae, specialized transport pathways and haem acquisition, uptake and processing functions. *P. gingivalis* LPS can be extensively modified. Four different lipid-A moieties and at least two distinct repeating oligosaccharides that attach to the core oligosaccharide, APS and OPS (also referred to as A-LPS and O-LPS), have been identified. *P. gingivalis* LPS is capable of being an agonist or antagonist for TLR2 and

TLR4, can cause alveolar bone resorption and is a important target for antimicrobial compounds. Change or loss of LPS structure may lead to different charges on the outer surface of the bacterial cell, or loss of a specific moiety, which in turn may result in loss of pigment. The fimbriae of *P. gingivalis*, both the major (*fimA*) and minor (*mfa1*) types, serve as attachment points to bridge bacteria-bacteria interactions as well as bacteria-host interactions (Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami, 2009). There are several fimbriae isotypes and expression varies on a strain-level basis, a different important enough that several studies have looked for correlations between fimbrial type and disease states. Fimbriae are purported to be anchored similarly as LPS, as such, a mutation may change cell surface structure and in turn the binding or entry sites of haem. Several haem acquisition, uptake and processing loci have been identified as being necessary for growth and virulence of *P. gingivalis*, which is thought to be dependent on exogenous sources of haem (Lewis, 2010). Limited haem acquisition or processing may lead to lack of sufficient haem to display on the cell surface. Additionally, modifications to haem or regulatory signals may be necessary for *P. gingivalis* to display haem on the cell surface, as the absence of red blood cells (RBCs) in growth media essentially abolishes colony pigmentation regardless of haem or protoporphyrin IX concentration.

Most known pigmentation-associated loci are connected through a singular point-- the attachment or complex of pigment to the outside of the cell. The actual attachment point as well as the full pathway of genesis and deposition of pigment have yet to be elucidated. As such, we chose to employ a recently-developed transposon mutagenesis system in *P. gingivalis* to further characterize the colony-pigmentation phenotype (Klein et al., 2012).

Revealing intermediate steps or convergent pathway points will aid in parsing out overlapping or pleiotropic phenotypes.

Determining the genes and mechanisms involved in pigmentation of *P. gingivalis* furthers understanding of a key biosynthetic and metabolic circuit, which can potentially be expanded to other black pigmenting *Bacteroidetes*, and may identify molecular targets for the keystone periodontopathic species.

4.3 Methods

4.3.1 Bacterial Strains, Media and Growth Conditions

Bacterial strains and PCR primers used in the study are listed in Table 1. *P. gingivalis* strains W83 and ATCC 33277 were grown on blood agar plates (BAPHK) containing trypticase soy agar supplemented with defibrinated sheep's blood (5% vol/vol), hemin (5 μg/ml), and menadione (0.5 μg/ml) as well as brain-heart infusion broth (BHIHK$S_{bc}S_{tg}$C) containing brain-heart infusion, yeast extract (1 mg/ml), hemin (5 μg/ml), and menadione (0.5 μg/ml), sodium bicarbonate (1 mg/ml), sodium thioglycolate (0.25 mg/ml), and cysteine (0.5 mg/ml) were used for solid and liquid culture of *P. gingivalis,* respectively. Gentamicin (25–50 μg/ml), erythromycin (2–10 μg/ml) and tetracycline (1-2 μg/ml) were used when appropriate for prevention of contamination as well as isolation and maintenance of *P. gingivalis* mutants. The strains were grown at 37°C in GasPak™ EZ Anaerobe Pouch System (BD Biosciences) for 24 hr or 48 hr for broth-based assays, and 7, 14 or 21 days for plate-based assays. Growth curves were performed using a Biotek Powerwave HT spectrophotometer fitted within an anaerobic chamber (COY). OD600 nm reads were taken every 15 min, following a 1min 'moderate' shake, over a 48 hr period.

Escherichia coli DH5α, TOP10 and S17-1 λpir were used for cloning, plasmid maintenance and conjugation. Ampicillin (100 μg/ml) was used when appropriate for prevention of contamination as well as isolation and maintenance of transformants containing plasmids. Strains were grown and maintained on LB [Lennox] agar or in LB [Lennox] broth (Invitrogen).

Table 4-1. Bacterial strains, plasmids, and primers for PCR, deletion and complementation.

Bacterial Strain / Plasmid / Primer	**Notes**
Porphyromonas gingivalis:	
ATCC 33277	Wild-type
W83	Wild-type
Tn-PGN_0361	ATCC 33277-background glycosyltransferase transposon insertion
Tn-PG0264	W83-background glycosyltransferase transposon insertion
Δ-PGN_0361	ATCC 33277-background glycosyltransferase deletion
Δ-PG0264	W83-background glycosyltransferase deletion
Tn-PGN_1302	ATCC 33277-background *waaL* transposon mutant
Tn-PG1051	W83-background *waaL* transposon mutant
Tn-1 Comp	Complemented W83-background PG0264 transposon mutant
Δ-1 Comp	Complemented W83-background PG0264 deletion mutant
Escherichia coli:	
TOP10	Cloning and plasmid maintenance
S17-1 λpir	Conjugation into *P. gingivalis*
pT-COW	Complementation vector without insert
pT-COW::PGN_0361	Complementation vector with ATCC 33277 PGN_0361 gene insert

4.3.2 Construction of P. gingivalis Transposon Mutant Libraries

Transposon mutant libraries were constructed in *P. gingivalis* strains ATCC 33277 and W83 using a Mariner-based pSAM_Bt system as described previously (Klein et al., 2012)(Klein, Duncan, & Hu, 2015). Briefly, wild-type *P. gingivalis* was cultured anaerobically on BAPHK and then used to inoculate BHIHK$S_{bc}S_{tg}$C for on overnight culture at 37ºC anaerobically in a Gaspak (BD). Overnight cultures of *P. gingivalis* were back-diluted into BHIHK$S_{bc}S_{tg}$C and incubated for 8 hr [to reach OD_{600} nm between 0.5-1.0]. While back-diluted *P. gingivalis* was growing, *E. coli* S17-1 λpir carrying vector pSAM_Bt was inoculated into LB broth containing Amp 100 μg/ml [to reach OD_{600} nm between 0.5-1.0]. *P. gingivalis* and *E. coli* were centrifuged to pellet, resuspended (together) in 1 ml of PBS, then 'puddled' on BAPHK lacking antibiotics and allowed to conjugate for 5 hr aerobically at 37ºC. Conjugation plates were resuspended in PBS and then plated onto BAPHK selection plates containing gentamicin and erythromycin. Mutant clones were pooled into BHI containing 15% glycerol following 10-14 days of growth and then frozen at -80ºC.

4.3.3 Construction of P. gingivalis Deletion and Complementation Mutants

Deletions of gene PGN_0361 in strain ATCC 33277 and PG0264 in strain W83 were carried out using 3-way stitching and homologous recombination. PCR primers (containing complementary overlaps to *ermG* of vector pSAM_Bt) were designed immediately flanking the gene PGN_0361 in *P. gingivalis* strain ATCC 33277 to generate 500 bp products upstream and downstream of the CDS (Table 4-1). ErmG PCR product (containing complementary overlaps to PGN_0361) was designed using vector pSAM_Bt *ermG*. PCR

products were generated using GoTaqLong Master Mix (Promega). An initial PCR for each three individual parts was performed, followed by a 'stitching' PCR for all three pieces together. Following PCR amplification and purification (Qiagen), the final stitched product was electroporated into electrocompetent wild-type ATCC 33277 and W83 cells. After overnight incubation at 37ºC anaerobically in a Gaspak (BD), cells were plated onto BAPHK containing gentamicin and erythromycin, and then incubated at 37ºC anaerobically in a Gaspak (BD). Transformants were confirmed via PCR and sequencing.

To generate complementation constructs for transposon mutants and deletions of PGN_0361 and PG0264, gene PGN_0361 was ligated into the replicating plasmid pT-COW at SphI and NheI restriction sites and then transformed into *E. coli* TOP10. Use of these restriction sites in pT-COW removes 335 bp from the middle of the *tet* (for *E. coli*) gene. Transformants were selected for ampicillin resistance generated by an insertion event. Clones were confirmed by restriction digest, PCR and sequencing. The complementation vector pT-COW::S0361 was then subcloned into *E. coli* S17-1 λpir. Deletion and transposon mutant strains in the *P. gingivalis* W83-background were conjugated with *E. coli* S17-1 λpir / pT-COW::S0361 without antibiotic selection (as above for transposon mutagenesis), and transformants were selected on BAPHK containing gentamicin and tetracycline following incubation at 37ºC anaerobically in a Gaspak (BD). Complementation was confirmed by growth on tetracycline, black colony pigmentation and sequencing.

4.3.4 Bioinformatic Analyses

Genome sequence FASTA and GenBank files were downloaded from the NCBI database. The Pfam and InterProScan databases and programs were used to determine the presence and characteristics of nucleic acid and protein motifs (Finn et al., 2014) (Jones et al., 2014). Query inputs were FASTA sequences from NCBI download files. For Pfam, an E-value of 1.0 and checking Pfam-B motifs were selected options prior to submission.

NCBI BLAST suites were utilized to find homologues and determine locations within *P. gingivalis* strains and in other species (Corpet, 1988). Query inputs were FASTA sequences from NCBI genome sequencing projects. Default settings for BLASTn and BLASTp were used, as well as BLASTn of WGS data.

PHYRE2 software platform was used to determine putative structure, structured domains and structure changes due to SNPs or truncations (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015).

The Geneious software platform (version R8) was used to generate sequence alignments (Kearse et al., 2012).

BROP, specifically the 'Genomics Tools for Oral Pathogens' and 'Microbial Transcriptome Database' sections of the resource, were used check for a determine levels of transcription for given pigmentation-associated *P. gingivalis* loci (Chen et al., 2005). Under the 'Genomics Tools for Oral Pathogens' subset, the 'GenomeViewer' function was used to compare genome areas under the three different nutrient conditions (the same conditions performed in the RNAseq) (Chen et al., 2004) (Høvik, Wen-Han, Olsen, & Chen, 2012).

4.3.5 Phenotypic Analyses

A colony pigmentation screen was performed using aliquots of the initial BAPHK-based pooled transposon mutant library. A single glycerol-stocked aliquot was thawed, diluted in BHIHK$S_{bc}S_{tg}$C, and then plated onto BAPHK to obtain single-fold coverage of the colonies contained in the library (~35,000). The screen was completed four times. Plates were incubated for 7-14 days, with clones selected throughout the incubation to acquire a variety of pigmentation defects ranging from complete, to partial, to delayed. All selected clones were individual stocked and then confirmed as pigmentation-defective following another plating. Patches were made of each selected clone for Tn-seq prep; each patch was divided into two and pooled in duplicate for gDNA preparation. Tn-seq was performed as described previously (Klein et al., 2012)(Klein et al., 2015).

Gingipain protease activity, both Rgp and Kgp, were measured using cleavable colorimetric substrates; Benzoyl-arginine-4 nitroanilide hydrochloride (BAPNA) for Rgp and *N*-α-benzyloxycarbonyl-l-lysine-*p*-nitroanilide (Z-Lys-*p*NA) for Kgp. Briefly, 20 µl of culture or culture supernatant from 24 hr or 48 hr BHI broth growth was added to a microtiter plate. 50 µl TCD solution as well as 50 µl of BAPNA solution were added to the microtiter plate. Assays were incubated immediately aerobically at 37˚C and monitored for yellow color development. Absorbance at 410 nm was measured and PBS-only negative control was subtracted from all samples (Potempa, Banbula, & Travis, 2000).

Gingipain protease-specific western blots were performed on cell extract from wild-type, transposon mutant, deletion and complementation *P. gingivalis* strains. SDS-PAGE was carried out in 4-15% mini-Protean gels using a BioRad mini-gel system. Blots were probed with antibody to r-RgpA adhesin domains, which has cross-reactivity with peptides from

RgpA, Kgp, and HagA (as well as higher molecular weight intermediates) (Chen & Duncan, 2004).

LPS preparations were generated two ways and results were equivalent between the two methods. Initially, cells were resuspended corresponding to 3 OD_{600} nm units of an overnight culture of *E. coli* (as control) or *P.gingivalis*, respectively, in 150 µl lysis buffer (2% SDS, 4% β-mercaptoethanol, 10% glycerol, 1M Tris HCl pH 6.8). Preparations were heated at 100°C for 10 min, and then incubated with 2 µl Proteinase K (20 mg/ml, Roche Applied Science) at 60°C for 2 hr. 150 µl 95% phenol was added and then incubated at 70°C for 15 min. Preparations were then cooled on ice for 10min and then centrifuged at 18,000 g for 10 min. The aqueous phase was then transferred to sterile tube. Finally, LPS was precipitated by the addition of 2.5 volumes of ethanol and then resuspended in 50 µl dH_2O. Subsequent confirmatory experiments using LPS isolations were prepared using the Intron LPS extraction kit (Intron Biotechnology, South Korea).

LPS silver, glycan and A-LPS stains and blots were carried out on purified LPS from wild-type, transposon mutant, deletion and complementation *P. gingivalis* strains. SDS-PAGE was carried out in 4-15% mini-Protean gels using a BioRad mini-gel system. Silver staining of gels was performed using the SilverQuest Silver Staining kit (Invitrogen/Life Technologies) according to manufacturer's instructions. Glycoprotein identification staining was performed using the Pierce glycoprotein staining kit (Invitrogen/Life Technologies) according to manufacturer's instructions. For A-LPS Western blotting, Immobilon-P nitrocellulose membranes were probed with monoclonal antibody MAb 1B5 (Paramonov, Aduse-Opoku, Hashim, Rangarajan, & Curtis, 2009).

Polymyxin B susceptibility of wild-type, transposon mutant, deletion and complementation *P. gingivalis* strains was measured via plate-to-broth growth assay. Strains were grown on BAPHK for 7-14 days at 37˚C in Gaspaks (BD). Single colonies were taken, resuspended in PBS, and used to inoculate BHI broth with increasing amounts of polymyxin B. Cultures were incubated anaerobically for 48 hrs at 37˚C. Growth was measured spectrophotometerically by absorbance at 600 nm.

4.4 Results

4.4.1 Transposon Mutant Library Colony Pigmentation Screen and Tn-seq

Preliminary colony pigmentation screens were performed with the ATCC 33277 strain and W83 strain background mutant libraries to optimize the selection timeframe and demonstrate that multiple insertions were responsible for pigmentation defects prior to a large screen. Clones selected in the preliminary screens had the transposon insertion sites determined by nested semi-random sequencing as previous described (Klein et al., 2012)(Klein et al., 2015). Known genes were hit and few identical insertion sites between clones were found (data not shown), suggesting that a large screen was appropriate. Identical sites were identified in mutants of both strain backgrounds, suggesting that secondary mutations were not the cause of pigmentation defects (data not shown). Growth and pigmentation rates of mutants as colonies on BAPHK suggested a range of 7-21 days as ideal for screening.

Four separate screens were performed, each using a single aliquot of the entire pooled *P. gingivalis* ATCC 33277 mutant library. Clones were selected between days 7 and 14 of growth, with plates only being exposed to ambient oxygen for an hour at maximum. Each

clone was re-struck, confirmed as having a pigmentation defect and then glycerol stocked such that each mutant could be prepared for Tn-seq in equal amounts.

Sequencing identified a total of 235 insertions that caused colony pigmentation defects (Tables 4-2 and 4-4). Of these 235 insertions, 60 are completely identical sequence reads. These identical hits correspond to regions within *kgp*, *hagA*, *rgpA*, 16S rRNA, 23S rRNA, CRISPR 30-36, *tonB*, *lysA* and *ISPg1*. Genes *kgp*, *hagA* and *rgpA* have highly-repetitive haemagglutinin domains and have each been previously shown to be involved in colony pigmentation. Genes *tonB* and *lysA* are each encoded twice within the genome at different sites. TonB is involved in the uptake of iron, haem, vitamin B12 and siderophores; providing energy necessary for potentially each of the three known haem uptake systems in *P. gingivalis*. LysA, the meso-diaminopimelate decarboxylase, is involved in peptidoglycan biosynthesis, thus potentially affecting the outer membrane components that garner and complex haem to the outer surface of the colony. Mutations within 16S rRNA and 23S rRNA loci are likely to affect cell growth rate in general, and since pigmentation seems to be tied slightly to colony growth these mutants may pigment more slowly than others within the pool. There are 31 full (intact) versions of *ISPg1* in strain ATCC 33277. The IS elements are transcribed and may have outward-reading promoters that affect upstream or downstream transcription. As such mutations within these regions are likely to cause pigmentation defects when associated with other protein coding sequences in the pigmentation network. The insertions into the CRISPR 30-36 region are located within the promoter region of an *ISPg1* element in addition to being within the CRISPR 30-36 coding region. As such, their effect on pigmentation may be indirect. Downstream of the *ISPg1* element is a MITE and phosphoglucomutase (*pgm*). Phosphoglucomutase is involved in carbohydrate metabolism

and mutations in *pgm* are pleiotropic. Additionally, *pgm* has been shown to be involved in LPS biosynthesis and modulation in *Yersinia pestis*, a connection that would potentially result in pigmentation changes in *P. gingivalis*.

Out of the total 235 insertions, 121 are within *kgp*, *hagA* and *rgpA*, 107 of which are unique. *Kgp*, *hagA* and *rgpA*, as well as the other gingipains and haemagglutinins comprise the largest protein coding sequences in the genome and are several of the most highly-saturated with transposon insertions in the mutant library. As such, insertions in these genes should be expected to comprise the majority of mutants selected in the screen. Even though these genes have been previously confirmed or suggested to be involved in colony pigmentation, the coverage of this Tn-seq screen may allow for identification of specific regions within the proteins required for the phenotype and also serves as a confirmation of the screen.

Genes of interest due to either gene function or number of insertions found in the screen include PGN_0361 (glycosyltransferase), PGN_0637 (heat shock-related protease *htrA*), PGN_1770 / PGN_0123 / PGN_2080 (PorSS C-terminal domain), and PGN_1116 and PGN_1234 (aminotransferase class I/classII) (Tables 4-2 and 4-4). The putative glycosyltransferase (PGN_0361) was identified twice [unique insertions] in the Tn-seq screen and also found independently in screens of the W83 strain background mutant library (data not shown) [the W83 putative glycosyltransferase locus is PG0264, and further references to this gene in either strain will be written as PGN_0361/PG0264 unless noted]. Other glycosyltransferases have previously been found to effect pigmentation through gingipain protease and LPS modifications (Yamaguchi et al., 2010)(Osbourne et al., 2010) (Paramonov et al., 2009). The heat shock-related protease *htrA* (PGN_0637) was identified

three independent times in the screen. HtrA and its homologues have been shown to affect protein maturation as well as outer membrane morphology. Importantly, another gene had previously been suggested to be the homologue of *E. coli htrA* and was shown not to affect pigmentation in *P. gingivalis* (PGN_1436). Two of the three Por Secretion System C-terminal domain-containing genes (PGN_1770, PGN_0123 and PGN_2080) identified in the screen are adjacent to known pigment or haem affecting loci. The Por Secretion System is a *Bacteroides*-specific secretion system that in *P. gingivalis* is responsible for gingipain protease secretion, and interruption of most PorSS genes leads to pigmentation defects (Sato, 2011). Since these PorSS C-terminal domain genes are located downstream of the previously identified pigment affecting genes, mechanisms involving protein transport or localization are plausible. The two aminotransferases (PGN_1116 and PGN_1234) are each located adjacent to known pigment affecting loci, a PorSS C-terminal domain protein and the PorR/PorS two component system, respectively. Given the gene orientations, polarity of insertions should not be a factor in the phenotype. Of note, aminotransferase domains are found within proteins that are part of or feed into tetrapyrrole biosynthesis pathways.

Table 4-2. *P. gingivalis* Colony Pigmentation-Associated Screen Tn-seq (Limited Detail). Genes and intergenic regions identified through Tn-seq analysis. The ATCC 33277 strain locus name, gene designation and protein functional characterization are provided. Protein functional characterizations were either found through NCBI or determined using Pfam. Gene designations were noted if NCBI listed or a previous article had characterized the gene in *P. gingivalis*.

Locus	Gene	Product
PGN_0026		Cytidine deaminase-like (IPR016193)
PGN_0048		PcfK-like protein
PGN_0100		Diaminopimelate decarboxylase LysA
PGN_0103	*tonB*	TonB
PGN_0123		PorSS C-terminal domain protein
PGN_0184	*fimD*	Minor component FimD
PGN_0215		Hypothetical protein
PGN_0287	*mfa1*	Mfa1 fimbrilin
PGN_0289	*mfa2*	Fimbrillin-A associated anchor protein Mfa2
PGN_0361		Glycosyl transferase family 2
PGN_0380	*nagC*	Transcriptional regulator/sugar kinase
PGN_0413	*gyrB*	DNA gyrase B subunit
PGN_0438		Hypothetical protein
PGN_0465	*relA*	GTP pyrophosphokinase
PGN_0504	*scpB*	Methylmalonyl-CoA decarboxylase beta subunit
PGN_0533	*nadA*	Quinolinate synthetase A (IPR003473)
PGN_0637	*htrA*	Heat shock-related protease HtrA protein
PGN_0715		Outer membrane efflux protein (IPR003423)
PGN_0778	*porT*	PorT
PGN_0789		TPR domain protein
PGN_0862		Type III restriction enzyme, res subunit
PGN_0903	*fimR*	Two-component system response regulator FimR
PGN_0946		Hypothetical protein
PGN_0949		ABC transporter ATP-binding protein
PGN_0950		ABC transporter ATP-binding protein
PGN_0952		Carboxyl-terminal processing protease
PGN_1105		FKBP-type peptidyl-prolyl cis-trans isomerase
PGN_1116		Aminotransferase, class I/classII (IPR004839)
PGN_1120		NADPH-NAD transhydrogenase/Alanine dehydrogenase/PNT
PGN_1234		Hypothetical protein
PGN_1272		Diaminopimelate decarboxylase LysA
PGN_1275	*tonB*	TonB
PGN_1282	*traN*	Conjugate transposon protein TraN

PGN_1284	*traP*	DNA primase involved in conjugation TraP
PGN_1302	*waaL*	O-antigen ligase
PGN_1303		Lipoprotein
PGN_1313		Hypothetical protein
PGN_1591		Hypothetical protein
PGN_1618		Methionine gamma-lyase
PGN_1643		PF07610 family protein / PapD-like (IPR008962)
PGN_1666	*purL*	Phosphoribosylformylglycinamidine synthase
PGN_1675	*porL*	Por secretion system protein PorL/GldL
PGN_1690		Alpha-L-fucosidase
PGN_1722		Phosphoribulokinase/uridine kinase (IPR006083)
PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp
PGN_1733	*hagA*	Haemagglutinin protein HagA
PGN_1770		PorSS C-terminal domain protein
PGN_1777		Peptidase C1B, bleomycin hydrolase (IPR004134)
PGN_1812	*ppk*	Polyphosphate kinase
PGN_1914		Carboxyl-terminal processing protease
PGN_1916		ABC transporter ATP-binding protein
PGN_1917		ABC transporter ATP-binding protein
PGN_1920		Membrane transport protein, MMPL domain (IPR004869)
PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA
PGN_1998		Outer membrane protein, OmpA/MotB, C-terminal (IPR006665)
PGN_2010		Secreted protein, YngK-like
PGN_2017		YjeF family protein
PGN_2047		Hypothetical protein
PGN_2080		Pectin lyase fold/virulence factor (IPR011050) / PorSS CTD
PGN_r0001		16S ribosomal RNA
PGN_r0002		23S ribosomal RNA
PGN_r0002		23S ribosomal RNA
PGN_r0004		16S ribosomal RNA
PGN_r0005		23S ribosomal RNA
PGN_r0008		23S ribosomal RNA
PGN_r0009		16S ribosomal RNA
PGN_r0010		16S ribosomal RNA
PGN_r0011		23S ribosomal RNA
Intergenic region		3' to *hagB*
Intergenic region		3' to PGN_0136
Intergenic region		3' to *murQ* and 5' to PGN_1195
Intergenic region		3' to PGN_0840 (*ispD*)
Intergenic region		3' to PGN_0326
Intergenic region		3' to PGN_1770

Intergenic region		5' to PGN_0329 (Transglutaminase)
Intergenic region		3' to PGN_1523 (Polysaccharide export protein)
Intergenic region		3' to PGN_1791 (Flavodoxin *fldA*)
Intergenic region		5' to PGN_0904 (*fimS*)
Intergenic region		3' to PGN_0174 (AraC transcriptional regulator)
Intergenic region		ISPg1
Intergenic region		rRNA-16S ribosomal RNA PGN16SrRNA09
Intergenic region		3' to PGN_0170 (L-asparaginase)
Intergenic region		5' to PGN_0889 (*tkrA*)
Intergenic region		CRISPR 30-36
Intergenic region		ISPg3 TIR

Several insertions that did not make our sequencing read count cutoff are repetitive, and when combined would aggregate as a hit. However, given that no individual insertion site had the reads required for a definition of a hit, they were not included in our final list.

From the screen we selected the putative glycosyltransferase PGN_0361/PG0264 for confirmation and further characterization. Five different insertions into PGN_0361 had been isolated as non-pigmenting strains, two in the large Tn-seq screen and three in prior optimization screens, as well as one strain from the W83 background. The insertion site within the W83 background isolate is identical to one of the ATCC 33277 background isolates, and these strains were used for all subsequent assays. The insertion is 174 nucleotides (58 amino acids) into the gene, which is a total of 1,068 nucleotides.

The PGN_0361/PG0264 gene is located immediately downstream of the tyrosyl-tRNA synthetase (*tyrS*). TyrS is essential, which we have previously shown, and presumably supplies a promoter for transcription of PGN_0361/PG0264 given the lack of space or a predictable promoter upstream of PGN_0361/PG0264. The first true protein coding sequence downstream of PGN_0361/PG0264 is the arginyl-tRNA synthetase (*argS*), another essential gene. However, between the 3' end of PGN_0361/PG0264 and the 5' end of *argS* is a non-autonomous transposable element that we identified during our characterization of Tn-seq data (manuscript submitted, see Chapter 3). As such, the PGN_0361/PG0264 gene is confined between two essential tRNA synthetases and adjacent to a repetitive non-autonomous transposable element.

To further confirm that PGN_0361/PG0264 is responsible for the non-pigmenting phenotype, we constructed targeted gene deletions and complementation *in trans*. Gene

deletions were constructed via insertion of the erythromycin resistance encoding *ermG* gene into the PGN_0361 and PG0264 genes by homologous recombination that removed the entire target gene. Multiple deletion clones were isolated that recapitulated the non-pigmenting phenotype (Fig.4-1). Complementation of the transposon insertion and the targeted deletion in the W83 strain background was carried out by cloning full length PGN_0361 into the replicating vector pT-COW and conjugating the resultant pT-COW::PGN_0361 back into *P. gingivalis* strains (Fig.4-2). Complemented strains recapitulated wild-type pigmentation when cultured on media containing tetracycline selective for the complement vector.

Upon removal of tetracycline selection and growth in BHI broth for 48 hr, and then inoculation onto BAPHK without tetracycline, about one-third of the population loses the pT-COW::PGN_0361 vector completely. Additionally, of the population that initially retains the plasmid a small percentage (>5%) form sectored colonies, while the remaining 60% of colonies begin by pigmenting from the center of the colony and gradually lose pigmentation by the edge of the colony.

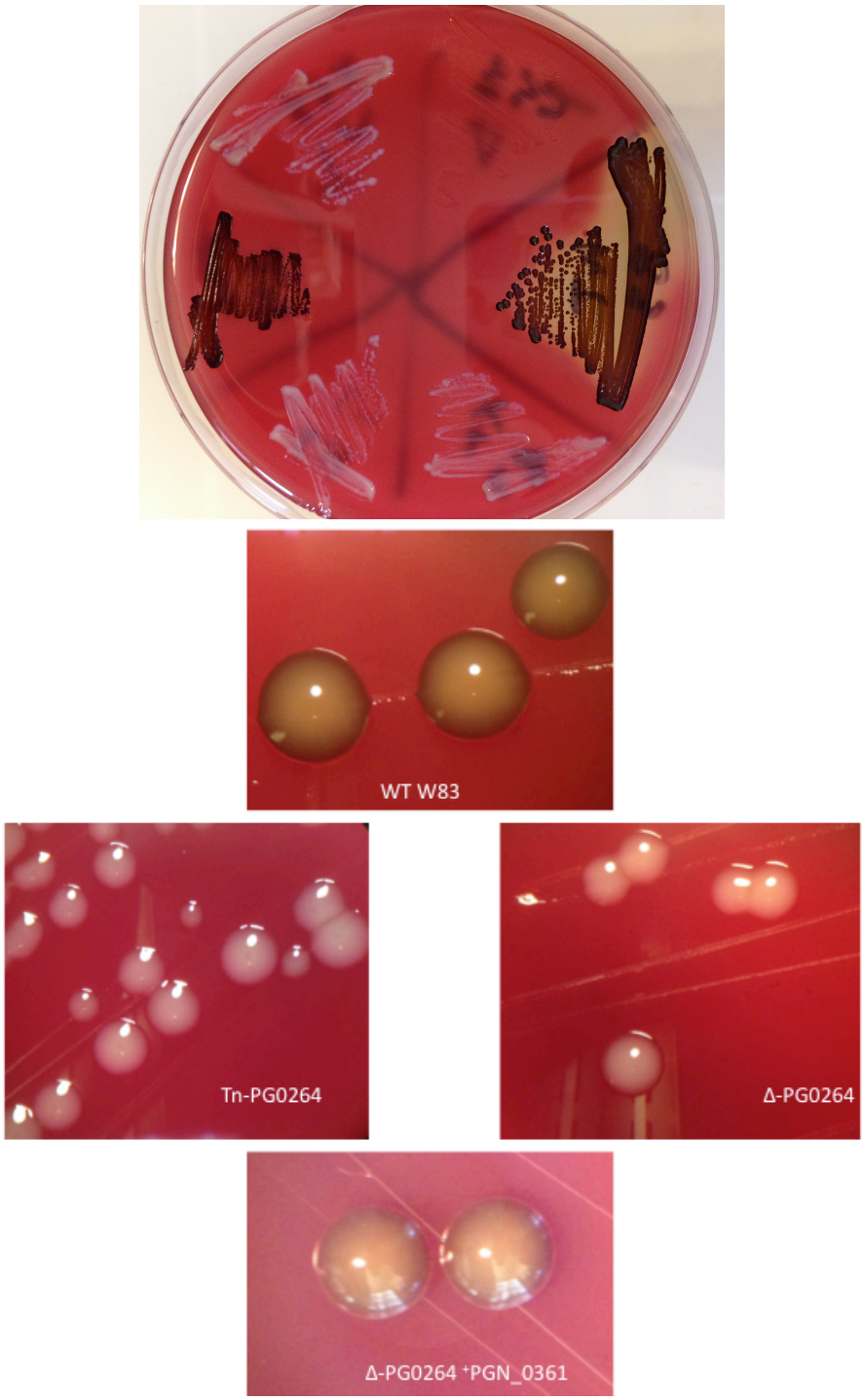


Figure 4-1. Pigmentation on Blood Agar of Wild-Type and Mutant Strains.
Top image shows wild-types, transposon mutants and deletion mutants in ATCC 33277 and W83 strain backgrounds; clockwise from left are wild-type ATCC 33277, transposon mutant in ATCC 33277, deletion mutant in ATCC 33277, wild-type W83, transposon mutant in W83 and deletion mutant in W83. The deletion mutant in the ATCC 33277 background displays more translucent colonies than other constructs. Bottom images show individual colonies of wild-type, insertional mutants and complemented insertion mutant from the W83 strain background following the same time of incubation.

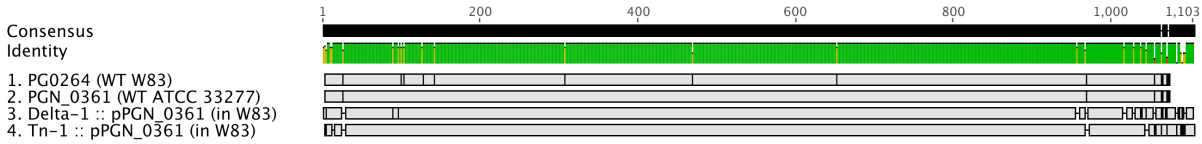


Figure 4-2. Wild-Type and Complemented Mutant Strains Sequence Alignment. Alignment of PGN_0361/PG0264 putative glycosyltransferase sequences of wild-types and complemented transposon and deletion mutants. Seven SNPs are present between wild-types, and these SNPs can be tracked in the complemented strains as the background for complements are W83 and the pT-COW::PGN_0361 contains the ATCC 33277 sequence. SNPs are seen as black lines within the grey alignment tracks. Generated using Geneious R8 platform.

4.4.2 PGN_0361/PG0264 Bioinformatic Analyses

After identification of PGN_0361/PG0264 as a target of interest, we chose to use *in silico* analyses to assist in characterization of the gene/protein. Pfam and InterProScan programs were used to identify the presence of known nucleic acid and protein motifs. Pfam identified a glycosyltransferase 2_3 domain that encompasses the N-terminal half of the protein (e-value $3.6x10^{-23}$); a motif that is widely distributed among prokaryotic and eukaryotic species. InterProScan confirmed the glycosyltransferase 2_3 domain, which was also recognized by GENE3D and PANTHER software. The C-terminal third of the protein contains no predicted domains.

The NCBI programs BLAST and conserved domain database (CDD) were queried to find homologues of PGN_0361/PG0264, the genomic landscape of PGN_0361/PG0264 homologues and potential active sites of the glycosyltransferase domain. Following BLASTn and BLASTp of PGN_0361/PG0264 (using default BLAST settings), all of the matches with e-values greater than 1x10-100 are within *Bacteroidetes*. *Porphyromonas gingivicanis* has maintained the tyrosyl-tRNA synthetase upstream of the PGN_0361/PG0264 homologue, but downstream has been rearranged and the protein homologue itself seems to be truncated with an early translational stop. Recombination and repair factor *recR* is the upstream gene in the majority of other *Porphyromonas*, *Parabacteroides, Haliscomenobacter, Sporocytophaga*, *Owenweeksia* and *Tannerella* species. *Fluviicola taffensis* homologue contains endoribonuclease L-PSP directly upstream; a gene that has a repetitive non-autonomous transposable element associated with it in *P. gingivalis*. Of note, the C-terminal region of the genes/proteins have less homology between strains and species than the N-terminal region that contains the glycosyltransferase 2_3 domain. The CDD queries revealed non-

overlapping predicted metal binding and transferase active sites, which were conserved throughout PGN_0361/PG0264 homologues. An 'unextendable partial coding region' is present in the *Barnesiella viscericola* DSM 18177 PGN_0361/PG0264 homologue, but not in *Barnesiella intestinihominis* YIT 11860. An 'unextendable partial coding region' is present in the *Porphyromonas crevioricanis* strain COT-253_OH1447 homologue, and is followed by a pseudogene involved in cell wall biosynthesis. In the instances of shortened sequences, the stop codon always occurs following in the C-terminal region after the glycosyltransferase domain. It is possible that these 'unextendable partial coding regions' may be due to sequeicng or assembly errors.

The PHYRE2 platform was queried in order to generate protein structure predictions as well as identify proteins that form similar structures to identify potential structure-function links. The best model template found was n-acetylgalactosaminyltransferase 2 from humans, which is involved in mucin type O-glycan biosynthesis. O-glycan biosynthesis involves the addition of carbohydrate moieties in a repeating fashion, similar to that of LPS biosynthesis in bacteria, and analogous O-glycosylation systems have been described in *Neisseria* and *Pseudomonas* species (Hug & Feldman, 2011). The carbohydrate active enzyme database (CAZY) glycosyltransferase type of human n-acetylgalactosaminyltransferase 2 is GT27, however, GT27 and GT2 are noted as having high similarities (when compared to all 97 GT types). A list of all *P. gingivalis* strain ATCC 33277 glycosyltransferases and their GT types are listed in Table 4-3.

Table 4-3. CAZY Identifications of Glycosyltransferases in *P. gingivalis* ATCC 33277. Each CAZY identified glycosyltransferase in *P. gingivalis* strain ATCC 33277 listed by locus, glycosyltransferase type (predicted or confirmed), gene designation if known, and whether the Mariner-based transposon mutant library harbored viable transposon mutant in the locus.

Locus	GT Type	Gene	Tn Library Viability
PGN_0225	GT2		
PGN_0232	GT2		
PGN_0361	GT2		
PGN_0777	GT2		No insertions in Tn library
PGN_1026	GT2	*wcaA-like*	No insertions in Tn library
PGN_1045	GT2	*lacZII*	
PGN_1239	GT2		
PGN_1628	GT2		
PGN_1651	GT2		No insertions in Tn library
PGN_1668	GT2		
PGN_1724	GT2		
PGN_1807	GT2	*wcaA-like*	
PGN_2087	GT2		No insertions in Tn library
PGN_1310	GT3	*glgA*	
PGN_0227	GT4	*rfaG-like*	
PGN_0242	GT4	*mtfB-like*	
PGN_0428	GT4		
PGN_1134	GT4	*rfaG-like*	No insertions in Tn library
PGN_1135	GT4	*rfaG-like*	
PGN_1240	GT4	*rfaG-like*	
PGN_1251	GT4	*gtfB*	
PGN_1736	GT5		No insertions in Tn library
PGN_1255	GT9	*rfaF-like*	
PGN_0206	GT19	*lpxB*	
PGN_0233	GT26		
PGN_0627	GT28	*murG*	No insertions in Tn library
PGN_0544	GT30	*waaA*	No insertions in Tn library
PGN_0733	GT35	*glgP*	
PGN_0817	GT51		
PGN_1627	GT83	*arnT*	

BROP 'Genomics Tools for Oral Pathogens' and 'Microbial Transcriptome Database' were used to visualize previous microarray and RNAseq data pertaining to *P. gingivalis* PG0264. Microarrays and matching RNAseq had been performed on blood, tryptic soy and minimal media (Chen et al., 2004) (Høvik et al., 2012). RNAseq data visualized through JBrowse software depicts lower transcription of PG0264 on blood medium as opposed to tryptic soy and minimal medium. Microarray data showed a significantly higher level of transcription on minimal as opposed to tryptic soy medium, with a p-value of 0.007 (Chen et al., 2004) (Høvik et al., 2012).

4.4.3 PGN_0361/PG0264 Phenotypic Characterization

To determine the effects and mechanism of action of the PGN_0361 we employed assays targeting haem metabolism, gingipain proteases activity and expression, and lipopolysaccharide visualization.

Four different types of growth experiments were carried out in order to rule out growth rate, final cell density or colony size as factors in the pigment phenotype. First, wild-type and mutant strains were grown anaerobically at 37°C for 48 hr in a chamber to observe growth rate. No significant defect in glycosyltransferase mutant growth with respect to wild-type under standard broth medium supplementation was apparent. Second, wild-type and mutant strains were grown anaerobically at 37°C for 72 hr in gaspaks. Cells from wild-type and PGN_0361/PG0264 mutants were visually identical with respect to cell density and turbidity at endpoint after static growth (Fig.4-3). However, the O-antigen ligase (*waaL*) mutant culture resulted in significant pelleting and a 'wall-climbing' phenotype that have previously been identified for *waaL* mutants in *P. gingivalis*. Third, selected Tn-seq data

from ongoing nutrient-variable experiments with the transposon mutant libraries was examined. Transposon mutants in PGN_0361 survived/grew best under hemin supplementation when compared to supplementation with either protoporphyrin IX or protoporphryin IX combined with vitamin B12 and serum (Fig.4-4). In the Tn-seq experiments both the number of total distinct insertions within the gene, as well as the number of reads from a given insertion, were greater under hemin supplementation than the other conditions. Lastly, wild-type W83, deletion of PG0264 and transposon mutant in *waaL* were grown on blood agar either with or without hemin supplementation. No colony morphology differences were apparent for any of the three strains due to exogenous hemin supplementation of blood agar (Fig.4-5).

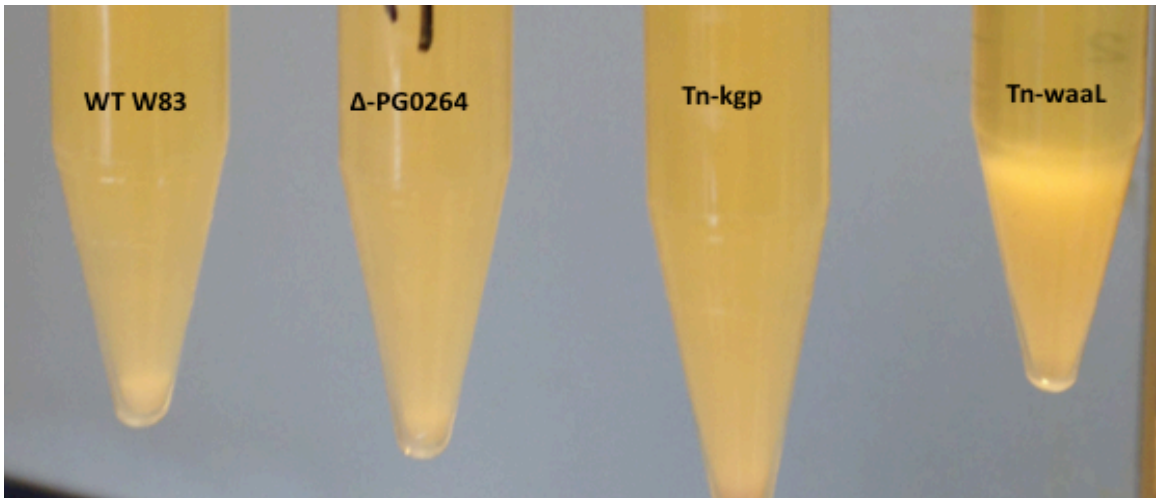


Figure 4-3. BHI Broth Growth of Wild-Type and Mutant Strains.
Wild-type, PG0264 deletion, transposon mutant in *kgp* gingipain and transposon mutant in *waaL* O-antigen ligase following 72 hr static BHI broth growth. The *waaL* transposon mutant aggregates at the bottom of the tube (volumetric number can be seen through less turbid culture). The *kgp* transposon mutant, which displays white colonies on blood agar due to inability to lyse RBCs and bind haem to the surface, as well as the PG0264 deletion do not aggregate, suggesting that the PG0264 mutation retains some form of O-antigen.

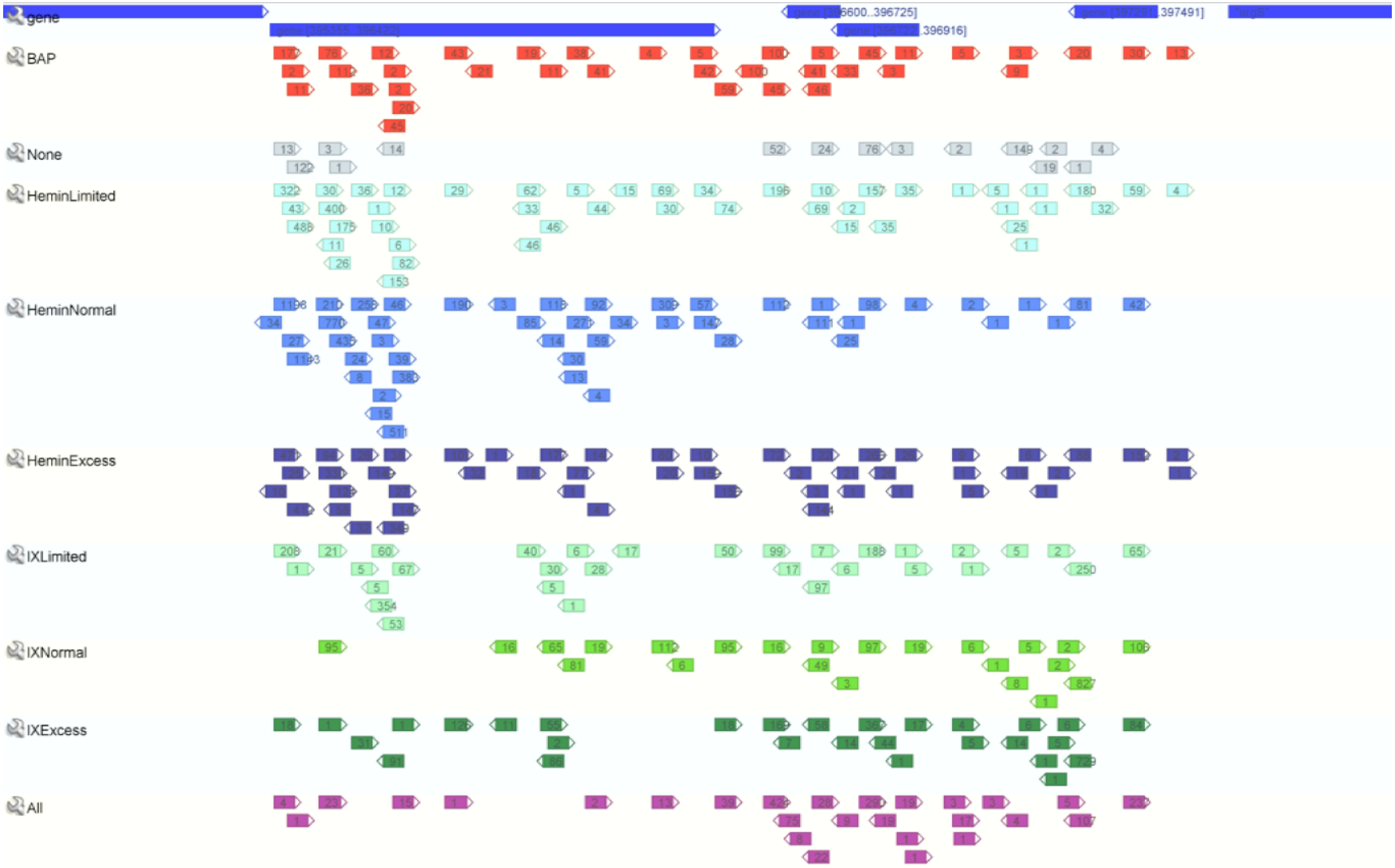

gene
BAP
None
HeminLimited
HeminNormal
HeminExcess
IXLimited
IXNormal
IXExcess
All

Figure 4-4. Tn-seq Under Exogenous Tetrapyrrole Supplementation. GenomeView-visualized Tn-seq results focused on PGN_0361-area. Selection conditions were varying levels of haem, protoporphyrin IX, vitamin B12 and serum. The top panel depicts the BAPHK-based original mutant library. The next four panels show the original mutant library after three passages in $BHIHKS_{bc}S_{tg}C$ with no added haem, $1/10^{th}$, 1 and 2 times the base haem level. The next three panels show the analogous experiments to haem except using protoporphyrin IX supplementation. The final panel shows supplementation with protoporphyrin IX, vitamin B12 and serum.

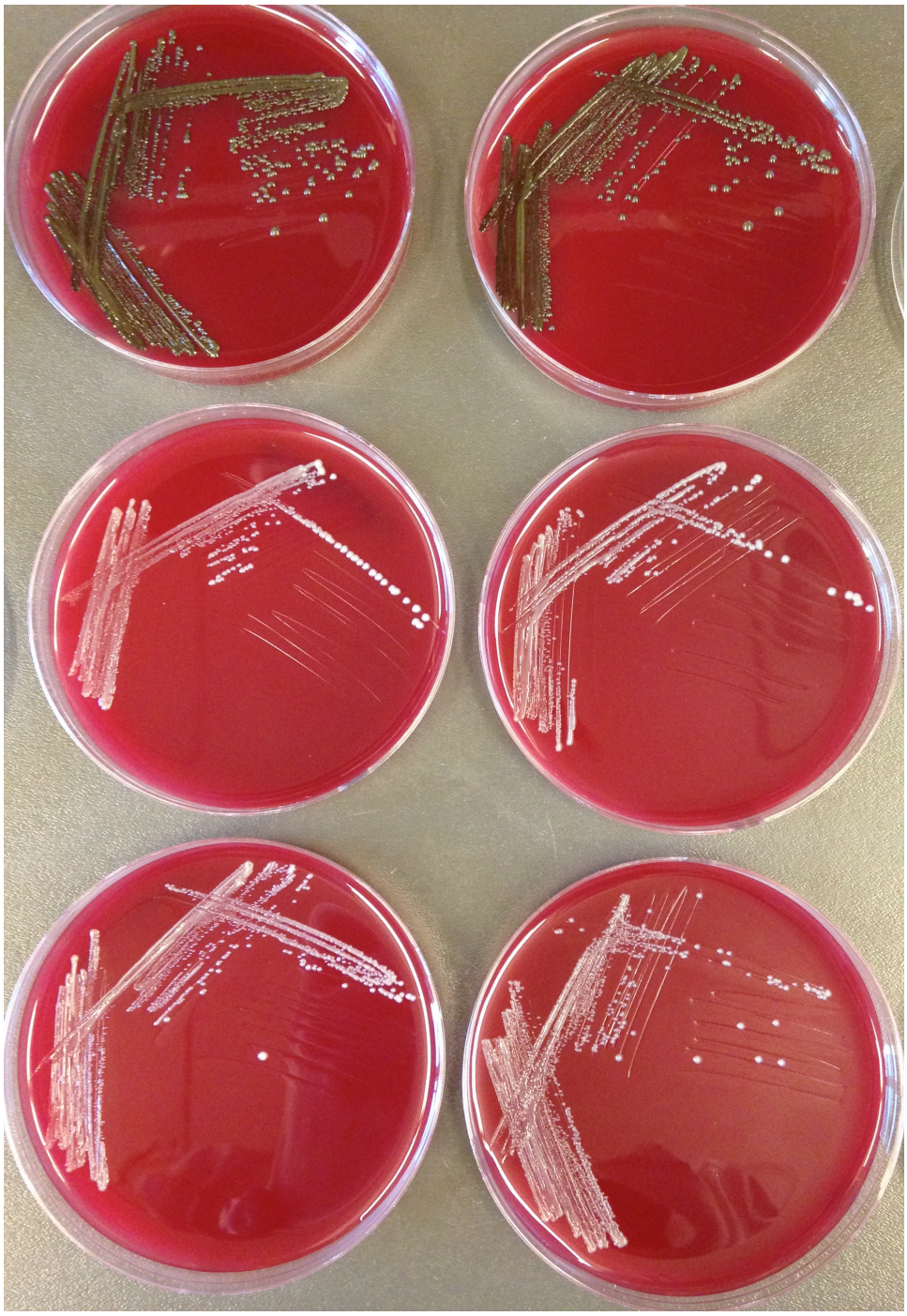


Figure 4-5. Pigmentation With Respect to Exogenous Haem of Wild-Type and Mutants. Wild-type (top), PG0264 glycosyltransferase deletion (middle) and W83-background transposon mutant in *waaL* O-antigen ligase (bottom) imaged after seven day growth on blood agar plates with (left) or without (right) exogenous haem supplementation. Colony morphology is unaffected by lack of exogenous haem in solid growth media as long as blood is present.

Gingipain protease expression and activity experiments were carried out. Expression of gingipain proteases was confirmed through commassie brilliant blue and silver stains. Reduced level of Rgp gingipain activity was found for transposon mutants and deletions of the putative glycosyltransferase in both wild-type backgrounds (Fig.4-6). The difference in cell-associated Rgp activity was greater than that of supernatant-associated activity. No difference in Kgp gingipain activity was apparent.

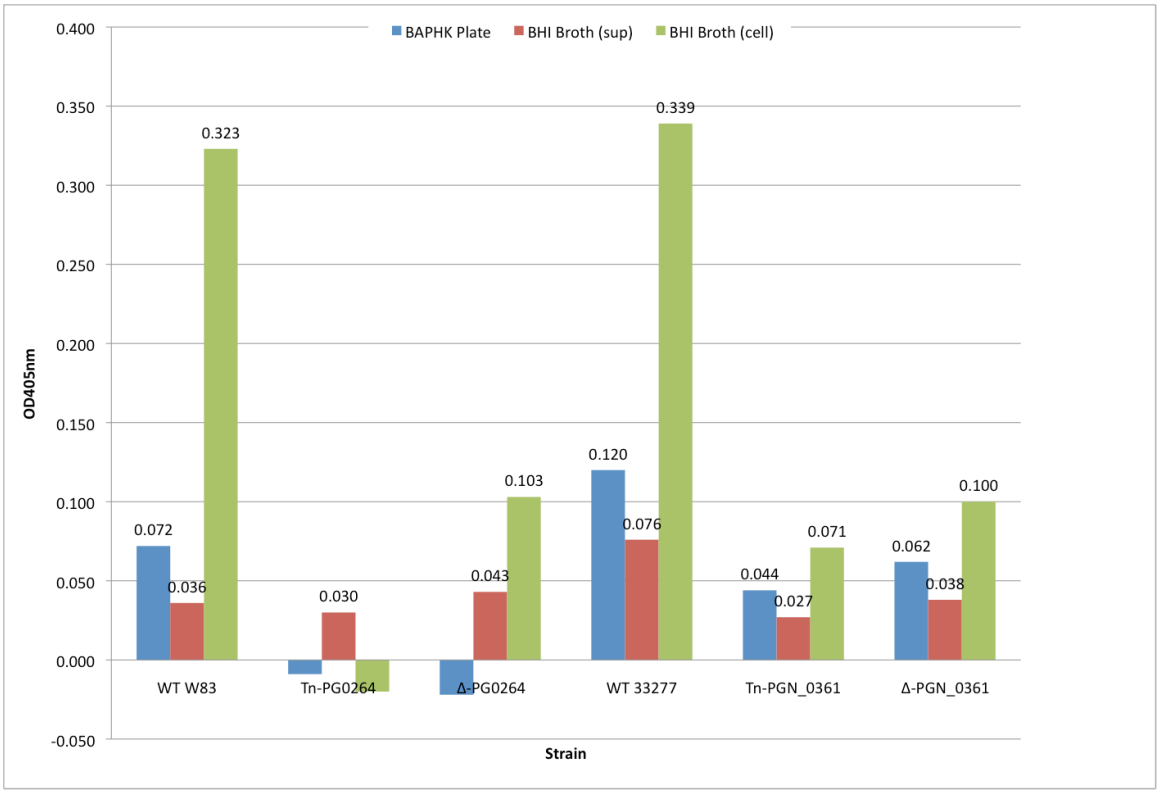


Figure 4-6. Gingipain Protease Activity.
Columns represent average of triplicates. Values reported are optical density measured at 405 nm and were normalized to cell density at OD600 nm. Samples contain colonies from BAPHK plates, supernatant from BHI broth and cells from BHI broth are indicated by blue, red and green columns, respectively. Reduction of cell-associated Rgp gingipain protease activity seen in PGN_0361/PG0264 mutants with respect to wild-type background.

Three different methods to analyze lipopolysaccharide were utilized. First, LPS preparations were examined via silver stain following SDS-PAGE. Variations in the banding pattern between wild-type and deletion mutant are apparent in both strain backgrounds (Fig.4-7A). Typical 'laddering' that is attributed to repeating O-antigen units was absent in the deletion strains. However, when examining the LPS via glycan stain following SDS-PAGE, an opposite laddering phenotype can be seen. The PGN_0361/PG0264 transposon insertion and deletion strains have laddering while it is absent in the wild-type strains (Fig.4-7B). The O-antigen ligase mutant lacks all ladder-type staining, which is expected because without the activity of O-antigen ligase only the lipid-A portion of LPS is expressed properly on the cell surface (Paramonov et al., 2009). Complementation of both the transposon mutant and the deletion mutant of PG0264 with the pT-COW plasmid expressing PGN_0361 shows the same LPS staining pattern as the wild-type (Fig.4-7C). Western blot performed with antibody 1B5 that is specific to A-LPS repeating unit of *P. gingivalis* shows a complete lack of A-LPS from both the transposon mutant and the deletion mutant of PG0264 (Fig.4-7D). Additionally, the *waaL* transposon mutant shows no A-LPS banding, which has been previously reported and serves as a control for the 1B5 staining. Complemented strain A-LPS were stained similarly by antibody 1B5 as wild-type.

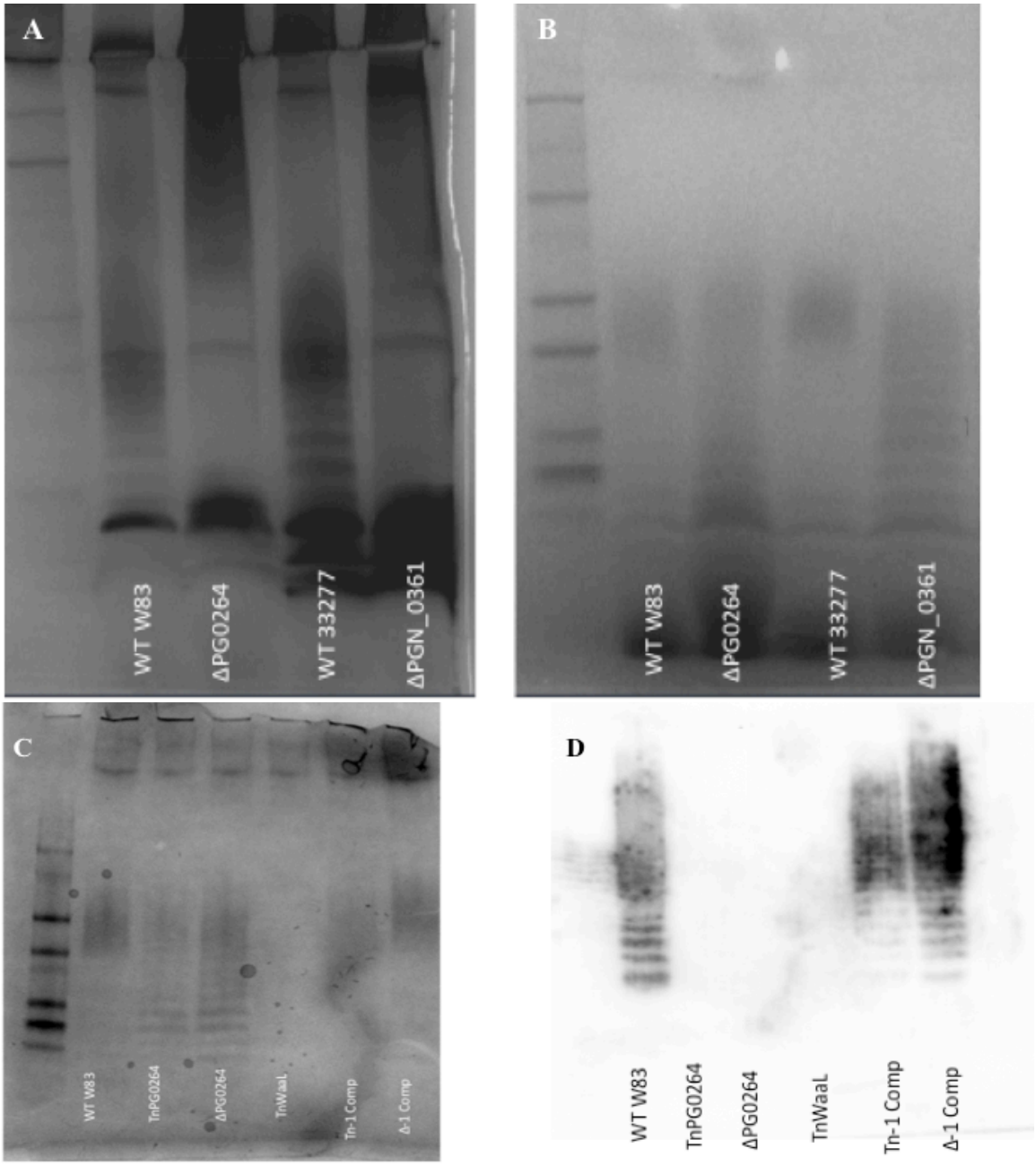


Figure 4-7 (A-D). Lipopolysaccharide Staining and A-LPS Western Blot.
Lipopolysaccharide preparations visualized by silver stain (A), glycan stain (B/C) and Western blot with monoclonal antibody 1B5 (D). Deletions of PGN_0361/PG0264 cause loss of 'laddering' in silver stain, yet deletion mutant display 'laddering' in glycan stain when wild-types do not. Lack of O-antigen ligase activity abolishes glycan staining and antibody 1B5 Western blot signal.

A polymyxin B antimicrobial susceptibility assay was carried out using wild type, PG0264 transposon, PG0264 deletion and O-antigen ligase mutants in the W83 strain background. Modifications of LPS have previously been demonstrated to modulate polymyxin B susceptibility in *P. gingivalis* and other species (Coats, To, Jain, Braham, & Darveau, 2009)(Díaz et al., 2015) (Tzeng et al., 2005). A trend of the PG0264 deletion and O-antigen ligase mutants showing reduced resistance, the PG0264 transposon mutant showing an intermediate phenotype and the complement of the PG0264 deletion showing increased resistance emerged. As the transposon insertion location within the mutant may not completely abolish function of the protein an intermediate phenotype could be expected. In support of this, following over 28 days of incubation on BAPHK, the transposon mutation strains in PGN_0361/PG0264 regain partial colony pigmentation, yet this is never observed in the targeted deletion mutants.

4.5 Discussion

Pigmentation has been categorized as a virulence factor of bacterial pathogens such as *Pseudomonas aeruginosa* and *Staphylococcus aureus*. While *P. gingivalis* black colony pigmentation is a similarly known virulence factor, unlike that of *P. aeruginosa* and *S. aureus* pigments *P. gingivalis* pigmentation is not generated through a well-defined and linear genetic route, nor is it transported by a universal system. Colony pigmentation of *P. gingivalis* involves a myriad of genes spread throughout the chromosome and a specialized transport system to deliver the proteins when and where they are required. To date, black colony pigmentation of *P. gingivalis* involves about thirty genes that affect LPS biosynthesis and transport, gingipain proteases glycosylation and transport, and fimbriae biosynthesis. Although many factors of colony pigmentation are known the complete network of genes and environmental stimuli required have yet to be determined.

With our transposon mutant library and sequencing strategy we aimed to identify new genes that are part of a global pigment network. The majority of our identified mutants from the Tn-seq screen were comprised of genes already known to be involved in colony pigmentation; *kgp*, *rgpA*, *hagA*, *mfa1* and *mfa2*. These serve to validate the screen. In addition, for genes where there are multiple insertions with different phenotypes, the exact insertion sites may aid in identification of specific regions within the proteins that are necessary for the pigment phenotype. Another group of hits from the screen are ribosomal RNA genes. Mutations in these regions can effect translation and growth rate in general, which may delay colony pigmentation. A third group of candidates identified in our screen involve mobile elements. Each of the eight genes from *TnPg17-A* and *TnPg17-B* were hit once, and none of the sequences were repetitive. The two mobile elements are distinct four

gene clusters located in separate parts of the genome comprised of two ABC transporter proteins, a carboxyl terminal processing protein and a membrane protein.

We chose to confirm and characterize the PGN_0361/PG0264 gene because multiple insertions were identified during the screen, the location of the gene on the chromosome and bioinformatic analyses suggested that previously noted pathways involving pigmentation may be involved.

A clue to the potential site of action of the putative glycosyltransferase came from Paramonov et al. noting in their characterization of *P. gingivalis* W50 O-antigen repeating subunit that loss of the phosphoethanolamine moiety would allow for periodate oxidation at that residue (Paramonov et al., 2001). Since the periodate oxidation seen in our glycan staining of LPS preparations matches the laddering property attributed to LPS O-antigen following silver staining, the prediction by Paramonov et al. may have helped to localize the site of action to the phosphoethanolamine attachment point. Whether the glycosyltransferase directly modifies O-antigen with phosphoethanolamine or if the O-antigen composition is modified such that phosphoethanolamine can no longer be added currently remains unknown.

A lack of A-LPS staining with the 1B5 antibody of PG0264 mutations suggests that potentially A-LPS is not produced by these strains, but more likely that the A-LPS is either no longer attached to cell surface or that it has been degraded. Because A-LPS and O-LPS are transported to the cell surface via the same system and some form of O-LPS is present, it is doubtful that the transport of A-LPS is inhibited in these strains. Additionally, since the broth growth phenotype of PG0264 mutations is not similar to *waaL* mutant, some form of O-LPS would be expected as in its absence there would be excessive aggregation and pelleting during liquid culture.

Glycan staining shows that PGN_0361/PG0264 phenotype is WaaL-dependent [O-antigen ligase] (PGN_1302/PG1051) and thus requires O-LPS or A-LPS to perform its function. Being O-antigen ligase-dependent does potentially compartmentalize where the PGN_0361/PG0264 modification takes place given that O-LPS and A-LPS are synthesized prior to export to the cell surface. If the O-LPS and A-LPS are the site of action, then WbaP (PGN_1896), Wzx (PGN_1033), Wzy (PGN_1242) and WzzP (PGN_2005) as well as LptA, LptC, LptD and LptO could be used to determine localization of PGN_0361/PG0264 action.

As with all screens several pitfalls are possible due to our methods. First, some genes involved in colony pigmentation may be missed due to essentiality or insertion preferences of the transposon vector. Second, experimenter bias during selection could have skewed the collection of clones. Lastly, more screen replicates could have been performed to collect additional colonies in order to saturate the screen.

4.6 Conclusions

In conclusion, we have utilized a Mariner transposon based mutagenesis system with *Porphyromonas gingivalis* in order to isolate novel colony pigmentation-affecting loci in the species. When screened *en masse* and coupled with high-throughput sequencing, we identified 235 sites within the mutant library, 67 genes and 15 intergenic regions, putatively involved in the colony pigmentation network. From these identified mutants, we confirmed and characterize a putative glycosyltransferase with a complete non-pigmenting phenotype. The putative glycosyltransferase PGN_0361/PG0264 affects pigmentation, changes hemolysis, modulates gingipain protease activity and alters lipopolysaccharide. The alteration in lipopolysaccharide is potentially located within the repeating oligosaccharide unit, likely involving a phosphoethanolamine moiety (Fig.4-8). Determining the saccharide

composition of the mutant strains as well as confirming the glycosyltransferase mechanism would help further define the site and mechanism of action for PGN_0361/PG0264.

The use of transposon mutant libraries in *P. gingivalis*, pSAM_Bt-based as well as other vector constructs, will allow for positive and negative selection of mutants in order to identify interaction networks and better understand virulence factors such as pigmentation. Several pressing questions that could potentially be parsed out by the other hits from the screen are if and how haem is processed within the cytoplasm for re-display on the bacterial outer-membrane and how heme that has been displayed on the outer-membrane can be re-incorporated into the cell during times of heme-limitation.

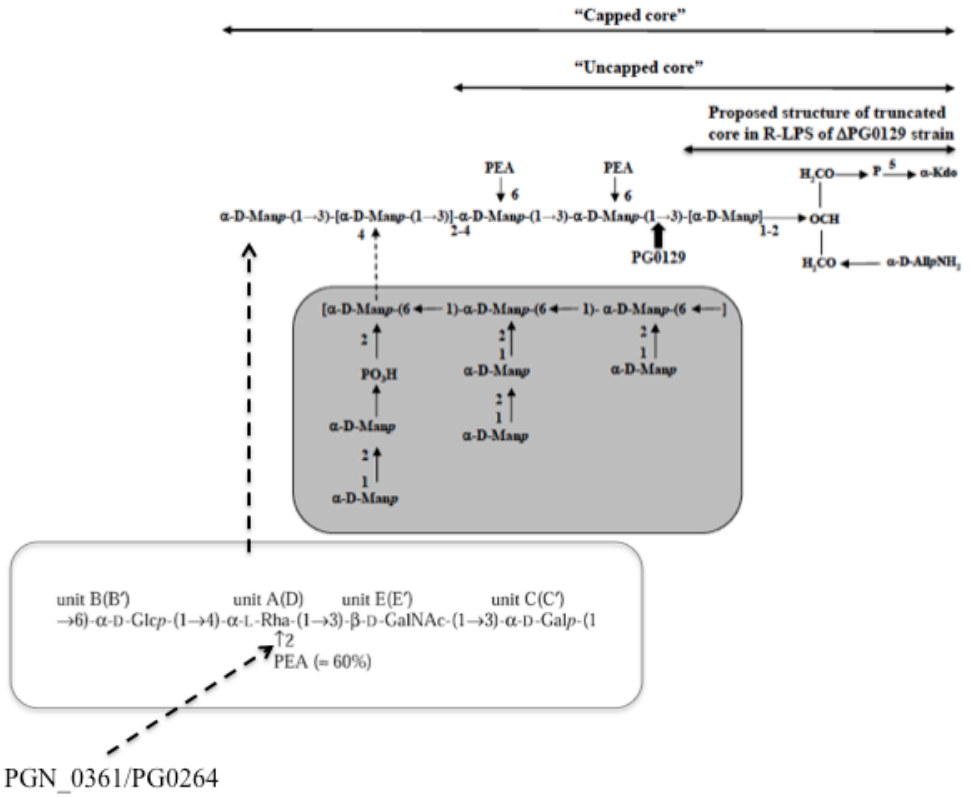


Figure 4-8. Model of Potential Site of Action for Glycosyltransferase PG0264.
Model modified from Paramonov et al., Eur. J. Biochem. 268, 4698-4707 (2001) and Paramonov et al., J. Bacteriol. doi:10.1128/JB.02562-14 (2015) (Paramonov et al., 2001) (Paramonov, Aduse-Opoku, Hashim, Rangarajan, & Curtis, 2015). Structures shown are the core oligosaccharide from *P. gingivalis* strain W50 LPS with the O-antigen repeating unit and A-LPS repeating unit, boxed with white and grey background respectively. The dotted arrow pointing to the O-antigen phosphoethanolamine residue denotes most likely site of action for glycosyltransferase PG0264 as determined from our initial characterizations.

Table 4-4. *P. gingivalis* Colony Pigmentation-Associated Screen Tn-seq (Expanded Information)

Genes and intergenic regions identified through Tn-seq analysis. The ATCC 33277 strain locus name, gene designation and protein functional characterization are provided. Protein functional characterizations were either found through NCBI or determined using Pfam. Gene designations were noted if NCBI listed or a previous article had characterized the gene in *P. gingivalis*. Sequencing read from Illumina output is shown, as well as if the sequence has an identical match somewhere else in the genome.

Position	Locus	Gene	Product	Sequence	Multiple Identical Seq?
33102	PGN_0026		Cytidine deaminase-like (IPR016193)	ATAATTCCCCGATCGGATTTCTCTGCCG	NO
59184	PGN_0048		PcfK-like protein	CATCTTGACGGAAGTCCAAAGAATGGGG	NO
64784	PGN_0057	*traP*	Conjugate transposon protein TraP	CTACATTCTTGGAAAGCAGAAACTCATT	NO
65060	PGN_0057	*traP*	Conjugate transposon protein TraP	ATACATCTGAGCCGTTACGCTTCAAGGA	NO
107228	PGN_0100		Diaminopimelate decarboxylase LysA	GATATGGGAGGCGGGCTTGGTATCAATT	YES
109125	PGN_0103	*tonB*	TonB	GTACTATGACATACACCCTTGTTTTATA	YES
127453	PGN_0123		PorSS C-terminal domain protein	GTATGCTGCCGAATGTATCTGTATCCCC	NO
141254			Intergenic region 3' to PGN_0136	ATGTTGCGACCGAACAGTCTTTCGGACG	NO
183642			Intergenic region 3' to PGN_0170 (L-asparaginase)	ATATATGAGGAGAATCGCAGAACGTCCT	NO
185308			Intergenic region 3' to PGN_0174 (AraC transcriptional regulator)	CGAGATCCGATCGGATCTTTCATTACTC	NO
195612	PGN_0184	*fimD*	Minor component FimD	TTGGGAGGAAAGGAAGAAAGCAGATGGA	NO
229239	PGN_0215		Hypothetical protein	AATCGAATTATCCAAAGGGTAATAATCG	YES
229519	PGN_r0001		16S ribosomal RNA	CAAGGAAATATTTATAGCTGTAAGATAG	YES
231572			PGN23SrRNA01 / PGN_r0002	AGTATACCGCCGGCGGCAATTGATTGCC	YES
232254	PGN_r0002		23S ribosomal RNA	TAATCAAGACAAGGAGCCGAAGCGAAAG	YES
232826	PGN_r0002		23S ribosomal RNA	TATCGGTAGGGGAGCATTCCAGCGACGT	YES
308700	PGN_0287	*mfa1*	Mfa1 fimbrilin	TATGTCTATGACGTTGTCCATGCCTATG	NO
308911	PGN_0287	*mfa1*	Mfa1 fimbrilin	GATCCTGCAACCCACAATGCCATTTTGA	NO
309047	PGN_0287	*mfa1*	Mfa1 fimbrilin	ATGCAGCAGACTTTGATGCTAAATTCAA	NO
309633	PGN_0287	*mfa1*	Mfa1 fimbrilin	TCAGCTTCAGGACGTGACTGGCGAATTG	NO
309760	PGN_0287	*mfa1*	Mfa1 fimbrilin	TTGGTTCGTGCGAAGTTTACTCCCAAGA	NO
309827	PGN_0287	*mfa1*	Mfa1 fimbrilin	ATACTGCAGTTCCTGAATATGTAGCAGG	NO
311865	PGN_0289	*mfa2*	Fimbrillin-A associated anchor protein Mfa2	CAGGAGAAAAAGTTGCCGAACATTTTTC	NO
312118	PGN_0289	*mfa2*	Fimbrillin-A associated anchor protein Mfa2	TTCCAATCAGACGGTTACGATCGGGGGA	NO
312605	PGN_0289	*mfa2*	Fimbrillin-A associated anchor protein Mfa2	TATCGTATCAAGACGCTGCCGGATAAGA	NO
356645			Intergenic region 3' to PGN_0326	CTATAAAGGGTAGAGGCTTTTAAGTATA	NO
361230			Intergenic region 5' to PGN_0329 (transglutaminase)	GGTGTTTTCTCCGTTCTTTCTATTCTCT	NO
395364	PGN_0361		Glycosyl transferase family 2	TCCATCGTTATTGTCAATTATCGTGTTC	NO
395532	PGN_0361		Glycosyl transferase family 2	GCTAACGAGGAGAATGTAGGTTTCTCTC	NO
415289	PGN_0380	*nagC*	Transcriptional regulator/sugar kinase	TTACACCTCTGCTACAGGAGTAGCCCGT	NO
449720	PGN_0413	*gyrB*	DNA gyrase B subunit	GTCCTAAGAACTATCAGTATTATCCTGC	NO

477203	PGN_0438		Hypothetical protein	AATCGAATTATCCAAAGGGTAATAATCG	YES
477483	PGN_r0004		16S ribosomal RNA	CAAGGAAATATTTATAGCTGTAAGATAG	YES
479536			PGN23SrRNA02 / PGN_r0005	AGTATACCGCCGGCGGCAATTGATTGCC	YES
480218	PGN_r0005		23S ribosomal RNA	TAATCAAGACAAGGAGCCGAAGCGAAAG	YES
480790	PGN_r0005		23S ribosomal RNA	TATCGGTAGGGGAGCATTCCAGCGACGT	YES
511990	PGN_0465	*relA*	GTP pyrophosphokinase	CATGTCCGTGATGATGGAATAGACATGC	NO
550469	PGN_0504	*scpB*	Methylmalonyl-CoA decarboxylase beta subunit	TCCCCTTCAAAGATGCCGGTCTGCAGAT	NO
582965	PGN_0533	*nadA*	Quinolinate synthetase A (IPR003473)	AGCTCCAAGGTTTCTATCCGGGCCAAAG	NO
659816		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
698150	PGN_0637	*htrA*	Heat shock-related protease HtrA protein	CGATGATCTACAGCCAAACGGGCAACTA	NO
698637	PGN_0637	*htrA*	Heat shock-related protease HtrA protein	TGGTGTTTCTTATGGTGTAGAGGTAACC	NO
698656	PGN_0637	*htrA*	Heat shock-related protease HtrA protein	GAGGTAACCTCTGTAAGCTCGGGGAAAT	NO
777121	PGN_0715		Outer membrane efflux protein (IPR003423)	ATATGCGTTTCCAACATTATCTCATCTG	NO
847805	PGN_0778	*porT*	PorT	GTAAATTCAGATTAGGCAGAAGGAACAT	NO
863227	PGN_0789		TPR domain protein	TTATCTGTTCTTTGAGTGCATCCAGTTC	NO
933066			Intergenic region 3' to PGN_0840 (IspD)	GCTACTCAATGTTGCACTGGTGAGTGGA	NO
934830		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
949294	PGN_0862		Type III restriction enzyme, res subunit	TTGATCTCGATCTCGTCGTATTGGAAGA	NO
985175			Intergenic region 5' to PGN_0889 (TkrA)	TTTAGGTCGTGACATCCGAAAAACAATA	NO
1005042	PGN_0903	*fimR*	Two-component system response regulator FimR	CGATTTCTGCCAAACCTTGCATTTGACT	NO
1007054		*fimS*	Intergenic region 5' to PGN_0904 (FimS)	ATTATGTTCAATTACGAGTGAGCCGATC	NO
1054999	PGN_0946		Hypothetical protein	TTCTGGACTATCCGCTCGATATGATGAC	NO
1058741	PGN_0949		ABC transporter ATP-binding protein	CCTCGCTTTTGGGACGCTACCCAAGGCC	NO
1060126	PGN_0950		ABC transporter ATP-binding protein	TAATCCTCAAACTGGGACTTGCCTCTGT	NO
1062335	PGN_0952		Carboxyl-terminal processing protease	AATCCACCACACCGAAAGTAAGCCAAGC	NO
1231667	PGN_1105		FKBP-type peptidyl-prolyl cis-trans isomerase	ATACAAGTCGTAGCTACAGGTCACAAAT	NO
1241487	PGN_1116		Aminotransferase, class I/classII (IPR004839)	GATGATAGAGCAGAACTGTCCTGTCTGC	NO
1247590	PGN_1120		NADPH-NAD transhydrogenase/Alanine dehydrogenase/PNT	GAGCTGATGCGCAAAGGTCAGTATCTGA	NO
1292851		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
1324580		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
1330777			Intergenic region 3' to MurQ and 5' to PGN_1195	GTAAGTTTCTATTTCTTGGGAGGCTTTT	NO
1378741	PGN_1234		Hypothetical protein	GATGTCGGGTGGAAACGTCCGTCTTGCT	NO
1421297	PGN_1272		Diaminopimelate decarboxylase LysA	GATATGGGAGGCGGGCTTGGTATCAATT	YES
1423194	PGN_1275	*tonB*	TonB	GTACTATGACATACACCCTTGTTTTATA	YES
1430756	PGN_1282	*traN*	Conjugate transposon protein TraN	CCTGCTCAAAGGGTTGTATTCGCACAAC	NO
1432269	PGN_1284	*traP*	DNA primase involved in conjugation TraP	TGTACCGTATTCGAGGGCTTTATGGATT	NO
1432545	PGN_1284	*traP*	DNA primase involved in conjugation TraP	GGGAACAATGTATAACGAACAACAAAAC	NO
1444904	PGN_1302	*waaL*	O-antigen ligase	ATAGTAAGCTACCAATAGAACAAGAATA	NO

1444973	PGN_1302	*waaL*	O-antigen ligase	ATAGGAAATACCACAGAAAAGCCATATC	NO
1445729	PGN_1303		Lipoprotein	GTATGATTGATAGCCTTGGCTCCGGCAT	NO
1457223	PGN_1313		Hypothetical protein	TGGTTACCAAAGTCTCCCCTAAAGCGCC	NO
1556843	PGN_r0008		23S ribosomal RNA	ATATATTATATCCCACGGCTTCGGTAAA	NO
1557415	PGN_r0008		23S ribosomal RNA	ATAACAATTAACCTTGCCACAAACAGTA	NO
1558097			rRNA-23S ribosomal RNA PGN_r0008	TTAAGCTTGTTGTGCTCTACTTAAATTC	NO
1560150	PGN_r0009		16S ribosomal RNA	GTAATATCATGCAATAATACAAGTGTAT	NO
1560430			rRNA-16S ribosomal RNA PGN16SrRNA09	TTATCAGGTAGTACACCTCGGTTTCTTT	NO
1598851		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
1705008			Intergenic region 3' to PGN_1523 (Polysaccharide export protein)	CATCAACAACCCCCTTTTTTAAGCCATA	NO
1783632	PGN_1591		Hypothetical protein	CTATTATATGCTCCGACAACGTAAAACG	NO
1810394	PGN_1618		Methionine gamma-lyase	AGCACGAAGGCGTACGCGTCATGGTGGA	NO
1810427	PGN_1618		Methionine gamma-lyase	CCTACTGCACGCCCTATATCTGCCGTCC	NO
1812498		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
1838451	PGN_1643		PF07610 family protein / PapD-like (IPR008962)	CTACAACCCCTTTGATGCGTAGCGTAAA	NO
1861468	PGN_1666	*purL*	Phosphoribosylformylglycinamidine synthase	TACCCCGAACACGAACTCAAGCACAAGC	NO
1865227		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
1870863	PGN_1675	*porL*	Por secretion system protein PorL/GldL	ATATGGTATTCAGGCCGGAGATATTGCG	NO
1888813	PGN_1690		Alpha-L-fucosidase	AGGGGCAGCGCACCGAGGCTTTCACTGT	NO
1922329	PGN_1722		Phosphoribulokinase/uridine kinase (IPR006083)	GGAGTAGCAGGCGGAAGTGCTTCCGGCA	NO
1928710	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGTCTTTCCTACAACCGTCAGCGTGT	NO
1929676	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACCATCACGATATACCGTGTAGGTAT	NO
1930063	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTATCAGGATTCAAAGCCATTCCATTCC	YES
1930361	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTACCGTCTTCTGGCGCCAAGTACCCTG	YES
1930390	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACGACCACGAATAGCTTCCGGCGAGC	YES
1930494	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACACCGCATAGTGCTCGGATGCATAA	YES
1930834	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTAAGCGTTACTTTCTGGCCGACTGCAC	YES
1930841	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTACTTTCTGGCCGACTGCACTACCGGT	YES
1930909	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAACGTCTTTACATACCTTCGGAGATA	YES
1931064	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GCTTGCAGGTGAATCGTCTTCGACTTCC	YES
1931267	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTACAACTTCACCCTGTCCTGTAACGAT	YES
1931293	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAATATTCTGTGTAGTAACAACAGGAT	YES
1931422	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATTGTGATCGGCATCCAACAAGAACT	YES
1931471	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACGTTGTCTGCTGCGAGCACAACCTT	YES
1931513	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACATCGTTTGCAGGTTCGATCGTAAC	YES
1931536	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAACGAAAAGACCGTCTCCGATCCGTT	YES
1931618	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTACTTTCTGACCCTGCGTTGTAGCAGT	NO
1931664	PGN_1728	*kgp*	Lysine-specific cysteine proteinase	GTAGGGGCTAGGCTCTCCTGCCTGAATT	NO

			Kgp		
1931725	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAACTACATCATAATTACCATTTTCCG	NO
1931729	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTACATCATAATTACCATTTTCCGTAAT	NO
1931736	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAATTACCATTTTCCGTAATCTGCTTA	NO
1931739	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATTACCATTTTCCGTAATCTGCTTAGTC	NO
1931740	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTACCATTTTCCGTAATCTGCTTAGTCA	NO
1931767	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATATTCACAGTCGCAACACCGCTGGCAT	NO
1931794	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTAGCAACACCTGTTCCATACAAAACTC	NO
1931837	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTACGTAAGAACCGGCAGAAGCCTGAAT	NO
1931841	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAAGAACCGGCAGAAGCCTGAATGCTA	NO
1931868	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGAAGCCTGATTCTGAGGCAGAGAA	NO
1931905	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATAAGTATTGGTCTTAGGCATTGCAC	NO
1931907	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGTATTGGTCTTAGGCATTGCACGA	NO
1931911	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATTGGTCTTAGGCATTGCACGATAAG	NO
1931967	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGCTTCCCAATAGTAATGAGCACCG	NO
1931982	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAATGAGCACCGATATGGGTAATATTG	NO
1932004	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATATTGCCGATATTTCCAGCATGAGTAG	NO
1932040	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTACCTGCCCACATAATAGAATTCACTG	NO
1932052	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAATAGAATTCACTGTGTTGTACGAAT	NO
1932055	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGAATTCACTGTGTTGTACGAATCTT	NO
1932072	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACGAATCTTCCAAGAATGTAGCATCA	NO
1932091	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAGCATCATAAGAACCCATAGACGTAC	NO
1932094	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GCATCATAAGAACCCATAGACGTACCTT	NO
1932099	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGAACCCATAGACGTACCTTCAAAA	NO
1932115	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACCTTCAAAAGTAGGCTGAACACCAA	NO
1932127	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAGGCTGAACACCAAATACGGCATTAG	NO
1932143	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACGGCATTAGCACCGACACTCCAATA	NO
1932168	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGTAGTCCTCGCCCCAATAAGAATTC	NO
1932208	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATATAGGCATAAGCACCTTTCTCCTTGA	NO
1932210	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGGCATAAGCACCTTTCTCCTTGACA	NO
1932216	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGCACCTTTCTCCTTGACACGAGTC	NO
1932286	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAACACAGCAGTTCCCAATAGCTAAGA	NO
1932315	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATTTGTCCTTATTTGTGAGTGCTTTC	NO
1932397	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATAGTTGGCAAAGCCGACACCGGTAT	NO

1932399	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGTTGGCAAAGCCGACACCGGTATTC	NO
1932435	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGCAGCCTGTATAAGGAGCTTTAGGG	NO
1932445	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATAAGGAGCTTTAGGGTAACTGTACA	NO
1932447	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGGAGCTTTAGGGTAACTGTACACA	NO
1932478	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATAGCCATGATCTTGATTGTAGTAAT	NO
1932480	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGCCATGATCTTGATTGTAGTAATAC	NO
1932498	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAGTAATACTGTACAGCATATTTGATG	NO
1932501	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAATACTGTACAGCATATTTGATGGTT	NO
1932504	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACTGTACAGCATATTTGATGGTTTGC	NO
1932509	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACAGCATATTTGATGGTTTGCTGGCC	NO
1932536	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTATCTTAGGATTCCAGTAGGAGTCAGC	NO
1932541	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTAGGATTCCAGTAGGAGTCAGCACCGG	NO
1932591	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGCTCTTATCCGGCATGGTAGCCTTT	NO
1932621	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACATCAATACCTTATCAATGATGTTC	NO
1932629	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACCTTATCAATGATGTTCGTCAGTTC	NO
1932676	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACGGAAAGTATACATTTCAGGGAAAT	NO
1932685	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATACATTTCAGGGAAATAGTCGCCAT	NO
1932687	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACATTTCAGGGAAATAGTCGCCATCG	NO
1932702	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGTCGCCATCGACTGCACTGTAATAC	NO
1932726	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACAAGTCGGTAACTTTTTTTGTTTTC	NO
1932817	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTAGCTGCCAATCCATCATTGTATTTCT	NO
1932837	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATTTCTTGTGAATAAATGCCTTGATA	NO
1932850	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAATGCCTTGATAGAGGCGTTTGTCG	NO
1932862	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGAGGCGTTTGTCGTTCCTACTTCAG	NO
1932881	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTACTTCAGCTTCGTCTGTGTAATGCAC	NO
1932991	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACGAACCGGCGTATTATACAAGTCGC	NO
1933003	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATTATACAAGTCGCCATGATCTGTAT	NO
1933006	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTATACAAGTCGCCATGATCTGTATAAA	NO
1933008	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACAAGTCGCCATGATCTGTATAAACA	NO
1933027	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATAAACATCTCTATTGAAGAGCTGTT	NO
1933029	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAACATCTCTATTGAAGAGCTGTTTA	NO
1933054	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTATAAGCTGTTTCGAAATAAGGGCTAA	NO
1933056	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGCTGTTTCGAAATAAGGGCTAAAA	NO
1933071	PGN_1728	*kgp*	Lysine-specific cysteine proteinase	ATAAGGGCTAAAAGAAGCATCATACAAA	NO

			Kgp		
1933078	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTAAAAGAAGCATCATACAAACGTTGTG	NO
1933092	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACAAACGTTGTGTAGCTACTTCATCA	NO
1933105	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAGCTACTTCATCAGCTCCTTGAAAGC	NO
1933109	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTACTTCATCAGCTCCTTGAAAGCTTAC	NO
1933133	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTACTTCAATTTCGATGTTGTTTCTAAC	NO
1933185	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATACTGAACAGGATTAATGGTAAGAGCT	NO
1933198	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTAATGGTAAGAGCTGCAATGCGAACAC	NO
1933247	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTACTTGGGTCAGTTCTTGTCCGACAAA	NO
1933287	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAAGCAGCAGCATTGTAAACGAAGGGA	NO
1933342	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	ATAGAGGGTTGATGTGGCATGAGTTTTT	NO
1933377	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTATTGGTTCAGAGAGTAAACTTGCTCG	NO
1933534	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACCACCTTTGGTCTCCACCTTTGTCA	NO
1933606	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	CTATTGTTCGTACATGTCGTTCGAGTAG	NO
1933615	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTACATGTCGTTCGAGTAGTCGGAGCAT	NO
1933648	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	TTAATCTTGGCGCTTTGGGCGTAAAGAC	NO
1933668	PGN_1728	*kgp*	Lysine-specific cysteine proteinase Kgp	GTAAAGACCAACTCCCAAAAGGGACGCC	NO
1936908	PGN_1733	*hagA*	Haemagglutinin protein HagA	ATCGTCTTTCCTACAACTGTCAGCGTGT	NO
1939075	PGN_1733	*hagA*	Haemagglutinin protein HagA	TTACTTTCTGGCCGACTGCACTACCGGT	YES
1940443	PGN_1733	*hagA*	Haemagglutinin protein HagA	TTACTTTCTGGCCGACTGCACTACCGGT	YES
1941811	PGN_1733	*hagA*	Haemagglutinin protein HagA	TTACTTTCTGGCCGACTGCACTACCGGT	YES
1943167	PGN_1733	*hagA*	Haemagglutinin protein HagA	TTACTTTCTGACCTACTGCACTACCGGT	NO
1943747	PGN_1733	*hagA*	Haemagglutinin protein HagA	ATAGTTGTCAGGACTCAAATCAATTTTT	NO
1985140			Intergenic region 3' to PGN_1770	CTTATTGGTGCTGCTCTTTTGGGAGCAG	NO
1985535	PGN_1770		PorSS C-terminal domain protein	TTACATAGTTGGTTATATTGCTAATACC	NO
1997843	PGN_1777		Peptidase C1B, bleomycin hydrolase (IPR004134)	ATACCGTGGTCATCGGTGAGGCTATACT	NO
2010600			Intergenic region 3' to PGN_1791 (Flavodoxin FldA)	GAGAACATTAATTACAAGCCCATGGCAG	NO
2035631	PGN_1812	*ppk*	Polyphosphate kinase	GTCCTTCCTACGAATGCCTTCCATCAGA	NO
2066554			Intergenic region or ISPg3 TIR	GCTACTCAATGTTGCACTGGTGGGTGGA	NO
2119640		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
2123357			Intergenic region 3' to *hagB*	ATATGATAATGGCGAATCACTATGCCAG	NO
2134392	PGN_1914		Carboxyl-terminal processing protease	TTACGGAAGTCTTTCCCATCGATAGTCA	NO
2136601	PGN_1916		ABC transporter ATP-binding protein	ATATATGAAGCATTAAGGACGGCACCGA	NO
2137986	PGN_1917		ABC transporter ATP-binding protein	GTATGAGTCGGGCTATAGTGGTCTTACC	NO
2141728	PGN_1920		Membrane transport protein, MMPL domain (IPR004869)	ATACCCATCAGACCTCCCACTACAATGG	NO
2149326		*ISPg1*	*ISPg1*	TTATGCTGTTTTTAATGCTCAAATATAC	YES
2181602			CRISPR 30-36	ATAGCCATTGACGAGTTTTAATTCCTGT	YES
2181668			CRISPR 30-36	ATAGCCATTGACGAGTTTTAATTCCTGT	YES

2204262	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	ATCGTCTTTCCAACAACTGTCAGCGTGT	NO
2204483	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	GCTTGCAGGTGAATCGTCTTCGACTTCC	YES
2204686	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	TTACAACTTCACCCTGTCCTGTAACGAT	YES
2204712	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	ATAATATTCTGTGTAGTAACAACAGGAT	YES
2204841	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	GTATTGTGATCGGCATCCAACAAGAACT	YES
2204890	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	ATACGTTGTCTGCTGCGAGCACAACCTT	YES
2204932	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	GTACATCGTTTGCAGGTTCGATCGTAAC	YES
2204955	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	GTAACGAAAAGACCGTCTCCGATCCGTT	YES
2205228	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	GTACCGTCACGATACACCGTGTAGGTAT	YES
2205615	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	TTATCAGGATTCAAAGCCATTCCATTCC	YES
2205913	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	CTACCGTCTTCTGGCGCCAAGTACCCTG	YES
2205942	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	ATACGACCACGAATAGCTTCCGGCGAGC	YES
2206046	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	ATACACCGCATAGTGCTCGGATGCATAA	YES
2206473	PGN_1970	*rgpA*	Arginine-specific cysteine proteinase RgpA	GTAACGTCTTTACATACCTTCGGAGATA	YES
2236755	PGN_1998		Outer membrane protein, OmpA/MotB, C-terminal (IPR006665)	TTACTGTCTGAATCGTAGCATCGGGTAC	NO
2249244	PGN_2010		Secreted protein, YngK-like	TGCCACGTTAGCTGAGTGGTGGGGGCGC	NO
2259456	PGN_2017		YjeF family protein	GTAAGGGTGTGTGTGGCTCTGATGATGG	NO
2295481	PGN_2047		Hypothetical protein	AATCGAATTATCCAAAGGGTAATAATCG	YES
2295761	PGN_r0010		16S ribosomal RNA	CAAGGAAATATTTATAGCTGTAAGATAG	YES
2297814			PGN23SrRNA04 / PGN_r00011	AGTATACCGCCGGCGGCAATTGATTGCC	YES
2298496	PGN_r0011		23S ribosomal RNA	TAATCAAGACAAGGAGCCGAAGCGAAAG	YES
2299068	PGN_r0011		23S ribosomal RNA	TATCGGTAGGGGAGCATTCCAGCGACGT	YES
2340345	PGN_2080		Pectin lyase fold/virulence factor (IPR011050) / PorSS CTD	CTTTCTCCAACTGCTATGCTCTCAATGT	NO

Chapter 5: Conclusions and Future Directions

5.1 Significance of This Work

The initial goal of this work, on which all further projects were to be based, was to generate a saturated, sequencing-adapted transposon mutant library in the periodontal pathogen *Porphyromonas gingivalis* for application in screens and selections. The mutant library was successfully created and validated. Several other research laboratories are now utilizing our library directly and we have assisted additional laboratories in generating similar libraries. Following the construction of the mutant library we described in detail the first set of essential genes of an anaerobic pathogen (Klein et al., 2012). Our essential gene set has been used for validation and comparison by multiple *in silico* studies (Wei, Ning, Ye, & Guo, 2013) (Veeranagouda, Husain, Tenorio, & Wexler, 2014). Through the colony pigmentation screen that we performed with the transposon mutant library we identified several novel pigmentation-affecting loci, one of which we have confirmed, as well as a novel miniature inverted-repeat transposable element (Fig.5-1). The BrickBuilt MITE had evaded detection by sequencing/assembly pipelines as well as multiple transposon, insertion sequence and repetitive element identifying programs, as well as researchers who characterized genes adjacent to these elements. Additionally, suggesting the importance of BrickBuilt to *P. gingivalis*, BrickBuilt MITE regions had in multiple cases been identified as significant results following microarray and RNAseq, although none were followed up on presumably due to the complex nature of the regions. Once sufficient understanding of the regions surrounding the putative glycosyltransferase identified in the pigment screen was achieved, that it flanked by two *in vitro* essential genes and immediately adjacent to repetitive element, we were able to confirm its involvement in colony pigmentation as well as provide preliminary characterization of its site and mechanism of action.

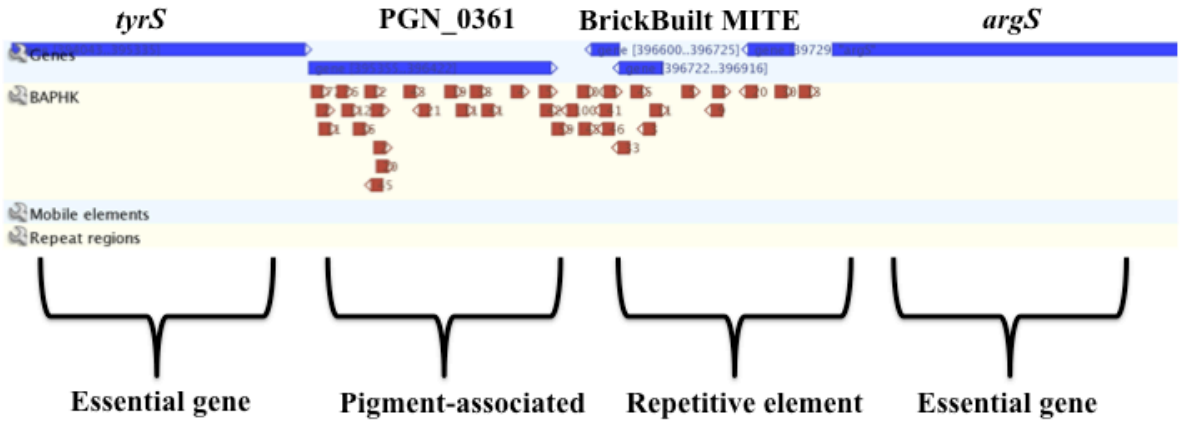


Figure 5-1. Gene Essentiality, BrickBuilt MITE and PGN_0361 Glycosyltransferase Model. GenomeView depiction of BAPHK transposon mutant insertions in the *P. gingivalis* ATCC 33277 strain background, focused on the PGN_0361 putative glycosyltransferase area. *TyrS* and *argS* do not contain transposon insertions because they are essential genes. The 'mobile elements' and 'repeat regions' tracks use the NCBI annotations for this strain, and as such had not identified the BrickBuilt MITE. We isolated multiple independent transposon insertion mutants of PGN_0361/PG0264 that were non-pigmented. This locus was not previously identified using the *Tn4400* or *Tn4351* transposon mutagenesis systems; potentially suggesting additional utility of the Mariner-based system.

5.2 Transposon mutagenesis and Essential Gene Determination

5.2.1 Modifications to pSAM System

The pSAM_Bt transposon mutagenesis vector, which was developed for *B. thetaiotaomicron* by Dr. Andrew Goodman, worked with *P. gingivalis* without any modification (Goodman et al., 2009). During our studies with *P. gingivalis* and work from Dr. Michael Malamy's laboratory with *B. fragilis*, it was determined that insertions of pSAM_Bt into both *P. gingivalis* and *B. fragilis* were having the *ermG* gene almost exclusively transcribed by native chromosomal promoters. This dependence on native chromosomal promoters became evident when following Tn-seq of the transposon mutant libraries the transposon insertions were frequently oriented unidirectionally within a protein coding sequence. Upon further inspection of the genic and intergenic regions throughout *P. gingivalis* that did have bidirectional insertion orientation, multiple promoters were frequently assumed or have been show to be present due to endogenous pseudogenes, Tn, IS or CTn elements. Additionally, overlapping convergent promoters from genes that lack transcriptional terminators also provide opportunities for transcriptional driving of the *ermG* gene present on the transposon. The lack of transcriptional terminators, overlapping transcripts and intergenic transcription has previously been identified through microarray and RNAseq studies with *P. gingivalis* (Høvik et al., 2012).

After our initial pSAM_Bt-based libraries were constructed in strains ATCC 33277 and W83, Dr. Jeff Gordon's laboratory, where Dr. Andrew Goodman first utilized pSAM_Bt in *B. thetaiotaomicron*, developed a second mutagenesis vector using the pSAM plasmid backbone (McNulty et al., 2013). In this version, which they entitled pSAM_WH2, a promoter for the *ermG* gene from *Bacteroides cellulosilyticus* WH2 was inserted where

either only a weak or potentially omitted promoter was previously located. Initial mutagenesis using strains ATCC 33277 and W83 with pSAM_WH2 was significantly more efficient than with pSAM_Bt, suggesting that the changes within pSAM_WH2 may give rise to more insertion mutants and/or allow for more mutants to survive the antibiotic selection used to separate wild-type from mutant strains following conjugation.

The utility and efficiency of pSAM-based vectors for transposon mutagenesis in *P. gingivalis* is evident, yet additional modification to the vectors may expand functionality. First, addition of an outward-reading inducible promoter would give the ability to show effects of knockout (technically interruption) at a given site in a genome as well as expression (or over expression) under a given condition. Second, a transcriptional reporter added to the transposon would allow for expression monitoring of individual mutants. Third, addition of a tag such a SNAP or CLIP tags (NEB) would allow for visualization and potentially collection of mutagenized proteins; the use of SNAP tags with *P. gingivalis* has previously been demonstrated (Nicolle et al., 2010).

Importantly, the original pSAM_Bt vector has additional functionality after coupling it with the pSAM_WH2 vector and the knowledge that, at least in *P. gingivalis* and potentially *B. fragilis*, pSAM_Bt insertions will be located exclusively downstream of native chromosomal promoters. Once saturated transposon mutant libraries are generated using both the pSAM_Bt and pSAM_WH2 vectors in the wild-type same strain background and then Tn-seq is performed on these pooled mutant libraries, detailed promoter mapping of a strain (and/or a species) could be carried out. Given that there has yet to be a consensus promoter sequence or architecture determined for any *Porphyromonas* species this could provide

valuable information with respect to genetic manipulations, transcriptional regulation and evolution.

5.2.2 In Silico, In Vitro and In Vivo Essentiality Analyses

The term 'gene essentiality' is only as useful and descriptive as the context in which it is used. The most basic definition of 'gene essentiality' is that of a region within a genome that is absolutely required for survival under a specific condition. Growth following survival is not required in the most basic definition given that removal of the test/selection condition generally follows, which allows for growth or outgrowth of the surviving population. Technically, all descriptions of essentiality are 'conditional' essentiality, given that each study used to determine what is essential must have a given set of defined parameters. The majority of gene essentiality studies to date have been performed *in vitro*, although several recent studies have taken to using *in vivo* model systems as well.

For *in vitro* and *in vivo* essential gene studies to be useful as standalone projects, able to provide inter-study comparisons, as well as be the foundation for *in silico* model development, conditional characteristics must be defined. For *in vitro* studies examples include strain of microbe within the species, growth medium, temperature, time, light exposure, humidity, oxygen concentration, time allowed for growth to be evident, method for the generation of mutants, level of saturation, sequencing methods and statistical methods. In addition to the *in vitro* examples *in vivo* studies should also include species and strain of animal, animal husbandry methods and methods for procuring microbial samples back from test animals.

In silico analyses of gene essentiality are now possible given that *in vitro* and *in vivo* studies have been performed. Three main types of *in silico* analyses can be done using prior studies as input; building and validation of a computational program, cross-species (or strain) comparison and extrapolation to other species, strains or conditions.

5.2.3 Utility of Gene Essentiality Studies

The utility of gene essentiality studies ranges from understanding the basic physiological and metabolic properties of organisms to the translational development of targeted therapeutics. *In vitro* studies can provide information relating to nutritional pathways and requirements, susceptibility to antimicrobial compounds and defined interspecies interactions. *In vivo* studies can expand upon *in vitro* information in potentially more realistic or physiologically relevant model systems. Additionally, *in vivo* studies provide a platform through which to preemptively study evolutionary and/or infection outbreak dynamics. *In silico* studies potentially provide their greatest benefit as platforms for prediction of gene essentiality for less studied or more difficult to manipulate organisms in which it is not yet feasible to perform such studies directly, as well as in biopharmaceutical research applications. An example of such an *in silico* study would be the overlay, and then if possible after an *in vitro* study, comparison of *P. gingivalis* and other *Porphyromonas* species such as *P. gulae* or *P. endodontalis*. Such extrapolations to other species could influence or help limit failed genetic manipulations directed at putatively essential loci. Other examples of valuable *in silico* studies would be the identification of species-specific essential genes or the comparison of shared essential genes with the focus on finding varying structural differences between species that can be exploited with targeted therapeutics. On a basic level, platforms

such as the Database of Essential Genes (DEG) could be mined and then coupled with PHYRE2 for protein structural prediction to determine if specific binding sites for cofactors or potential inhibitor binding pockets vary between species (Luo, Lin, Gao, Zhang, & Zhang, 2014) (Kelley et al., 2015).

Other future applications of essentiality using transposon mutagenesis systems would be to define essential domains in non-essential genes, domains within genes causing essentiality, as well as using species-specific, genus-specific and phyla-specific essentiality in an evolutionary context.

5.3 Identification of a Miniature Inverted-repeat Transposable Element

5.3.1 MITEs in Porphyromonas gingivalis

MITEs were initially identified and characterized in eukaryotic species, primarily plants, but have begun to be identified in archaeal and bacterial species. The majority of the bacterial MITEs have been identified in 'pathogenic' and 'opportunistic' species, yet this is likely due to factors such as the presence and more numerous whole genome sequencing projects, greater number of laboratories and more government funding to study such species, genetic tractability, and greater desire to understand the intergenic and repetitive regions of a genome where transcriptional regulation and recombination may contribute to pathogenicity and dissemination.

There are now four different non-autonomous transposable element versions of the miniature inverted-repeat transposable element class within *P. gingivalis*. Two of the four versions, MITE464 (MITEPgRS) and BrickBuilt, contain tandem direct repeats in the middle of the element. Neither the 41 bp MITE464 (MITEPgRS) repeat or the 23 bp BrickBuilt

repeat have been identified in a genome outside of *P. gingivalis* or *P. gulae* to date. When compiled, the total percent of the *P. gingivalis* genome encoded by MITEs is 1%; 0.44% from BrickBuilt elements, 0.39% from MITEPgRS elements, and the rest from MITE700 and MITE239 family elements. General copy numbers of the four MITE versions are similar, holding around 10-20. The number of full copies and partial or fragment copies of each element differs slightly between genomes within the species. The terminal inverted repeats (TIRs) of BrickBuilt and MITEPgRS elements are almost identical, as are the TIRs of MITE293 and MITE700 elements. Each of the MITE-like elements in *P. gingivalis* share either identical or within one nucleotide TIRs with those of full-length IS elements within the *P. gingivalis* genomes; *ISPg1*, *ISPg3*, *ISPg4* and *ISPg9*. The matching full-length ISPg elements are all categorized within the IS5 family. BrickBuilt and MITEPgRS TIRs are most similar to those of *ISPg1* and *ISPg9* (which share identical TIRs), and MITE293 and MITE700 TIRs match with *ISPg3*.

5.3.2 Origin of BrickBuilt MITE

With the sequencing projects to date, including full genomes, partial genomes, targeted regions, and plasmids, BrickBuilt elements are only present in *Porphyromonas gingivalis* and *Porphyromonas gulae*. In our initial identification and characterization of BrickBuilt we did not aggressively attempt to identify the origin (e.g. viral, bacterial or eukaryotic) of the element. Phages have not yet been isolated and demonstrated to infect *P. gingivalis*, nor have endogenous plasmids been found within any strain of the species.

Given that BrickBuilt and MITEPgRS elements share similar TIRs, one part of the origin of BrickBuilt MITE elements is probably tied to the presence and origin of

MITEPgRS elements as well. Importantly, since aside from the TIRs little homology is found between the MITEs or the autonomous IS elements in *P. gingivalis* genomes, the majority of the MITE sequences within *P. gingivalis* must have originally either been from exogenous sources such as phages or plasmids, horizontal transfer, or degraded over the evolution of the species. Degradation of the original sequences seems unlikely given the lack of homologous element portions within *Bacteroidetes* aside from *P. gulae*. Horizontal gene transfer also seems unlikely for how the element would have first come into *P. gingivalis* because genetic manipulation of and passage to *P. gingivalis* has been difficult due to low competency, multiple restriction modification systems and a lack of endogenous plasmids. Horizontal transfer could, however, be a manner in which BrickBuilt elements are passed between *P. gingivalis* strains, as populations can be non-clonal within a host and have been shown to exchange genetic information. Phages or potentially plasmids are most likely the source of BrickBuilt elements within *P. gingivalis*. Several of the BrickBuilt elements are flanked by genes associated with phages. Additionally, all BrickBuilt elements are located intergenically, do not carry extra CDS, and are frequently associated with repetitive CDS. Intergenic locations are areas where exogenous elements can insert into and potentially thrive because of generally lower selection pressure and not inactivating host genes. Repetitive host proteins would potentially allow for limited disruption of host physiology following insertion, but could also reveal an insertion site preference for the original element.

The oral cavity, which is constantly bathed with food, drink and the surrounding environment as the entryway to the gastrointestinal tract, is filled with many distinct species of bacteria as well as high bacterial loads. As a site with easy access and many host cells, the oral cavity should be a good niche for bacteriophage to thrive. From several studies focusing

on isolating phage from the oral cavity, although several phage have been isolated and shown to have lytic cycles in non-oral community species, no phage have been isolated specific to any oral pathogens. Although a company focused on phage biology for therapeutic use (PhageWorks) has previously claimed to have isolated a phage that infects *P. gingivalis*, they have not provided any detail nor have they published on or shown a product based on this purported finding (initially in 2012). Present and active CRISPR systems are a likely reason why the isolation of phages specific for oral pathogens may be difficult. Several studies of *P. gingivalis* CRISPR systems have been carried out and showed that all sequenced strains carried these anti-infective systems, that the systems were actively adapting, and that many sequences from exogenous origin were already captured by the systems (Burmistrz et al., 2015). No BrickBuilt sequences were found within the CRISPR systems. Cases of self-targeting inclusions in CRISPR systems have been found (Stern, Keren, Wurtzel, Amitai, & Sorek, 2010). The lack of BrickBuilt being selected by the CRISPR systems may suggest that its presence is not deleterious to the host.

5.3.3 BrickBuilt in Porphyromonas gulae

The identification of multiple BrickBuilt elements in a species other than *Porphyromonas gingivalis*, especially given its phylogenetic relation, is interesting and worthy of further study. Twelve *P. gulae* genomes are currently publically available. The twelve strains used for sequencing were isolated in two different studies; one strain from a wolf in Canada and eleven strains from domestic client-owned dogs at a veterinary dental clinic in the United Kingdom. Importantly, at this point none of the sequencing projects are categorized as complete chromosomes; all are either scaffold or contig level assemblies.

Two key findings emerged following our work characterizing BrickBuilt in *P. gulae*; several BrickBuilt elements are located in the same loci as they are in *P. gingivalis* and there are significant strain-level differences. At a single locus in *P. gulae*, BrickBuilt_5 in *P. gingivalis*, four strains contain a BrickBuilt element and seven strains do not. Within the four *P. gulae* strains that do have a BrickBuilt element at that locus, there are three different versions present that are variable in the 23 bp tandem repeat segment. Interestingly, within the ~6 kb segment several SNPs align the four BrickBuilt-containing *P. gulae* with the *P. gingivalis* strains, and many SNPs differentiate the BrickBuilt-lacking *P. gulae* from both the BrickBuilt-containing *P. gulae* as well as the *P. gingivalis*. Additionally, the strain isolated from a wolf, which was designated the type-strain for *P. gulae* does not contain a BrickBuilt element at this locus.

5.3.4 MITE Expansion: A Genomic and Application Perspective

On a genetic and genomic level, the identification, sequencing/assembly and further characterization of BrickBuilt elements in *P. gingivalis* and *P. gulae* could lead to key insights on intra-genus and intra-species evolution via transposable element dynamics. Full assemblies of sequence data, preferably without the use of previously-sequenced scaffold genomes, will be necessary to reliably study MITEs. Coupling multiple sequencing platforms may aid gap closures, as might the use of Nanopore-type technologies.

Experimental evolution studies and footprinting or pull-down assays could help determine the mobility and function of BrickBuilt elements, respectively. Serial passaging of several strains and several clones of *P. gingivalis* in a single growth medium, followed by

whole genome sequencing and alignment, could demonstrate the frequency at which BrickBuilt elements recombine or acquire mutations. Using multiple growth media and environmental stresses may allow for identification of regulators of or triggers for BrickBuilt activation.

MITEs may have therapeutic potential as antimicrobial targets. Three main reasons for targeting MITEs, especially BrickBuilt, are apparent. First, MITEs tend to have limited host range. Species-specificity as opposed to broad-spectrum actions is gaining praise given the risk of antimicrobial resistance. With a keystone pathogen of a polymicrobial disease such as periodontitis, singling out *P. gingivalis* without harming the rest of the ecological niche is important. Second, MITEs are frequently found in high copy number. Successfully mutating multiple high copy number elements before the activity of an oligo-specific antimicrobial causes damage has low probability. Several CRISPR targeting experiments have already been carried out and the ability to kill a species, even when within a mixed population, was demonstrated; accentuating the killing by using multiple identical sites within a target host would make the process more efficient and have a higher probability of success. Third, MITE sequences could easily be screened against the human genome in order to make sure no off target hits are found within the host.

5.4 Mapping and Characterization of Colony Pigmentation-associated Loci

5.4.1 Why Pigment?

Four key aspects help to answer the question of why to study microbial pigments. First, pigmentation is an easily identifiable macroscopic phenotype. Second, pigments have previously been shown to have important physiological, biomedical and industrial functions

or applications. Third, although simple to physically identify, the underlying genetic and metabolic pathways associated with pigments provide ample opportunities to study regulatory and enzymatic mechanisms in depth. Lastly, the previous three aspects can be worked on while also maintaining a lay relevance or 'hook' for basic, clinical and funding purposes.

Finding a microbial phenotype to study that can be seen by the unaided eye is frequently not a possibility. However, if a species of interest produces a pigment, or a species involved in a community of interest produces a pigment, a visual, assayable and targetable topic of study is evident. Pigments can be extracted and isolated. Different methods may be required to obtain the pigment compound depending upon its makeup, but the compound should be able to be traced throughout the process macroscopically or through spectrophotometry and/or chromatography. After the identification and isolation of a pigment, the structure and chemical composition are usually targeted, followed by the mechanisms and metabolites required for its synthesis in conjunction with the physiological relevance.

Microbial pigments can vary between being expressed on an almost constitutive basis to that of a highly limited occurrence. In the case of *P. gingivalis*, the routine *in vitro* plating medium of blood agar leads to the characteristic black colony pigmentation. Although pigmentation of *P. gingivalis* is used as a diagnostic of the species and is considered a major virulence factor, the pathways, metabolites and regulation of pigmentation involve considerable questions still to be answered. One of the simplest questions that still remains unanswered for *P. gingivalis* pigmentation is why leaving blood out of plate-based *in vitro* growth media causes colonies to never pigment, even if excess porphyrin compounds are

added to the medium as supplementation. Multiple studies have suggested that gingipains liberate the haem molecules from red blood cells (RBCs) to start the pigment production cascade, yet the supplementation of haem in the place of RBCs does not rescue pigmentation. Maybe a 'trigger' molecule is required in addition to the haem that results in pigmentation. Or possibly parts of degraded RBCs are physically necessary for complexing haem to the bacterial cell surface. Another remaining and pressing two-part question relates to if *P. gingivalis* is a haem auxotroph even though most of the haem biosynthesis pathway is still present and functional in the species. Six of the canonical eight genes in haem biosynthesis are present in *P. gingivalis*. Several species have already been shown to navigate around the loss of a single gene within the pathway by using a different enzyme from another pathway or gaining a second mechanism for a protein already in the pathway to complement the loss. Metabolite supplementing studies have shown that *P. gingivalis* enzymes in the porphyrin pathway are functional, yet since endogenous haem has not been reported, either the system is non-functional or the regulation of the system is unknown.

5.4.2 Limitations of and Complications with Pigment Screening

All screens and selections have intrinsic limitations and complications. Inherent in our colony pigmentation screen of *P. gingivalis* are four main factors that may influence the results. First, the wild-type strain background of *P. gingivalis* will influence the overall pigment phenotype as well as the loci identified by our screen method. Second, the method of mutant library generation will dictate what insertion mutants are generated. Third, the screening and selection methods with respect to colony/clone selection will bias toward

certain mutants. Fourth, sequencing and subsequent data analysis will affect what is found and considered significant.

With respect to wild-type strain background selection and influencing the screen results, strain ATCC 33277 is known to pigment dark black between 3-7 days, lacks or maintains insignificant capsule, generate both types of fimbriae and already had a completed genome sequencing and assembly project. We also initially worked in tandem with strain W83 because it also had a completed genome sequencing and assembly project and slightly different phenotypic properties than strain ATCC 33277 (elaborates capsule, lacks major fimbriae and survives better without exogenous haem), but did not select clones for Tn-seq analysis. However, dozens of individual pigment-deficient clones were subjected to nested semi-random sequencing for identification of transposon insertions affecting colony pigmentation. Our previous work with strains ATCC33277 and W83 *in vitro* had revealed non-pigmented spontaneous mutants during single broth growths and plating onto BAP only at extremely low frequency. Strains such as HG66, which is relatively non-pigmenting, and beige or brown pigmented strains were not selected because intrinsic mutations already affecting pigmentation are unknown.

Concerning our pSAM_Bt-based conjugation method for generating the mutant library, biases of the transposon, vector construct matters and growth or survival complications during and after transposition will influence downstream results. Mariner transposons are known to target 'TA' dinucleotide sites with a suggested frequency of 98-99% under most *in vitro* conditions. Highly unusual DNA sequences or DNA topology, as well as temperature differences may influence the frequency of non-TA insertion sites. As previously mentioned above with respect to pSAM_Bt, the insertions into *P. gingivalis* using

this vector construct only occurred (following antimicrobial selection) in the genome where chromosomally encoded promoters were present to drive the transcription of the *ermG* gene. Insertions could have occurred where promoters were not present or not strong enough to clear the 2-5 μg/ml antimicrobial selection level but would not have made it into the pooled mutant library. The methods used with respect to the growth media and incubation times for generating the transposon mutants were described in our original article and subsequent methods article (Klein et al., 2015). Of note, mutagenesis was carried out during a 5 hr aerobic conjugation with *E. coli* carrying the pSAM_Bt vector. *P. gingivalis* survives well for that length of aerobic exposure, but it is still a stress on the obligate anaerobe species. Certain mutants with altered resistance to oxygen exposure may survive worse during the aerobic conjugation period and thus never make it into the mutant pool. Additionally, clones that divided during the 5-hour period may lead to selection of siblings in downstream applications or chances for second conjugation events. Anaerobic conjugation is a possibility, yet overall efficiency may be lower due to regulation in *E. coli*. Furthermore, using BAPHK medium for selection may not allow for all mutants to survive (or survive well enough to make visible colonies). BAPHK does not contain vitamin, mineral or serum supplements. Addition of such supplements may increase the number of clones within the library as well as allow for the identification of genes involved in colony pigmentation if connections to vitamin and serum-supplied nutrients are involved in the pathway.

With regard to the screen and selection methods, human visual error, experimenter bias and total clone selection numbers will affect target identification. Pigmentation defects were screened for between days 5-14 of growth. Defects that manifest earlier or later in growth would not be included in the final pools (e.g. a fast pigmenting or a strain intrinsically

prone to spontaneous mutations that would manifest initially as pigmented and then as sectored). Additionally, the mutants were selected by two different experimenters that may each have slightly different concepts on pigmentation defects. Finally, the total number of clones selected could have been larger. However, three separate screens of single-fold library coverage were carried out, such that if genes were present in the library that affect pigmentation but were not recovered they would likely be at extremely low numbers.

Concerning sequencing and data analysis, the presence of repetitive sequences, length of sequencing reads and necessity for cut-off levels in high-throughput sequencing data can influence results. The Illumina HiSeq2000 sequence reads that we obtained from our Tn-seq experiments were no longer than 50 bp. Other sequencing programs within Illumina as well as other sequencing platforms can garner longer sequencing reads. Connected to sequence read length is the matter of repetitive sequences within a genome. If a sequence involved in colony pigmentation has a repetitive sequence, the true locus will not be able to be differentiated between the identical loci via Tn-seq alone. Longer sequencing reads may alleviate this problem, however, individual confirmation may be necessary regardless of read length or sequencing platform. At a low frequency 'random' sequence reads will appear in the results following Tn-seq even if sequence was not in the initial mutant pool. As such, a cutoff must be introduced to eliminates such sequence reads as being regarded as 'true' hits. Reasons for such events occurring include applying multiple bar-coded samples to a single sequencing lane, resulting in 'bleed through' between the barcodes, as well as having samples of low complexity, resulting in highly similar sequences that occasionally get miss-sequenced and end up within the results.

5.4.3 Conservation of Pigment and/or Pigment-Associated Genes

Of the genes affecting pigmentation identified in *P. gingivalis* to date the genomic location of such genes could be described as 'abundant yet scattered'. While several operons have been identified to affect pigmentation such as the three haem/iron acquisition systems, and occasionally genes such as gingipains and haemagglutinins involved in pigmentation are located closely, genes involved in pigmentation are distributed regularly throughout the genome. Other species occasionally cluster virulence factors into pathogenicity islands or plasmids, and perhaps chromosomally interspersed nature of pigmentation genes in *P. gingivalis* confers an evolutionary advantage such as avoiding plasmid loss and incompatibility, or gives greater control of transcriptional regulation.

Thirteen of fifteen species within *Porphyromonas* display colony pigmentation. The two species of *Porphyromonas* that do not or have high strain variation in pigment production are *P. bennonis* and *P. catoniae*. Strains of *P. bennonis* are of human origin, yet comprise isolates from abscesses throughout the body; not within or near the oral cavity. Strains of *P. catoniae* are of human origin and isolated from the oral cavity, yet unlike all other *Porphyromonas* they are saccharolytic. Several possibilities for the lack or variable pigmentation of these species are plausible. With *P. bennonis* samples being isolated from abscesses the microenvironment or niche could be drastically different in each sample cases as well as from *Porphyromonas* species niches. Patients with abscesses are likely to have been given antimicrobials and may also be immunosuppressed, necessitating or driving the loss of pigmentation. The saccharolytic nature of *P. catoniae* strains may account for the lack of colony pigmentation. Given that the majority of genes involved in pigmentation involve gingipain proteases, which are the proteins required to extract haem from RBCs, are involved

in complexing haem to the cell surface and generate the peptides necessary for growth by the asaccharolytic *P. gingivalis*, the use of sugars for energy may have alleviated the need for or required the loss of pigmentation.

5.4.4 What More For Pigment?

To gain a complete picture of pigmentation in *P. gingivalis* many more studies will need to be carried out. With the development of sequencing technologies, many more methods are available to identify novel aspects of the pigmentation pathway.

The original experiments that isolated pigment-deficient mutants, where wild-type strains were grown in a chemostat for varying lengths of time, could be reassessed using whole genome sequencing. In fact, to this day no one has determined the exact mutations that the input strain in the chemostat experiments underwent in order to become the beige or brown lineages. To accentuate the mutation frequency, a mutagen could be applied to the wild-type strain and pigmentation-affected clones could be analyzed by whole genome sequencing as well. In each of these types of experiments, wild-type and several mutant strains would need to be analyzed to determine the frequency of background mutations such that target identification can occur more quickly. Additionally, using known pigment-deficient strains one could screen for suppressor or phenotype reversion mutants. If screening for suppressor, RNAseq may prove necessary if the suppression is not a fixed chromosomal DNA mutation. Furthermore, performing RNAseq on pigment-affected strains without any phenotype reversion may reveal how strains lacking such a pleiotropic factor adapt to the environment.

Although the ability to display pigment has been shown as a key virulence factor in numerous *in vitro* and *in vivo* assays, the elaboration of pigment has never been visualized at static points or in real time *in vivo*. As such, an assay to determine when pigmentation occurs and/or is important in vivo could be carried out by tracking transcription of genes in the pigmentation pathway as well as monitoring haem deposition on the cell surface. *In vivo* transcriptional tracking could be carried out using reporters or RNAseq, and haem deposition could potentially be monitored using animal host models that have altered haem biosynthesis or via a labeling system that specific targets μ-oxo bis haem dimers.

With respect directly to the putative glycosyltransferase PGN_0361/PG0264 and its mechanism of action, effects on pigmentation and physiological relevance, several additional experiments could be performed. First, the monosaccharide composition of LPS from the mutant strain could be determined and compared to the wild-type parent strains through GC-MS. Second, NMR spectroscopy could be employed for structural determination of LPS moieties from wild-type and mutant strains. If alterations to lipid-A moieties are suspected, LC-MS of that specific fragment may provide insight as well (MS of whole intact LPS is currently not an option). Third, *in vitro* screens against antimicrobial compounds could be applied to determine if the gene modulates susceptibility against potential therapeutics. LPS-based alterations frequently affect antimicrobial compounds due to changes in surface structure and target molecule.

Chapter 6: References

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., & Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *Journal of Clinical Microbiology, 43*(11), 5721-5732.

Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., Judson, N., & Mekalanos, J. J. (2002). A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proceedings of the National Academy of Sciences of the United States of America, 99*(2), 966-971.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research, 40*(Web Server issue), W597-603. doi:10.1093/nar/gks400 [doi]

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The keio collection. *Molecular Systems Biology, 2*

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W.W., Noble, W. S. (2009). MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Research, 37*(Web Server issue), W202-8. doi:10.1093/nar/gkp335 [doi]

Bainbridge, B. W., Hirano, T., Grieshaber, N., & Davey, M. E. (2015). Deletion of a 77-base-pair inverted repeat element alters the synthesis of surface polysaccharides in *Porphyromonas gingivalis*. *Journal of Bacteriology, 197*(7), 1208-1220.

Bellaousov, S., Reuter, J. S., Seetin, M. G., & Mathews, D. H. (2013). RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Research, 41*(Web Server issue), W471-4. doi:10.1093/nar/gkt290 [doi]

Bennetzen, J. L., & Wang, H. (2014). *The contributions of transposable elements to the structure, function, and evolution of plant genomes*

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research, 27*(2), 573-580. doi:gkc131 [pii]

Bertels, F., & Rainey, P. B. (2011). Within-genome evolution of REPINs: A new family of miniature mobile DNA in bacteria. *PLoS Genetics, 7*(6), e1002132. doi:10.1371/journal.pgen.1002132 [doi]

Brunner, J., Wittink, F. R. A., Jonker, M. J., De Jong, M., Breit, T. M., Laine, M. L., De Soet, J.J, Crielaard, W. (2010). The core genome of the anaerobic oral pathogenic bacterium *Porphyromonas gingivalis*. *BMC Microbiology, 10*

Bryan, G., Garza, D., & Hartl, D. (1990). Insertion and excision of the transposable element mariner in *Drosophila*. *Genetics, 125*(1), 103-114.

Bureau, T. E., & Wessler, S. R. (1992). Tourist: A large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell, 4*(10), 1283-1294.

Bureau, T. E., & Wessler, S. R. (1994). Stowaway: A new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell, 6*(6), 907-916.

Burmistrz, M., Dudek, B., Staniec, D., Rodriguez Martinez, J. I., Bochtler, M., Potempa, J., & Pyrc, K. (2015). Functional analysis of *Porphyromonas gingivalis* W83 CRISPR-cas systems. *Journal of Bacteriology, 197*(16), 2631-2641. doi:10.1128/JB.00261-15 [doi]

Califano, J. V., Arimoto, T., & Kitten, T. (2003). The genetic relatedness of *Porphyromonas gingivalis* clinical and laboratory strains assessed by analysis of insertion sequence (IS) element distribution. *Journal of Periodontal Research, 38*(4), 411-416. doi:665 [pii]

Califano, J. V., Kitten, T., Lewis, J. P., Macrina, F. L., Fleischmann, R. D., Fraser, C. M., Duncan, M.J., Dewhirst, F. E. (2000). Characterization of *Porphyromonas gingivalis* insertion sequence-like element ISPg5. *Infection and Immunity, 68*(9), 5247-5253.

Cameron, D. E., Urbach, J. M., & Mekalanos, J. J. (2008). A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America, 105*(25), 8736-8741.

Chandad, F., Mayrand, D., Grenier, D., Hinode, D., & Mouton, C. (1996). Selection and phenotypic characterization of nonhemagglutinating mutants of *Porphyromonas gingivalis*. *Infection and Immunity, 64*(3), 952-958.

Chang, T. H., Huang, H. Y., Hsu, J. B., Weng, S. L., Horng, J. T., & Huang, H. D. (2013). An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics, 14 Suppl 2*, S4-2105-14-S2-S4. Epub 2013 Jan 21. doi:10.1186/1471-2105-14-S2-S4 [doi]

Chaudhuri, R. R., Allen, A. G., Owen, P. J., Shalom, G., Stone, K., Harrison, M., Burgis, T.A., Lockyer, M., Garcia-Lara, J., Foster, S.J., Pleasance, S.J., Peters, S.E., Maskell, D.J., Charles, I. G. (2009). Comprehensive identification of essential *Staphylococcus aureus* genes using transposon-mediated differential hybridisation (TMDH). *BMC Genomics, 10*

Chen, S., & Li, X. (2007). Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes. *BMC Evolutionary Biology, 7*, 46.

doi:1471-2148-7-46 [pii]

Chen, S. L., & Shapiro, L. (2003). Identification of long intergenic repeat sequences associated with DNA methylation sites in *Caulobacter crescentus* and other alpha-proteobacteria. *Journal of Bacteriology, 185*(16), 4997-5002.

Chen, T., Abbey, K., Deng, W., & Cheng, M. (2005). The bioinformatics resource for oral pathogens. *Nucleic Acids Research, 33*, W734-W740.

Chen, T., Dong, H., Tang, Y. P., Dallas, M. M., Malamy, M. H., & Duncan, M. J. (2000). Identification and cloning of genes from porphyromonas gingivalis after mutagenesis with a modified Tn4400 transposon from *Bacteroides fragilis*. *Infection and Immunity, 68*(1), 420-423.

Chen, T., Dong, H., Yong, R., & Duncan, M. J. (2000). Pleiotropic pigmentation mutants of *Porphyromonas gingivalis*. *Microbial Pathogenesis, 28*(4), 235-247.

Chen, T., & Duncan, M. J. (2004). Gingipain adhesin domains mediate *Porphyromonas gingivalis* adherence to epithelial cells. *Microbial Pathogenesis, 36*(4), 205-209.

Chen, T., Hosogi, Y., Nishikawa, K., Abbey, K., Fleischmann, R. D., Walling, J., & Duncan, M. J. (2004). Comparative whole-genome analysis of virulent and avirulent strains of *Porphyromonas gingivalis*. *Journal of Bacteriology, 186*(16), 5473-5479.

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., Fero, M.J., McAdams, H.H., Shapiro, L. (2011). The essential genome of a bacterium. *Molecular Systems Biology, 7*

Coates, B. S., Kroemer, J. A., Sumerford, D. V., & Hellmich, R. L. (2011). A novel class of miniature inverted repeat transposable elements (MITEs) that contain hitchhiking (GTCY)n microsatellites. *Insect Molecular Biology, 20*(1), 15-27.

Coats, S. R., To, T. T., Jain, S., Braham, P. H., & Darveau, R. P. (2009). *Porphyromonas gingivalis* resistance to polymyxin B is determined by the lipid-A 4'-phosphatase, PGN_0524. *International Journal of Oral Science, 1*(3), 126-135.

Coenye, T., & Vandamme, P. (2005). Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes, 12*(4), 221-233. doi:12/4/221 [pii]

Coil, D. A., Alexiev, A., Wallis, C., O'Flynn, C., Deusch, O., Davis, I., Horsfall, A., Kirkwood, N., Jospin, G., Eisen, J.A., Harris, S., Darling, A. E. (2015). Draft genome sequences of 26 *Porphyromonas* strains isolated from the canine oral microbiome. *Genome Announcements, 3*(2), 10.1128/genomeA.00187-15. doi:10.1128/genomeA.00187-15 [doi]

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research, 16*(22), 10881-10890.

Curtis, M. A., Zenobia, C., & Darveau, R. P. (2011). The relationship of the oral microbiotia to periodontal health and disease. *Cell Host & Microbe, 10*(4), 302-306. doi:10.1016/j.chom.2011.09.008 [doi]

Darmon, E., & Leach, D. R. (2014). Bacterial genome instability. *Microbiology and Molecular Biology Reviews : MMBR, 78*(1), 1-39. doi:10.1128/MMBR.00035-13 [doi]

Darveau, R. P. (2010). Periodontitis: A polymicrobial disruption of host homeostasis. *Nature Reviews Microbiology, 8*(7), 481-490.

De Berardinis, V., Vallenet, D., Castelli, V., Besnard, M., Pinet, A., Cruaud, C., Samair, S., Lechaplais, C., Gyapay, G., Richez, C., Durot, M., Kreimeyer, A., Le Fevre, F., Schachter, V., Pezo, V., Doring, V., Scarpelli, C., Medigue, C., Cohen, G.N., Marliere,

P., Salanoubat, M., Weissenbach, J. (2008). A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Molecular Systems Biology, 4*

De Gregorio, E., Silvestro, G., Petrillo, M., Carlomagno, M. S., & Di Nocera, P. P. (2005). Enterobacterial repetitive intergenic consensus sequence repeats in *Yersiniae*: Genomic organization and functional properties. *Journal of Bacteriology, 187*(23), 7945-7954. doi:187/23/7945 [pii]

Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. *Genome Biology and Evolution, 3*(1), 959-973.

Deng, H., Shu, D., Luo, D., Gong, T., Sun, F., & Tan, H. (2013). Scatter: A novel family of miniature inverted-repeat transposable elements in the fungus *Botrytis cinerea*. *Journal of Basic Microbiology, 53*(10), 815-822.

Díaz, L., Hoare, A., Soto, C., Bugueño, I., Silva, N., Dutzan, N., Venegas, D., Salinas, D., Pérez-Donoso, J.M., Gamonal, J, Bravo, D. (2015). *Changes in lipopolysaccharide profile of Porphyromonas gingivalis clinical isolates correlate with changes in colony morphology and polymyxin B resistance*

Duncan, M. J. (2003). Genomics of oral bacteria. *Critical Reviews in Oral Biology and Medicine, 14*(3), 175-187.

Fattash, I., Rooke, R., Wong, A., Hui, C., Luu, T., Bhardwaj, P., & Yang, G. (2013). Miniature inverted-repeat transposable elements: Discovery, distribution, and activity. *Genome, 56*(9), 475-486.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics, 9*(5), 397-405.

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: Where

genetics meets genomics. *Nature Reviews.Genetics, 3*(5), 329-341. doi:10.1038/nrg793 [doi]

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J., Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research, 42*(Database issue), D222-30. doi:10.1093/nar/gkt1223 [doi]

French, C. T., Lao, P., Loraine, A. E., Matthews, B. T., Yu, H., & Dybvig, K. (2008). Large-scale transposon mutagenesis of *Mycoplasma pulmonis*. *Molecular Microbiology, 69*(1), 67-76.

Gallagher, L. A., Ramage, E., Jacobs, M. A., Kaul, R., Brittnacher, M., & Manoil, C. (2007). A comprehensive transposon mutant library of *Francisella novicida*, a bioweapon surrogate. *Proceedings of the National Academy of Sciences of the United States of America, 104*(3), 1009-1014.

Gao, F., & Zhang, R. R. (2011). Enzymes are enriched in bacterial essential genes. *Plos One, 6*(6)

Gawronski, J. D., Wong, S. M., Giannoukos, G., Ward, D. V., & Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proceedings of the National Academy of Sciences of the United States of America, 106*(38), 16422-16427.

Gelfand, Y., Rodriguez, A., & Benson, G. (2007). TRDB--the tandem repeats database. *Nucleic Acids Research, 35*(Database issue), D80-7. doi:gkl1013 [pii]

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balezsi, G., Ravasz, E., Daugherty, M. D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., Bhattacharya, A., Kapatral,

V., D'Souza, M., Baev, M.V., Grechkin, Y., Mseeh, F., Fonstein, M.Y., Overbeek, R., Barabasi, A.-L., Oltvai, Z.N., Osterman, A. L. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology, 185*(19), 5673-5684.

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison III, C.A., Smith, H.O., Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 425-430.

Gonzalez, J., & Petrov, D. (2009). MITEs - the ultimate parasites. *Science, 325*(5946), 1352-1353.

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., Gordon, J. I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host & Microbe, 6*(3), 279-289.

Guo, Y., Nguyen, K. A., & Potempa, J. (2010). Dichotomy of gingipains action as virulence factors: From cleaving substrates with the precision of a surgeon's knife to a meat chopper-like brutal degradation of proteins. *Periodontology 2000, 54*(1), 15-44.

Høvik, H., Wen-Han, Y., Olsen, I., & Chen, T. (2012). Comprehensive transcriptome analysis of the periodontopathogenic bacterium *Porphyromonas gingivalis* W83. *Journal of Bacteriology, 194*(1), 100-114.

Hajishengallis, G. (2015). Periodontitis: From microbial immune subversion to systemic inflammation. *Nature Reviews.Immunology, 15*(1), 30-44. doi:10.1038/nri3785 [doi]

Hajishengallis, G., & Lamont, R. J. (2014). Breaking bad: Manipulation of the host response by porphyromonas gingivalis. *European Journal of Immunology, 44*(2), 328-338.

Hajishengallis, G., Liang, S., Payne, M. A., Hashim, A., Jotwani, R., Eskan, M. A., McIntosh, M.L., Alsam, A., Kirkwood, K.L., Lambris, J.D., Darveau,R.P., Curtis, M. A. (2011). Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement. *Cell Host and Microbe, 10*(5), 497-506.

Halász, J., Kodad, O., & Hegedus, A. (2014). Identification of a recently active prunus-specific non-autonomous mutator element with considerable genome shaping force. *Plant Journal,*

Han, Y., & Korban, S. S. (2007). Spring: A novel family of miniature inverted-repeat transposable elements is associated with genes in apple. *Genomics, 90*(2), 195-200.

Hayes, F. (2003). *Transposon-based strategies for microbial functional genomics and proteomics*

Hensel, M., & Holden, D. W. (1996). Molecular genetic approaches for the study of virulence in both pathogenic bacteria and fungi. *Microbiology, 142*(5), 1049-1058.

Hikosaka, A., & Kawahara, A. (2004). Lineage-specific tandem repeats riding on a transposable element of MITE in *Xenopus* evolution: A new mechanism for creating simple sequence repeats. *Journal of Molecular Evolution, 59*(6), 738-746.

Hirano, T., Beck, D. A., Demuth, D. R., Hackett, M., & Lamont, R. J. (2012). Deep sequencing of *Porphyromonas gingivalis* and comparative transcriptome analysis of a LuxS mutant. *Frontiers in Cellular and Infection Microbiology, 2*, 79.

Holt, S. C., Ebersole, J., Felton, J., Brunsvold, M., & Kornman, K. S. (1988). Implantation of bacteroides gingivalis in nonhuman primates initiates progression of periodontitis. *Science (New York, N.Y.), 239*(4835), 55-57.

Holt, S. C., Kesavalu, L., Walker, S., & Genco, C. A. (1999). Virulence factors of *Porphyromonas gingivalis*. *Periodontology 2000, 20*, 168-238.

Hoover, C. I., Abarbarchuk, E., Ng, C. Y., & Felton, J. R. (1992). Transposition of Tn4351 in *Porphyromonas gingivalis*. *Plasmid, 27*(3), 246-250.

Hoover, C. I., & Yoshimura, F. (1994). Transposon-induced pigment-deficient mutants of *Porphyromonas gingivalis*. *FEMS Microbiology Letters, 124*(1), 43-48.

Høvik, H., Wen-Han, Y., Olsen, I., & Chen, T. (2012). Comprehensive transcriptome analysis of the periodontopathogenic bacterium *Porphyromonas gingivalis* W83. *Journal of Bacteriology, 194*(1), 100-114.

Hug, I., & Feldman, M. F. (2011). *Analogies and homologies in lipopolysaccharide and glycoprotein biosynthesis in bacteria*

Igboin, C. O., Griffen, A. L., & Leys, E. J. (2009). *Porphyromonas gingivalis* strain diversity. *Journal of Clinical Microbiology, 47*(10), 3073-3081.

Ilyina, T. S. (2010). Miniature repetitive mobile elements of bacteria: Structural organization and properties. *Molecular Genetics, Microbiology and Virology, 25*(4), 139-147.

Imamura, T. (2003). The role of gingipains in the pathogenesis of periodontal disease. *Journal of Periodontology, 74*(1), 111-118.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics (Oxford, England), 30*(9), 1236-1240. doi:10.1093/bioinformatics/btu031 [doi]

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J.,

Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., Caspi, R. (2010). Pathway tools version 13.0: Integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics, 11*(1), 40-79. doi:10.1093/bib/bbp043 [doi]

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P, Drummond, A. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England), 28*(12), 1647-1649. doi:10.1093/bioinformatics/bts199 [doi]

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols, 10*(6), 845-858. doi:10.1038/nprot.2015.053 [doi]

Khatiwara, A., Jiang, T., Sung, S., Dawoud, T., Kim, J. N., Bhattacharya, D., Kim, H.B., Ricke, S.C., Kwon, Y. M. (2012). Genome scanning for conditionally essential genes in *Salmonella enterica* serotype typhimurium. *Applied and Environmental Microbiology, 78*(9), 3098-3107.

Klein, B. A., Duncan, M. J., & Hu, L. T. (2015). *Defining essential genes and identifying virulence factors of porphyromonas gingivalis by massively parallel sequencing of transposon libraries (tn-seq)*

Klein, B. A., Tenorio, E. L., Lazinski, D. W., Camilli, A., Duncan, M. J., & Hu, L. T. (2012). Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*. *BMC Genomics, 13*, 578-2164-13-578.

Knuth, K., Niesalla, H., Hueck, C. J., & Fuchs, T. M. (2004). Large-scale identification of

essential *Salmonella* genes by trapping lethal insertions. *Molecular Microbiology, 51*(6), 1729-1744.

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Dabarbouilla, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., le Coq, D., Masson, A., Maual, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F.M.L., Sekiguchi, J., Sekowska, A., Saror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., van Dijl, J.M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U, Ogasawara, N. (2003). Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences of the United States of America, 100*(8), 4678-4683.

Koressaar, T., & Remm, M. (2012). Characterization of species-specific repeats in 613 prokaryotic species. *DNA Research, 19*(3), 219-230.

Kuboniwa, M., Hendrickson, E. L., Xia, Q., Wang, T., Xie, H., Hackett, M., & Lamont, R. J. (2009). Proteomics of porphyromonas gingivalis within a model oral microbial

community. *BMC Microbiology, 9*, 98-2180-9-98. doi:10.1186/1471-2180-9-98 [doi]

Kuramitsu, H. K. (2003). Molecular genetic analysis of the virulence of oral bacterial pathogens: An historical perspective. *Critical Reviews in Oral Biology and Medicine : An Official Publication of the American Association of Oral Biologists, 14*(5), 331-344.

Lamont, R. J., & Jenkinson, H. F. (1998). Life below the gum line: Pathogenic mechanisms of *Porphyromonas gingivalis*. *Microbiology and Molecular Biology Reviews : MMBR, 62*(4), 1244-1263.

Lampe, D. J., Akerley, B. J., Rubin, E. J., Mekalanos, J. J., & Robertson, H. M. (1999). Hyperactive transposase mutants of the Himar1 mariner transposon. *Proceedings of the National Academy of Sciences of the United States of America, 96*(20), 11428-11433.

Lampe, D. J., Churchill, M. E. A., & Robertson, H. M. (1996). A purified mariner transposase is sufficient to mediate transposition in vitro. *EMBO Journal, 15*(19), 5470-5479.

Langridge, G. C., Phan, M., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., . . . Turner, A. K. (2009). Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Research, 19*(12), 2308-2316.

Lau, G. W., Ran, H., Kong, F., Hassett, D. J., & Mavrodi, D. (2004). *Pseudomonas aeruginosa* pyocyanin is critical for lung infection in mice. *Infection and Immunity, 72*(7), 4275-4278. doi:10.1128/IAI.72.7.4275-4278.2004 [doi]

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osteras, M., Schrenzel, J, Francois, P. (2009). Metagenomic study of the oral microbiota by illumina high-throughput sequencing. *Journal of Microbiological Methods, 79*(3), 266-271.

Lewis, J. P. (2010). Metal uptake in host-pathogen interactions: Role of iron in

Porphyromonas gingivalis interactions with host organisms. *Periodontology 2000, 52*(1), 94-116. doi:10.1111/j.1600-0757.2009.00329.x [doi]

Lewis, J. P., & Macrina, F. L. (1998). IS195, an insertion sequence-like element associated with protease genes in *Porphyromonas gingivalis*. *Infection and Immunity, 66*(7), 3035-3042.

Lin, Y., Gao, F., & Zhang, C. (2010). Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochemical and Biophysical Research Communications, 396*(2), 472-476.

Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews.Genetics, 14*(1), 49-61. doi:10.1038/nrg3374 [doi]

Liu, D., Zhou, Y., Naito, M., Yumoto, H., Li, Q., Miyake, Y., Liang, J., Shu, R. (2014). Draft genome sequence of *Porphyromonas gingivalis* strain SJD2, isolated from the periodontal pocket of a patient with periodontitis in china. *Genome Announcements, 2*(1), 10.1128/genomeA.01091-13. doi:10.1128/genomeA.01091-13 [doi]

Liu, G. Y., & Nizet, V. (2009). Color me bad: Microbial pigments as virulence factors. *Trends in Microbiology, 17*(9), 406-413.

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W., & Kuang, H. (2012). Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa. *Molecular Biology and Evolution, 29*(3), 1005-1017.

Luo, H., Lin, Y., Gao, F., Zhang, C., & Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research, 42*(D1), D574-D580.

Maeda, K., Nagata, H., Ojima, M., & Amano, A. (2014). Proteomic and transcriptional analysis of interaction between oral microbiota *Porphyromonas gingivalis* and *Streptococcus oralis*. *Journal of Proteome Research, 14*(1), 82-94.

Maley, J., & Roberts, I. S. (1994). Characterisation of IS1126 from *Porphyromonas gingivalis* W83: A new member of the IS4 family of insertion sequence elements. *FEMS Microbiology Letters, 123*(1-2), 219-224.

Mazurkiewicz, P., Tang, C. M., Boone, C., & Holden, D. W. (2006). Signature-tagged mutagenesis: Barcoding mutants for genome-wide screens. *Nature Reviews Genetics, 7*(12), 929-939.

McKee, A. S., McDermid, A. S., Wait, R., Baskerville, A., & Marsh, P. D. (1988). Isolation of colonial variants of *Bacteroides gingivalis* W50 with a reduced virulence. *Journal of Medical Microbiology, 27*(1), 59-64.

McLean, J. S., Lombardo, M., Ziegler, M. G., Novotny, M., Yee-Greenbaum, J., Badger, J. H., Tesler, G., Nurk, S., Lesin, V., Brami, D., Hall, A.P., Edlund, A., Allen, L.Z., Durkin, S., Reed, S., Torriani, F., Nealson, K.H., Pevzner, P.A., Friedman, R., Venter, J.C., Lasken, R. S. (2013). Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome Research, 23*(5), 867-877.

McNulty, N. P., Wu, M., Erickson, A. R., Pan, C., Erickson, B. K., Martens, E. C., Pudlo, N.A., Muegge, B.D., Henrissat, B., Hettich, R.L., Gordon, J. I. (2013). Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome. *PLoS Biology, 11*(8), e1001637. doi:10.1371/journal.pbio.1001637 [doi]

Metris, A., Reuter, M., Gaskin, D. J. H., Baranyi, J., & van Vliet, A. H. M. (2011). In vivo and in silico determination of essential genes of *Campylobacter jejuni*. *BMC Genomics, 12*

Molzen, T. E., Burghout, P., Bootsma, H. J., Brandt, C. T., Der Gaast-De Jongh, C. E. V., Eleveld, M. J., Verbeek, M.M., Frimodt-Maller, N., Ostergaard, C., Hermans, P. W. M. (2011). Genome-wide identification of *Streptococcus pneumoniae* genes essential for bacterial replication during experimental meningitis. *Infection and Immunity, 79*(1), 288-297.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., Okumoto, Y., Tanisaka, T., Wessler, S. R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature, 461*(7267), 1130-1134.

Naito, M., Hirakawa, H., Yamashita, A., Ohara, N., Shoji, M., Yukitake, H., Nakeyama, K., Toh, H., Yoshimura, F., Kuhara, S., Hattori, M., Hayashi, T., Nakayama, K. (2008). Determination of the genome sequence of *Porphyromonas gingivalis* strain ATCC 33277 and genomic comparison with strain W83 revealed extensive genome rearrangements in *P. gingivalis*. *DNA Research, 15*(4), 215-225.

Naito, M., Sato, K., Shoji, M., Yukitake, H., Ogura, Y., Hayashi, T., & Nakayama, K. (2011). Characterization of the *Porphyromonas gingivalis* conjugative transposon CTnPg1: Determination of the integration site and the genes essential for conjugal transfer. *Microbiology, 157*(7), 2022-2032.

Nakayama, K. (2003). Molecular genetics of *Porphyromonas gingivalis*: Gingipains and other virulence factors. *Current Protein & Peptide Science, 4*(6), 389-395.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., Finn, R. D. (2015). Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Research, 43*(Database issue), D130-7. doi:10.1093/nar/gku1063 [doi]

Nelson, K. E., Fleischmann, R. D., DeBoy, R. T., Paulsen, I. T., Fouts, D. E., Eisen, J. A., Daugherty, S.C., Dodson, R.J., Durkin, A.S., Gwinn, M., Haft, D.H., Kolonay, J.F., Nelson, W.C., Mason, T., Tallon, L., Gray, J., Granger, D., Tettelin, H., Dong, H., Galvin, J.L., Duncan, M.J., Dewhirst, F.E., Fraser, C. M. (2003). Complete genome sequence of the oral pathogenic bacterium *Porphyromonas gingivalis* strain W83. *Journal of Bacteriology, 185*(18), 5591-5601.

Nelson, W. C., Bhaya, D., & Heidelberg, J. F. (2012). Novel miniature transposable elements in thermophilic *Synechococcus* strains and their impact on an environmental population. *Journal of Bacteriology, 194*(14), 3636-3642.

Nicolle, O., Rouillon, A., Guyodo, H., Tamanai-Shacoori, Z., Chandad, F., Meuric, V., & Bonnaure-Mallet, M. (2010). Development of SNAP-tag-mediated live cell labeling as an alternative to GFP in *Porphyromonas gingivalis*. *FEMS Immunology and Medical Microbiology, 59*(3), 357-363.

Nishikawa, K., Yoshimura, F., & Duncan, M. J. (2004). A regulation cascade controls expression of *Porphyromonas gingivalis* fimbriae via the FimR response regulator. *Molecular Microbiology, 54*(2), 546-560. doi:MMI4291 [pii]

O'Brien-Simpson, N.M, Veith, P.D, Dashper, S.G,Reynolds, E.C. (2003). *Porphyromonas gingivalis* gingipains: The molecular teeth of a microbial vampire. *Current Protein and Peptide Science, 4*(6), 409.

Olczak, T., Simpson, W., Liu, X., & Genco, C. A. (2005). Iron and heme utilization in *Porphyromonas gingivalis*. *FEMS Microbiology Reviews, 29*(1), 119-144.

Osbourne, D. O., Aruni, W., Roy, F., Perry, C., Sandberg, L., Muthiah, A., & Fletcher, H. M. (2010). Role of vimA in cell surface biogenesis in *Porphyromonas gingivalis*. *Microbiology (Reading, England), 156*(Pt 7), 2180-2193.

Ozmeric, N., Preus, N. R., & Olsen, I. (2000). Genetic diversity of *Porphyromonas gingivalis* and its possible importance to pathogenicity. *Acta Odontologica Scandinavica, 58*(4), 183-187.

Padeken, J., Zeller, P., & Gasser, S. M. (2015). Repeat DNA in genome organization and stability. *Current Opinion in Genetics and Development, 31*, 12-19.

Paramonov, N., Aduse-Opoku, J., Hashim, A., Rangarajan, M., & Curtis, M. A. (2015). Identification of the linkage between A-polysaccharide and the core in the A-lipopolysaccharide of *Porphyromonas gingivalis* W50. *Journal of Bacteriology, 197*(10), 1735-1746. doi:10.1128/JB.02562-14 [doi]

Paramonov, N., Bailey, D., Rangarajan, M., Hashim, A., Kelly, G., Curtis, M. A., & Hounsell, E. F. (2001). *Structural analysis of the polysaccharide from the lipopolysaccharide of Porphyromonas gingivalis strain W50*

Paramonov, N. A., Aduse-Opoku, J., Hashim, A., Rangarajan, M., & Curtis, M. A. (2009). Structural analysis of the core region of O-lipopolysaccharide of *Porphyromonas gingivalis* from mutants defective in O-antigen ligase and O-antigen polymerase. *Journal of Bacteriology, 191*(16), 5272-5282.

Parkhill, J., Achtman, M., James, K. D., Bentley, S. D., Churcher, C., Klee, S. R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R.M., Davis, P., Devlin, K.,

Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M.A., Rajandream, M.A., Rutherford, K.M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B.G, Barrell, B. G. (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature, 404*(6777), 502-506. doi:10.1038/35006655 [doi]

Petrillo, M., Silvestro, G., Di Nocera, P. P., Boccia, A., & Paolella, G. (2006). Stem-loop structures in prokaryotic genomes. *BMC Genomics, 7*, 170.

Phillips, P., Progulske-Fox, A., Grieshaber, S., & Grieshaber, N. (2014). Expression of *Porphyromonas gingivalis* small RNA in response to hemin availability identified using microarray and RNA-seq analysis. *FEMS Microbiology Letters, 351*(2), 202-208.

Picardeau, M. (2010). Transposition of fly mariner elements into bacteria as a genetic tool for mutagenesis. *Genetica, 138*(5), 551-558.

Piégu, B., Bire, S., Arensburger, P., & Bigot, Y. (2015). A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. *Molecular Phylogenetics and Evolution, 86*, 90-109.

Pihlstrom, B. L., Michalowicz, B. S., & Johnson N.W. (2005). Periodontal diseases. *Lancet, 366*, 1809-1820.

Potempa, J., Banbula, A., & Travis, J. (2000). Role of bacterial proteinases in matrix destruction and modulation of host responses. *Periodontology 2000, 24*, 153-192.

Potempa, J., Sroka, A., Imamura, T., & Travis, J. (2003). Gingipains, the major cysteine proteinases and virulence factors of *Porphyromonas gingivalis*: Structure, function and assembly of multidomain protein complexes. *Current Protein and Peptide Science, 4*(6), 397-407.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R. D. (2012). The pfam protein families database. *Nucleic Acids Research, 40*, D290-D301.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Research, 33*, W116-W120.

Reznikoff, W. S., & Winterberg, K. M. (2007). *Transposon-based strategies for the identification of essential bacterial genes*

Roper, J. M., Raux, E., Brindley, A. A., Schubert, H. L., Gharbia, S. E., Shah, H. N., & Warren, M. J. (2000). The enigma of cobalamin (vitamin B12) biosynthesis in *Porphyromonas gingivalis*: Identification and characterization of a functional corrin pathway. *Journal of Biological Chemistry, 275*(51), 40316-40323.

Sakamoto, M. (2013). *The family porphyromonadaceae*

Salama, N. R., Shepherd, B., & Falkow, S. (2004). Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *Journal of Bacteriology, 186*(23), 7926-7935.

Sampath, P., Murukarthick, J., Izzah, N. K., Lee, J., Choi, H., Shirasawa, K., Choi, B.S., Liu, S., Nou, I.S, Yang, T. (2014). Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in brassica rapa and B. oleracea. *Plos One, 9*(4)

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2003). Genes required for *Mycobacterial* growth defined by high density mutagenesis. *Molecular Microbiology, 48*(1), 77-84.

Sato, K. (2011). Por secretion system of *Porphyromonas gingivalis*. *Journal of Oral Biosciences, 53*(3), 187-196.

Satovic, E., & Plohl, M. (2013). Tandem repeat-containing MITEs in the clam *Donax trunculus*. *Genome Biology and Evolution, 5*(12), 2549-2559.

Sawada, K., Kokeguchi, S., Hongyo, H., Sawada, S., Miyamoto, M., Maeda, H., Nishimura, F., Takashiba, S., Murayama, Y. (1999). Identification by subtractive hybridization of a novel insertion sequence specific for virulent strains of *Porphyromonas gingivalis*. *Infection and Immunity, 67*(11), 5621-5625.

Schneider, K., & Beck, C. F. (1986). Promoter-probe vectors for the analysis of divergently arranged promoters. *Gene, 42*(1), 37-48. doi:0378-1119(86)90148-4 [pii]

Scholle, M. D., & Gerdes, S. Y. (2007). *Whole-genome detection of conditionally essential and dispensable genes in Escherichia coli via genetic footprinting*

Seers, C. A., Slakeski, N., Veith, P. D., Nikolof, T., Chen, Y. -., Dashper, S. G., & Reynolds, E. C. (2006). The RgpB C-terminal domain has a role in attachment of RgpB to the outer membrane and belongs to a novel C-terminal-domain family found in *Porphyromonas gingivalis*. *Journal of Bacteriology, 188*(17), 6376-6386.

Siddiqui, H., Yoder-Himes, D. R., Mizgalska, D., Nguyen, K. A., Potempa, J., & Olsen, I. (2014). Genome sequence of *Porphyromonas gingivalis* strain HG66 (DSM 28984). *Genome Announcements, 2*(5), 10.1128/genomeA.00947-14. doi:10.1128/genomeA.00947-14 [doi]

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega.

Molecular Systems Biology, 7, 539. doi:10.1038/msb.2011.75 [doi]

Siguier, P., Gourbeyre, E., & Chandler, M. (2014). Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiology Reviews, 38*(5), 865-891.

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews.Genetics, 8*(4), 272-285. doi:nrg2072 [pii]

Smalley, J. W., Silver, J., Marsh, P. J., & Birss, A. J. (1998). The periodontopathogen *Porphyromonas gingivalis* binds iron protoporphyrin IX in the μ-oxo dimeric form: An oxidative buffer and possible pathogenic mechanism. *Biochemical Journal, 331*(3), 681-685.

Song, J., Ko, K. S., Lee, J., Baek, J. Y., Oh, W. S., Yoon, H. S., Jeong, J., Chun, J. (2005). Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Molecules and Cells, 19*(3), 365-374.

Stahl, M., & Stintzi, A. (2011). Identification of essential genes in *C. jejuni* genome highlights hyper-variable plasticity regions. *Functional and Integrative Genomics, 11*(2), 241-257.

Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. (2010). Self-targeting by CRISPR: Gene regulation or autoimmunity? *Trends in Genetics, 26*(8), 335-340.

Takada, K., & Hirasawa, M. (1998). Tn4351-generated non-haemolytic and/or non-pigmented mutants of *Porphyromonas gingivalis*. *Microbios, 95*(380), 35-44.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D.M., Mazumder, R., Smirnov, S., Nikolskaya, A.N., Rao, B.S., Mekhedov, S.L., Sverlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D. A. (2003). The COG database: An updated vesion includes eukaryotes. *BMC Bioinformatics, 4*

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research, 28*(1), 33-36.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E. V. (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research, 29*(1), 22-28.

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., & Tolstoy, I. (2015). RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Research, 43*(7), 3872. doi:10.1093/nar/gkv278 [doi]

Treangen, T. J., Abraham, A. L., Touchon, M., & Rocha, E. P. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews, 33*(3), 539-571.

Tribble, G. D., Kerr, J. E., & Wang, B. (2013). Genetic diversity in the oral pathogen *Porphyromonas gingivalis*: Molecular mechanisms and biological consequences. *Future Microbiology, 8*(5), 607-620.

Tzeng, Y. L., Ambrose, K. D., Zughaier, S., Zhou, X., Miller, Y. K., Shafer, W. M., & Stephens, D. S. (2005). Cationic antimicrobial peptide resistance in *Neisseria meningitidis*. *Journal of Bacteriology, 187*(15), 5387-5396.

Van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods, 6*(10), 767-772.

Van Opijnen, T., & Camilli, A. (2010). Genome-wide fitness and genetic interactions determined by tn-seq, a high-throughput massively parallel sequencing method for

microorganisms. *Current Protocols in Microbiology,* (SUPPL.19)

Veeranagouda, Y., Husain, F., Tenorio, E. L., & Wexler, H. M. (2014). Identification of genes required for the survival of *B. fragilis* using massive parallel sequencing of a saturated transposon mutant library. *BMC Genomics, 15*, 429-2164-15-429. doi:10.1186/1471-2164-15-429 [doi]

Wandersman, C., & Delepelaire, P. (2004). Bacterial iron sources: From siderophores to hemophores. *Annual Review of Microbiology, 58*, 611-647. doi:10.1146/annurev.micro.58.030603.123811 [doi]

Wang, C., Bond, V. C., & Genco, C. A. (1997). Identification of a second endogenous *Porphyromonas gingivalis* insertion element. *Journal of Bacteriology, 179*(11), 3808-3812.

Wang, X., Tan, J., Bai, Z., Su, H., Deng, X., Li, Z., Zhou, C., Chen, J. (2013). Detection and characterization of miniature inverted-repeat transposable elements in "*Candidatus liberibacter asiaticus*". *Journal of Bacteriology, 195*(17), 3979-3986.

Wang, X. G., Lin, B., Kidder, J. M., Telford, S., & Hu, L. T. (2002). Effects of environmental changes on expression of the oligopeptide permease (opp) genes of *Borrelia burgdorferi*. *Journal of Bacteriology, 184*(22), 6198-6206.

Watanabe, T., Maruyama, F., Nozawa, T., Aoki, A., Okano, S., Shibata, Y., Oshima, K., Kurokawa, K., Hattori, M., Nakagawa, I., Abiko, Y. (2011). Complete genome sequence of the bacterium *Porphyromonas gingivalis* TDC60, which causes periodontal disease. *Journal of Bacteriology, 193*(16), 4259-4260.

Wei, W., Ning, L., Ye, Y., & Guo, F. (2013). Geptop: A gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *Plos One, 8*(8)

Westesson, O., Skinner, M., & Holmes, I. (2013). Visualizing next-generation sequencing data with JBrowse. *Briefings in Bioinformatics, 14*(2), 172-177. doi:10.1093/bib/bbr078 [doi]

Wilks, A., & Burkhard, K. A. (2007). Heme and virulence: How bacterial pathogens regulate, transport and utilize heme. *Natural Product Reports, 24*(3), 511-522. doi:10.1039/b604193k [doi]

Wong, S. M. S., & Akerley, B. J. (2007). *Identification and analysis of essential genes in Haemophilus influenzae*

Xia, Q., Wang, T., Taub, F., Park, Y., Capestany, C. A., Lamont, R. J., & Hackett, M. (2007). Quantitative proteomics of intracellular *Porphyromonas gingivalis*. *Proteomics, 7*(23), 4323-4337. doi:10.1002/pmic.200700543 [doi]

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L.V., Gordon, J. I. (2003). A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science, 299*(5615), 2074-2076.

Xu, P., Ge, X., Chen, L., Wang, X., Dou, Y., Xu, J. Z., Patel, J.R., Stone, V., Trinh, M., Evans, K., Kitten, T., Bonchev, D., Buck, G. A. (2011). Genome-wide essential gene identification in *Streptococcus sanguinis*. *Scientific Reports, 1*

Yamaguchi, M., Sato, K., Yukitake, H., Noiri, Y., Ebisu, S., & Nakayama, K. (2010). A *Porphyromonas gingivalis* mutant defective in a putative glycosyltransferase exhibits defective biosynthesis of the polysaccharide portions of lipopolysaccharide, decreased gingipain activities, strong autoaggregation, and increased biofilm formation. *Infection and Immunity, 78*(9), 3801-3812.

Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y.,

Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., Datsenko, K.A., Nakayashiki, T., Tomita, M., Wanner, B.L., Mori, H. (2009). Update on the keio collection of *Escherichia coli* single-gene deletion mutants. *Molecular Systems Biology, 5*

Yang, H. P., & Barbash, D. A. (2008). Abundant and species-specific DINE-1 transposable elements in 12 *Drosophila* genomes. *Genome Biology, 9*(2), R39-2008-9-2-r39. Epub 2008 Feb 21. doi:10.1186/gb-2008-9-2-r39 [doi]

Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface components of *Porphyromonas gingivalis*. *Journal of Periodontal Research, 44*(1), 1-12.

Zhang, C., & Zhang, R. (2007). *Gene essentiality analysis based on DEG, a database of essential genes*

Zhang, H. H., Xu, H. E., Shen, Y. H., Han, M. J., & Zhang, Z. (2013). The origin and evolution of six miniature inverted-repeat transposable elements in *Bombyx mori* and *Rhodnius prolixus*. *Genome Biology and Evolution, 5*(11), 2020-2031.

Zhang, R., & Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Research, 37*, D455-D458.

Zhang, R., Ou, H., & Zhang, C. (2004). DEG: A database of essential genes. *Nucleic Acids Research, 32*(DATABASE ISS.), D271-D272.

Zhou, F., Tran, T., & Xu, Y. (2008). Nezha, a novel active miniature inverted-repeat transposable element in *Cyanobacteria*. *Biochemical and Biophysical Research Communications, 365*(4), 790-794.

Zhou, K., Aertsen, A., & Michiels, C. W. (2014). The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiology Reviews, 38*(1), 119-141. doi:10.1111/1574-

6976.12036 [doi]

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research, 31*(13), 3406-3415.