Genomic Analysis of the Classical *Bordetella*

Eric T. Harvill, Dept. of Veterinary and Biomedical Sciences, Penn State

Vivek Kapur, Dept. of Veterinary and Biomedical Sciences, Penn State

Ying Zhang, Dept. of Veterinary and Biomedical Sciences, Penn State

Karen Register, National Animal Disease Center, USDA/ARS, Ames, IA

Tracy Nicholson, National Animal Disease Center, USDA/ARS, Ames, IA

James M. Musser, Department of Pathology and Laboratory Medicine, Weill Cornell Medical College of Cornell University

Andrew Preston, Veterinary Public Health University of Bristol, Univ. of Bristol, U.K.

I.) INTRODUCTION

Nine Gram-negative species of bacteria comprise the genus *Bordetella*, a group of common respiratory pathogens. *B. bronchiseptica, B. pertussis*, and *B. parapertussis* are the three most commonly studied species within this genus and are referred to as the "classical *Bordetella*." *B. pertussis* and *B. parapertussis* are the causative agents of whooping cough or pertussis in humans, while *B. bronchiseptica* colonizes a broad range of mammalian hosts and causes various disease severities ranging from lethal pneumonia to asymptomatic respiratory infections (Goodnow RA, 1980). *B. pertussis* and *B. parapertussis* are believed to have evolved independently from a *B. bronchiseptica*-like progenitor (Diavatopoulos DA et al., 2005), and strong genetic similarities between three sequenced representative strains suggest the classical *Bordetella* be classified as subspecies (Preston A et al., 2004).

Whooping cough or pertussis is prevalent throughout the world, resulting in an estimated 16 million cases and 195,000 deaths per year in developing countries (WHO, 2008). Despite high vaccination rates in industrialized countries, whooping cough cases have been steadily increasing over the past three decades, leading to its recent classification by the CDC as a re-emerging disease (Celentano, L. P. et al. 2005, CDC. 2002, de Melker, H. E., et al. 2000). The most recent outbreak of whooping cough in 2010 killed 10 babies in California as reported by California Department of Public Health, and whooping cough remains one of the major threats to newborns worldwide. Although *B. pertussis* is considered the main cause of whooping cough, *B. parapertussis* causes clinical symptoms that are indistinguishable from *B. pertussis* infections (Heininger, U. et al, 1994). *B. parapertussis* has been shown to be just as prevalent within some populations as *B. pertussis*, if not more prevalent (Watanabe, M. and M. Nagai, 2004). Additionally, *B. parapertussis* has been broken down into two clades, one restricted to humans (*B. parapertussis*$_{hu}$) and the other observed only within sheep (*B. parapertussis*$_{ov}$) populations, often associated with pneumonia (Porter, J.F et al., 1994).

Similarly, *B. bronchiseptica* can cause lethal pneumonia, but the vast majority of *B. bronchiseptica* infections result in little if any noticeable disease. Once acquired, the nasal cavity generally remains colonized by *B. bronchiseptica* for the life of the animal (Harvill ET et al., 1999). Although not considered a human pathogen, *B. bronchiseptica* is frequently associated with respiratory

disease in domesticated and some wild animals and has been recovered from humans on a regular basis, suggesting that there is frequent spillover or zoonotic transmission into humans (Register KB and Harvill ET, 2010). In Massachusetts, the only state currently performing culture-based techniques that are able to distinguish among the three classical *Bordetella*, 9 cases of human-associated *B. bronchiseptica* over 17 years (between 1991 and 2008) have been detected, confirming the zoonotic transmission from infected animals to humans is occurring (Jennie Lavine, unpublished data). Each of these "spillover" events provides an opportunity for the novel adaptation of an animal pathogen to humans and, therefore, the emergence of new human infectious diseases (Daszak P et al., 2000). An estimated 75% of emerging diseases in humans arise from zoonotic sources, either through direct transmission, such as SARS and West Nile virus, or by mutations that result in a "species jump", such as HIV, highly pathogenic avian influenza H5N1, and the most recent pandemic of novel H1N1 influenza (Taylor LH et al., 2001). Domesticated animal populations throughout the world serve as an underappreciated and overlooked source of emerging zoonotic diseases and serve as an invaluable resource for investigative studies specifically addressing pathogen evolution. *B. bronchiseptica,* which colonizes a broad range of animal populations and humans, serves as an excellent model system to investigate zoonotic transmission. Importantly, our recently published data (supported by unpublished data below) provide evidence that there is horizontal gene transfer amongst the Bordetellae and/or with other organisms in their environment. The frequent spillover of *B. bronchiseptica* from animal sources to humans provides a source of additional genes that may be acquired by *B. pertussis*, including important virulence factors that we and others have characterized in *B. bronchiseptica*, but are missing from the genome of *B. pertussis*. Thus these spillover events are a source of concern both as potential zoonotic sources of new disease, or as a source of genes that could exacerbate *B. pertussis* epidemiology and/or disease.

Earlier work has shown that *B. pertussis* and *B. parapertussis* appear to have evolved from *B. bronchiseptica*-like progenitors primarily by genome reduction. On the basis of the currently used Multi-Locus Sequence Typing (MLST) scheme, in which partial sequences of seven conserved housekeeping genes are compared, *B. pertussis* and *B. parapertussis*$_{hu}$ are distinguished by only one or two single nucleotide polymorphisms (SNPs)(Diavatopoulos DA et al., 2005). Amongst the surprising findings from our analysis of a second *B. pertussis* genome is the observation that the rearrangements that distinguish *B. pertussis* from *B. bronchiseptica* appear to have happened very recently, and that they also distinguish one strain of *B. pertussis* from another. In fact there are approximately as many rearrangements distinguishing the two *B. pertussis* strains from each other as there are distinguishing either from *B. bronchiseptica*. There have been multiple recent publications describing *B. pertussis* strains that lack expression of, or genes encoding, a growing list of the most important known virulence factors. Thus, rather than being well-adapted to the human host, and therefore relatively genetically stable, *B. pertussis* appears to be evolving and adapting on an ongoing basis at a rate that is determined by mutations as well as by an elevated rate of recombination.

B. bronchiseptica lineages are more diverse and have been divided into two complexes (Complex I and IV), with Complex IV being more closely related to *B. pertussis* and containing a higher proportion of human isolates. The distribution of human isolates across the *B. bronchiseptica* phylogenetic tree suggests that many, perhaps all, lineages are able to infect and cause disease in humans. If all *B. bronchiseptica* strains have the necessary virulence factors that *B. pertussis* utilizes to cause whooping cough, as is true of strain RB50, then it is simply a matter of time before some new *B. bronchiseptica* lineage expressing that combination spills over into human populations from its natural animal source. Conversely, if many *B. bronchiseptica* lineages lack one or more key virulence factors then the chance of a new epidemic of zoonotic origin is much smaller. Thus, to understand the risk of some lineage of *B. bronchiseptica* emerging as a rapidly transmitted human pathogen we must know the genetic repertoire of the various lineages.

We also observed that lineages across the MLST-based tree differ in some of the most interesting characteristics of bacterial pathogens, including their apparent host range, level of virulence, and ability to cause short-term (acute) or life-long persistent infections. The especially powerful *Bordetella* animal infection model allows the association of striking *in vivo* pathogenesis phenotypes with specific genomic

changes (Forward Genetics), as we have demonstrated recently (Buboltz AM et al., 2008 and Buboltz AM et al., 2009). Functional genetic tools also allow for the manipulation of specific genes to define their contributions to phenotypic differences (Reverse Genetics). Thus each new genome provides a set of genes that can be correlated to the phenotypes of that strain to identify candidate genes involved, and the genes can be modified to prove the relationship to the phenotype.

The reemergence of pertussis in developed countries including US, the interesting phylogenetic relationship between the classical *Bordetella* species and amongst *B. bronchiseptica* lineages, the remarkable ability of these organisms to infect a large and diverse set of hosts, and the continual spillover of many of these lineages to diverse hosts including humans are compelling reasons why the addition of more genomic sequence information about the various lineages will be valuable. In addition, the recent recognition that a substantial proportion of whooping cough in the USA is caused by *B. parapertussis*, as well as our recently published evidence that the new acellular vaccines actually increase *B. parapertussis* colonization levels (Long GH et al, 2010), are compelling reasons to explore this lineage in particular. The available animal model systems and genetic tools combined with the genomic sequences of a diverse set of strains that we propose to sequence in this proposal will create a rich opportunity for unique discoveries to reveal the basis for emergence of zoonotic disease. Cumulatively, *Bordetella* species represent a unique paradigm in the study of bacterial genome evolution in that host restriction or adaption of *B. pertussis* and *B. parapertussis* appears to be driven by gene loss rather than gene acquisition (Parkhill J et al., 2003). However, the very different sets of genes lost by the two organisms, which cause the same disease in the same host (whooping cough in humans), suggests a more complex evolutionary process.

In order to understand the speciation and evolution of the *Bordetellae*, and the adaptation of multiple lineages to infect humans, we propose to sequence several strains of *B. bronchiseptica*, *B. parapertussis* and *B. pertussis*.

II.) Background and Rationale

A.) Current Sequencing and Analysis of the Classical *Bordetella* Genomes

In 2003, a representative genome sequence from each of the classical *Bordetella* species was published (Parkhill et al., 2003). Of these three, the *B. bronchiseptica* strain RB50 has the largest genome (5,339,179bp; 5007 predicted genes), while *B. parapertussis* 12822 (4,773,551bp; 4,404 genes) and *B. pertussis* Tohama I (4,086,189bp; 3,816 genes) have smaller genomes, apparently due to large-scale gene loss. Genes that are either lost or inactivated in the two important human pathogens (*B. pertussis* and *B. parapertussis*) include many predicted to be involved in membrane transport, small-molecule metabolism, regulation of gene expression, and synthesis of surface structures. It was argued that RB50 does not have the insertion sequences (IS) associated with gene loss in the other two, and appears to have a largely intact genome most similar to the progenitor of all three organisms. However, our more recent sequencing of several additional *B. bronchiseptica* strains reveal that each has a large number of novel genes not present in RB50, and that sets of genes have been transferred between *B. bronchiseptica* lineages and/or acquired from exogenous sources.

In 2009, we helped select seven more *Bordetella* genomes for sequencing: *B. bronchiseptica* strains 253, 1289, MO149, R77, and D445, *B. parapertussis* ovine strain BPP5, and *B. pertussis* strain 18323. These strains were selected based on their MLST, host specificity, and virulence phenotypes (Table 1). In collaboration with the Sanger Institute and other prominent investigators, we are leading the effort to assemble, annotate, and compare these genomes, which is approaching completion. A first paper is under preparation to describe the comparisons amongst these genomes. Of particular interest is the difference between the rates of point mutation and rates of recombination along different lineages, and how that might contribute to the evolution and the regulation of virulence factors toward very different host interactions. The annotation information of these strains are shown in Table 2.

Gene Bank Acc. Num.	Isolate	ST	Host	Species	Status	Note
BX470250	RB50	12	Rabbit	*B. bronchiseptica*	Finished	Parkhill et. al, 2003
BX470248	Tohama I	1	Human	*B. pertussis*	Finished	Parkhill et. al, 2003
BX470249	12822	19	Human	*B. parapertussis*	Finished	Parkhill et. al, 2003
	18323	24	Human	*B. pertussis*	Finished	Highly virulent in the mouse intracerebral test
	253	27	Dog	*B. bronchiseptica*	Draft	Hypovirulent compared to RB50, loss of *cya* locus
	1289	32	Monkey	*B. bronchiseptica*	Draft	Hypervirulent compared to RB50
	MO149	15	Human	*B. bronchiseptica*	Finished	Human *B. bronchiseptica* isolate
	Bpp5	16	Sheep	*B. parapertussis*	Finished	Does not colonize the mouse respiratory tract unlike others
	D445	17	Human	*B. bronchiseptica*	Draft	Human *B. bronchiseptica* isolate
	R77	18	Human	*B. bronchiseptica*	Draft	Human *B. bronchiseptica* isolate

Table 1. Summary of strain information of sequenced classical *Bordetella* strains.
NCBI ascension number, isolate name, sequence type based on MLST, species name, genome status, and publications associated with sequencing or strain isolation or virulence phenotypes are summarized in this table. The first three genomes were published in 2003, and the remaining seven are still currently under analysis.

Species	*B. bronchiseptica*						*B. parapertussis*			*B. pertussis*	
Strain Name	RB50	253	1289	MO149	R77	D445	12822	Bpp5	Bpp5 plasmid	Tohama1	18323
Contig (C) /Scaffold (S) Number	1	4 C	1	1	16 S	11 S	1	1	1	1	1
Genome Size (bp)	5,339,179	5,264,383	5,208,522	5,091,817	5,115,717	5,243,194	4,773,551	4,887,379	12,195	4,086,189	4,043,846
Gaps (Ns)	0	2	628	0	146,101	251,924	0	0	0	0	0
G+C content (%)	68.49	68.64	68.65	68.86	68.51	68.23	68.43	68.15	61.39	68.12	68.11
# of predicted ORFs	5009	4845	4785	4669	4667	4775	4402	4558	15	3806	3746
Pseudogenes	12	60	39	44	76	256	217	389	0	359	369
ORFs with Ns	NA	NA	NA	NA	320	473	NA	NA	NA	NA	NA
Average gene size (bp)	983	989	1000	1002	974	953	1000	986	534	983	986
All coding sequences (%)	92.23	91.04	91.90	91.88	91.51	91.21	92.23	92.16	65.6	91.62	91.38
rRNA operons	3	3	3	3	3	1?	3	3		3	3
tRNA	55	54	54	54	54	54	54	54		51	51
IS481	0	0	0	0	0	0	0	0		230	239
IS1001	0	0	0	0	0	0	22	27		0	0
IS1002	0	0	0	0	0	0	9	0		5	7
IS1663	0	0	0	7	0	9	0	0		17	18

Table 2. Summary of annotation information of sequenced classical *Bordetella* strains.
Strain name, number of contigs or scaffolds, genome size, number of gaps, G+C content, number of predicted open reading frames (ORFs), number of pseudogenes, number of ORFs with Ns, average gene size, % of all coding sequences, number of rRNA operons, tRNAs, and IS elements are summarized in this table. The published genomes are highlighted with dark blue.

Based on these new genomes we have begun to calculate the pan-genome (the full complement of genes), the core genome (genes that are present in all strains), the dispensable genome (genes present in at least two strains, but not all strains), and unique genes (specific to a single strain) (Medini D et al., 2008). Initial comparisons suggest the pan-genome may be much larger than initially estimated. These data suggest that a more accurate pan-genome, representing the true diversity of *Bordetella* species, could be established with additional genome sequences (Medini D et al., 2005).

B.) Phylogenetic Analysis of Classical *Bordetella* Lineages

The classical bordetellae are closely related to each other based on DNA hybridization studies and metabolic characteristics (Preston A et al., 2004). A few single nucleotide polymorphisms (SNPs) within nearly 3,000bp of seven housekeeping genes were previously used to create a minimum spanning tree (Diavatopolous DA et al., 2005). Recently, we expanded this tree to include 26 new sequence types (Figure 1). We find clusters of *B. pertussis* and *B. parapertussis* interspersed amongst numerous isolates

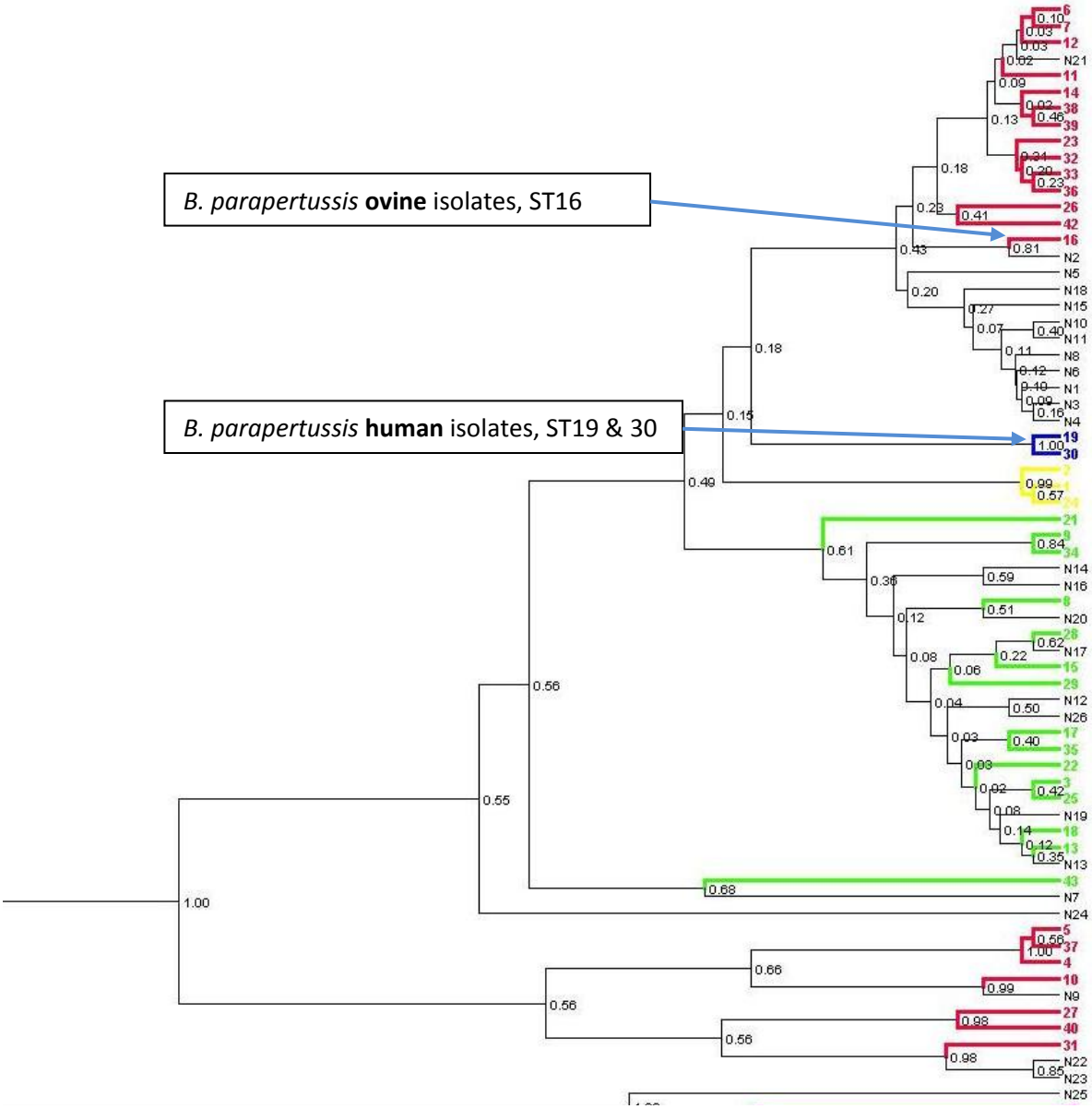

Figure 1. Phylogenetic tree of MLST housekeeping genes.
This tree was constructed by the UPGMA method with concatenated partial nucleotide sequences of 7 MLST housekeeping genes (*adk, glyA, icd, pgm, fumC, tyrB* and *pepA*). Each number represents the Sequence Type (ST) based on MLST typing and new ST has N in front of the number. Branches and STs are colored based on the conventional complex I (red; *B. bronchiseptica*), II (yellow; *B. pertussis*), III (blue; *B. parapertussis*), and IV (green; *B. bronchiseptica*) definition (Diavatopoulos, DA et al., 2005). *B. bronchiseptica* that do not cluster with known complexes (hot pink) and new sequence types (no color) are also identified.

of *B. bronchiseptica*. Interestingly, *B. parapertussis* ovine strain (ST16) is not clustered with *B.* parapertussis human isolates (ST19), from which it gets its name. Using the whole genome sequences of a single strain from each sequence type will allow us to overcome the limitations of partial sequences and generate a more robust phylogenetic tree, indicative of the strongest relationships between the classical bordetellae.

C.) Genome-wide Study of Single Nucleotide Polymorphism

As described above, the MLST-based phylogeny discriminated strains with only a few SNPs present in 7 housekeeping genes. More recently, we address phylogenetic relationships between *Bordetella* species utilizing a whole genome SNP analysis approach (Harris S et al., 2010). Interestingly, these two different analyses yield vastly different results. For example, whole genome SNP analysis has shown that *B. parapertussis* ovine (labeled as BPP5) and human isolates (labeled as BPP-12822) are more closely related to each other, a completely different result from the MLST-based analysis (Figure. 2). Although more powerful than the MLST approach, the whole genome SNP analysis is based on comparison to a single reference genome, in this case that of RB50; whole genome sequencing would allow us to replace the reference genome with the pan-genome, resulting in an even more representative tree. Understanding the absolute relationships between these species will require whole genome sequencing.

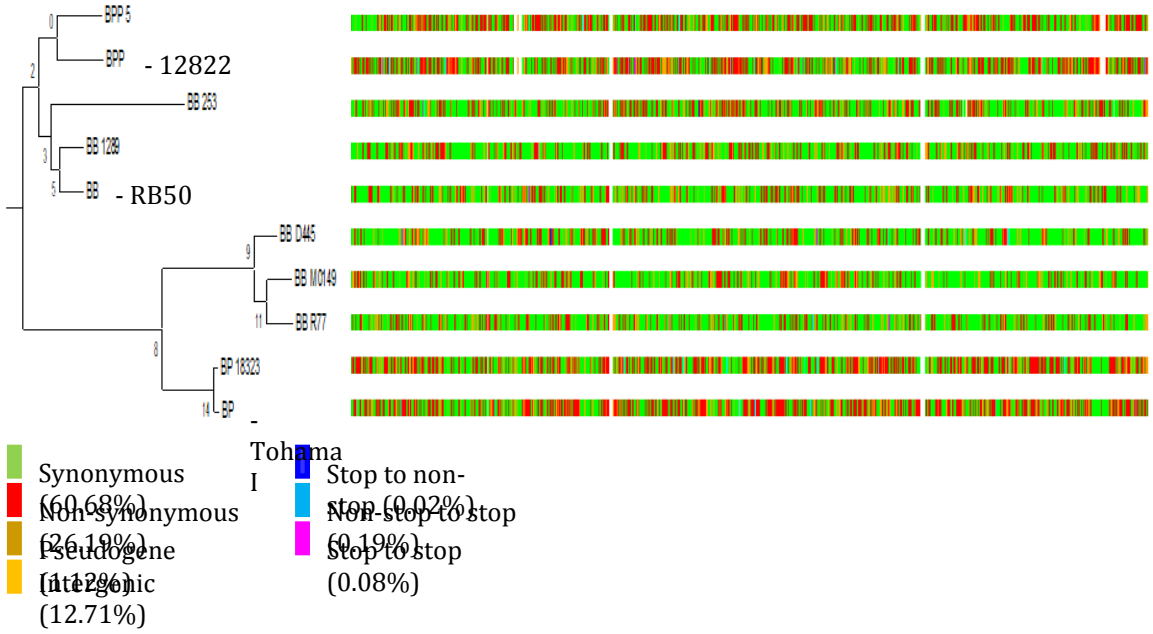

Figure 2. Maximum likelihood phylogenetic tree based on genome-wide SNPs
The whole genome sequence is shredded into short reads and remapped to the reference genome (*B. bronchiseptica* RB50), and candidate SNPs were identified. The phylogenetic tree was estimated with RAxML for all SNP sites in the genomes. Different kinds of SNPs have been color-coded as shown.

D.) Observed heterogeneity amongst various *Bordetella* strains

In addition to the various identified MLST types, we have documented that *B. bronchiseptica* strains are heterogeneous in various phenotypes, including pertactin type, O-antigen type, virulence and host range. Utilizing the previously sequenced genomes as a starting point, we have begun several studies to examine species diversity in various characteristics relevant to pathogenesis. In the process we have

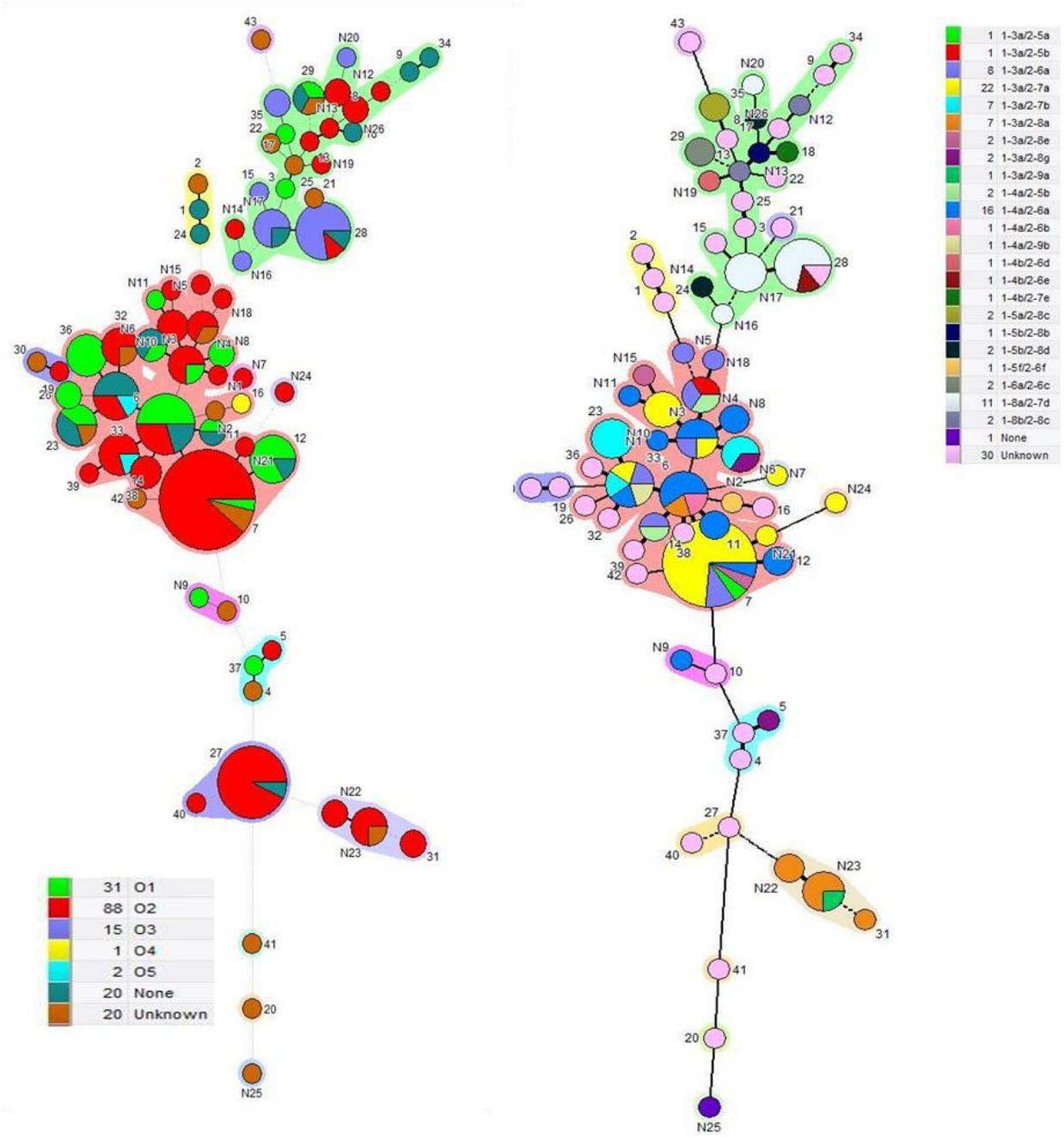

Figure 3. Minimum spanning tree of MLST housekeeping genes with *wbm* locus or *prn* repeats associated with sequence type.
This tree was constructed by the minimum spanning method with concatenated partial nucleotide sequences of 7 MLST housekeeping genes (*adk, glyA, icd, pgm, fumC, typrB,* and *pepA*). Each circle represents a sequence type (ST) the size of which is related to the number of isolates within that particular ST. Colors within circles indicate *wbm* locus (A) or *prn* repeats (B) distribution among *Bordetella* isolates.

identified evidence of transfer of a genetic locus involved in assembly of the major protective antigen, O-antigen, as described below.

Previously, it was shown that *B. bronchiseptica* and *B. parapertussis* evade complement-mediated killing and immunity induced by *B. pertussis* through production of an O-antigen, a large repeated glycan unit that attaches to the lipopolysaccharide on the outer membrane of Gram-negative bacteria (Gobel, 2008). Antigenically distinct types of O-antigen-encoding loci were identified among circulating *B. bronchiseptica* strains of the same ST, providing evidence for HGT events within the classical *Bordetella* (Figure 3A). This study found *B. bronchiseptica* O-antigen types 1 and 2 are distributed throughout the phylogenetic tree, while a recently discovered O-antigen type (O3) is only observed in *B. bronchiseptica* strains of Complex IV. Several strains lacking detectable O-antigen (labeled as None), and others having new antigenically distinct O-antigen types (O4 and O5), were identified across the phylogenetic tree as well (Figure 3A). This proposed genome sequencing will be able to identify other genes that are possibly horizontally transferred from a source outside the genus or within the genus, and to determine how these HGTs contribute to the evolution of these organisms.

A similar analysis comparing the local heterogeneity in the repeat region of the gene encoding pertactin (*prn*), a prominent vaccine antigen believed to be under positive selection for variation, was completed. This again illustrated the broad variation within these *Bordetella* strains (Figure 3B). Although this study provided no evidence for HGT, it nonetheless revealed considerable diversity within pertactin repeat regions; a total of 23 different alleles were identified. Among the three most prevalent genotypes (1-3a/2-7a; 1-4a/2-6a; 1-8a/2-7d), the first two were found only in *B. bronchiseptica* Complex I, while the third was only identified within Complex IV. Interestingly, a newly identified ST (N25) appears to lack the pertactin gene all together. Both O-antigen and pertactin studies provide evidence of differential selection pressures amongst various *Bordetella* lineages and emphasize the power of comparative sequencing to reveal evolutionary pressures and events.

The proposed sequencing of this diverse set of strains will identify possible HGT events, define the variation amongst virulence factors, and identify genes under positive and negative selection. Knowledge of the genome diversity amongst such a vast set of strains will permit a better understanding of the environmental pressures in the different ecological niches and hosts, and will shed light on the basis for some of the most interesting attributes of bacterial pathogens: host range, virulence, types of pathology induced and establishment of acute versus chronic/persistent infections.

This evidence of gene transfer contradicts the simple assumption of a "closed genome", and raises the possibility that multiple other genetic loci may have been acquired by individual lineages of bordetellae.

E.) Evidence for an Open Genome

Heaps' law defines that the number n of distinct attributes grows as the number N of entities increases with a sub-linear power law ($n \sim N^{\gamma}$ with $0 < \gamma < 1$). Tettelin et al. have used this law in the context of genomes and genes (Tettelin et al., 2008). As the number of sequenced genomes increases, discovering a new gene becomes harder. A power law ($N^{(\gamma-1)} = N^{-\alpha}$ with $\alpha = 1 - \gamma$) is a good candidate for this general observation. For $\alpha > 1$ or $\gamma < 0$, the size of the pan-genome approaches a constant as more genomes are sequenced, that is the definition of a closed pan-genome. For $\alpha \leq 1$ or $0 \leq \gamma < 1$, the size of the pan-genome increases as the number of sequenced genomes increases, that is the definition of an open pan-genome. Tettelin et al. have predicted that *B. anthracis* has a closed pan-genome based on 5 available genomes ($\alpha = 5.6 > 1$) for this species; we have used 5 *B. bronchiseptica* strains (253, 1289, D445, R77, and MO149) that were sequenced and annotated at Sanger and Penn State as well as published *B. bronchiseptica* RB50 to predict whether the pan-genome of *B. bronchiseptica* is open or closed. As it is shown in Figure 1, the pan-genome of *B. bronchiseptica* seems to be open with $\gamma = 0.0733 \pm 0.0008 > 0$. We also have plotted the number of new genes that are discovered for additional genomes being sequenced (data now shown). By power law regression, *B. bronchiseptica* appears to have an open pan-

genome with $\alpha = 0.935 \pm 0.036 < 1$, as *Prochlorococcus marinus* has been predicted to have an open pan-genome with $\alpha = 0.80 < 1$ (Tettelin et al., 2008). This preliminary data suggests that the pan-genome of *B. bronchiseptica* appears to be open; thus, sequencing additional **strains** of *B. bronchiseptica* would be valuable to identify core and pan-genome and to understand the diversity of *B. bronchiseptica* genomic characters. The identification of such genes will provide a genetic basis for the lineage-associated phenotypes.

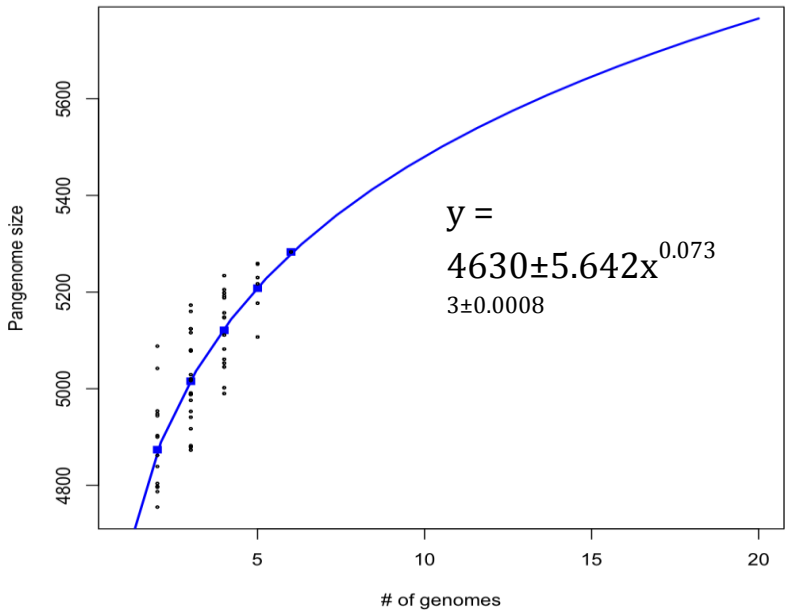

Figure 4. The pan-genome of *B. bronchiseptica* using means and a power law fit. The total number of genes found is shown for increasing values of the number of *B. bronchiseptica* genomes sequenced. The exponent $\gamma = 0.0733 > 0$ indicates that *B. bronchiseptica* is an open pan-genome species.

F.) Animal Models of *B. bronchiseptica* Infection Used in Forward Genetics

The study of most human pathogens suffers from the lack of a human infection model or from alternative models that accurately mirror events in the human host. For example, when pathogens require large numbers to colonize or do not grow within an animal host, it is difficult to determine with confidence that the infectious process is accurately modeled. There are various ways to estimate how closely an experimental infection mimics natural infection; one measure is the dose necessary to colonize a host. The infectious dose of many pathogens in the mouse model is measured in orders of magnitude, requiring an animal to ingest 10^3, 10^6 or even 10^9 CFU to initiate colonization. In contrast, the infectious dose of the prototype strain of *B. bronchiseptica* (RB50) in mice, as well as rats and rabbits, is less than 10 CFU (Harvill ET et al., 1999). *B. bronchiseptica* can be delivered in a droplet of PBS or water touched to the end of the noses of mice and persists for life in every animal inoculated (Harvill ET et al., 1999). In these and many other measures, *B. bronchiseptica* is exceptional in how efficiently it mimics a natural infection in laboratory animals. This allows the many aspects of the infectious process to be probed with the combined tools of bacterial genetics and mouse molecular immunology. We have used this approach to first identify roles for individual bacterial factors and then define the specific host immune functions with which they interact *in vivo*. Here, we propose to use genomic analysis results

with this model to better understand the genes involved in the most important and relevant aspects of bacterial pathogenesis.

We have demonstrated that specific *in vivo* phenotypes can be mapped onto the phylogenetic tree (Buboltz AM et al., 2008 and Buboltz AM etlal., 2009) and further shown that we can use sequence data to associate evolutionary events (e.g. deletion of toxin genes) with a specific *in vivo* phenotype (e.g. hypovirulence; Buboltz AM et al., 2008). Here, we propose to extend this approach to strains across the phylogenetic tree, evaluating the association between a number of measurable *in vivo* phenotypes and changes in the genome, which will lead to identify the specific sets of genes that are biologically significant. One major problem with the common approach to comparative genomics is that even if it identifies novel genes represented in different strains, there is often little clue as to the phenotype those genes might affect. The lack of appropriate infection models for many pathogens greatly compounds the problem, making it even harder for the researchers to connect the genomic differences to the phenotypes. The advantage of our approach is that it starts with a relevant and important *in vivo* phenotypes and uses newly sequenced genomes to identify the genes associated with that phenotype. It can then be a straight-forward application of reverse genetics (gene mutation) to define the role of each associated gene in the phenotype with a good infection mouse model. This combined approach of forward genetics with sequencing can do what no amount of sequencing alone can do: identify possible functions for genes.

At the time strains were selected for the current (nearly complete) sequencing effort, we had phenotypic data on two of these strains, identified as hyper-virulent and hypo-virulent in the mouse model. The sequence analysis of these two strains contributed directly to our two papers describing the very different characteristics of each, such as loss of functional genes. It also led us to investigate transcriptomic differences that are involved in those different phenotypes (Buboltz AM et al., 2008 and Buboltz AM et al., 2009). Since that time we have identified important phenotypes in the infection model for many strains that we propose to sequence in this proposal. Importantly, the strengths of this model, including the very high efficiency of infection, persistence for life and the demonstrated roles of many different *B. bronchiseptica* immunomodulatory factors, create a unique opportunity to define novel roles for bacterial genes involved in the infection process.

III.) Research Plan

Project 1: Sequence 120 Classical *Bordetella* Strains Preselected for Diversity.

1.) Strain Selection and Prioritize the sequencing bins

We have a network of collaborators and continue to collect additional strains of various Bordetella species. Combining our collection of strains with those of Dr. Register, Dr. Mooi, and Dr. Musser, we have access to the full diversity of the four largest classical *Bordetella* libraries in the world. The set of 120 strains we have chosen includes isolates from each major continent and from a broad range of host animals and disease manifestations. A brief description of the libraries from which these strains were selected follows:

Dr. Karen Register (USDA) has collected over 400 strains of *B. bronchiseptica* from a wide range of hosts and geographic locations. She has typed all of these strains based on pertactin type, identifying over 100 types. Based on her observed diversity, we chose one strain from each pertactin type she had identified to perform Multi-Locus Sequence Typing (MLST) in the lab of Dr. Harvill. The strains represented many of the known Sequence types and identified 26 new Sequence Types, confirming their diversity.

Another collaborator, Dr. James Musser, has a large strain collection containing isolates from a very wide range of different animals and different locations and times. This collection was analyzed by Multi-Locus Enzyme Electrophoresis (MLEE) typing, identifying 38 MLEE types.

The Harvill lab has collaborated with the Collaborative Pulmonary Critical Care Research Network (CPCCRN) in the analysis of recent *B. pertussis* isolates from severely ill babies. We have access to the strains that CPCCRN collected and could be shipped to us via FedEx. From this source we have 19 strains from the recent (2010-2011) California outbreak, 6 of which are from children with critical pertussis, one of who died. We have access to extensive patient information on most of these patients. We also have worked with Maria Tondella, at the CDC, who has offered to provide additional strains. We include *B. pertussis* strains in the prioritized list of strains to be sequenced below. Importantly, our current collaboration with the Sanger Institute to sequence a larger number of *B. pertussis* isolates does not contain any from the recent California outbreak or from the very recent critical/fatal pediatric infections collected by CPCCRN. In fact, recent US infections are very poorly represented in that set of strains. Multiple groups have reported that strains are actively changing over time, possibly in response to vaccination-induced immune pressures, and vary substantially from one country to another. Therefore, an important outcome of this project will be improved understanding of the strain(s) currently causing this major outbreak in California, and those causing the most severe and lethal disease in newborns in the USA. The Harvill lab also maintains extensive interactions with the Massachusetts Health Department, which monitors more closely for *Bordetella* infections than any other state. We have access to the thousands of strains of *B. pertussis* they have collected. In addition, we have collaborated with them to publish several work shown that *B. parapertussis* is a substantial contributor to disease in this state, and therefore probably across the US. From their collection we will chose four geographically and temporally distributed isolates of *B. parapertussis*.

In making our selections, we used four major criteria to select representative isolates from these collections for sequencing: (I) isolates recovered from a wide range of animals (land and sea mammals representing wild and domesticated populations as well as avian species, e.g. dogs, pigs, turkeys, and humans) (II) well-characterized strains with defined *in vitro* (*prn* types and O-antigen types) and *in vivo* (LD50 and ID50) phenotypes, (III) one isolate representing each MLST (Table 2) and MLEE (Table 3) type and (IV) relevant to recent and severe disease in the US, including *B. pertussis* isolates from the recent outbreak in California, *B. parapertussis* isolates from Massachusetts and recent fatal pediatric infections. Using these criteria, we propose here to sequence a total of 120 strains. The full list of the 120 strains is present in Table 3 and Table 4.

We propose to sequence the complete collection of *Bordetella* with several levels of priorities. Our first priority is to sequence 5 strains of *B. pertussis* recently isolated from severe/lethal disease during the 2010 California epidemic, together with 2 strains from each of the six non-classical species of *Bordetella* as well as the most distantly related MLST-based Sequence Types (ST 20, 41, and N25). After that, 20 strains representing the most distantly related MLST-based Sequence Types, which are ST N23, 31, 10, 21, 43, N4, 6, 42, 26, 39, 23, 14, 11, 9, 28, 8, 36, 37, N24, and 35. After that, we will further sequencing 20 strains that are secondly most distantly related and so on. The priority bins are summarized in Table 5.

ST	ET	Isolates	Host	O-Antigen	Prn	Complex	Location	Year
1	37	TohamaI	Human	None	Unknown	2	Japan	1952
1								
2								
3		446	Human	1	Unknown	4	USA	Unknown
4								
5		Mbord782	Cat	2	1-3a / 2-8g	1	Netherlands	Unknown
6	1	Mbord674	Guinea Pig	1	1-4a / 2-6b	1	Germany	Unknown
7		FosterP1	Pig	2	1-3a / 2-7a	1	Scotland	Unknown
8		310	Human	2	Unknown	4	USA	Unknown
9		309	Human	None	Unknown	4	USA	Unknown
10								
11		M11/06/1	Seal	Unknown	1-4a / 2-6a	1	England	2006
12		RB50	Rabbit	1	1-4a / 2-6a	1	USA	Unknown
12		761	Human	1	Unknown	1	USA	2001
13		7E71	Horse	2	1-8b / 2-8c	4	Oklahoma, USA	Unknown
14		C4	Rabbit	2	Unknown	1	Pennsylvania, USA	2006

15		444	Human	3	Unknown	4	USA	Unknown
15								
16		207	Sheep	4	Unknown	1	NewZealand	Unknown
16								
17		445	Human	1	Unknown	4	USA	Unknown
17								
18		R77	Human	Unknown	Unknown		Unknown	
18	14	Mbord675	Human	Unknown	1-4b / 2-7e	4	Germany	Unknown
19		12822	human	2	Unknown	3	Germany	Unknown
19								
20								
21		345	Human	Unknown	Unknown	4	USA	1996
22								
23		448	Human	1	Unknown	1	USA	Unknown
24	38	18323	Human	None	Unknown	2	USA	1947
24								
25								
26		2115	Human	1	Unknown	1	Netherlands	2000
27		253	Dog	2	Unknown	1	USA	Unknown
27		752	Human	2	Unknown	1	USA	1998
28		978	Unknown	None	Unknown	4	Unknown	Unknown
29	3	Mbord670	Guinea Pig	1	1-6a / 2-6c	4	UnitedStates	Unknown
30								
31	8	Mbord732	Dog	2	1-3a / 2-8a	1	Denmark	Unknown
32		1289	Monkey	2	Unknown	1	South America	Unknown
32		S308	Squirrel Monkey	2	Unknown	1	USA	2002
33		M105/00/1	Seal	5	1-4a / 2-6a	1	CaspianSea	2000
34		758	Human	None	Unknown	4	USA	2000
35		OSU553	Turkey	3	1-5a / 2-8c	4	Ohio, USA	1983
36		804	GuineaPig	1	Unknown	1	Unknown	Unknown
37		980	Unknown	1	Unknown	1	Unknown	Unknown
38		C3	Rabbit	2	Unknown	1	Pennsylvania, USA	2006
39		C1	Rabbit	2	Unknown	1	Pennsylvania, USA	2006
40		974	Unknown	2	Unknown	1	Unknown	Unknown
41								
42								
43								
N1		Care161	Pig	2	1-4a / 2-6a	1	South Dakota, USA	1996
N2	1	Mbord635	Cat	2	1-5f / 2-6f	1	USA	Unknown
N3	16	Mbord626	Leopard	1	1-3a / 2-6a	1	USA	Unknown
N4	1	Mbord681	Koala	2	1-4a / 2-5b	1	Australia	Unknown
N5	1	Mbord854	Guinea Pig	2	1-3a / 2-6a	1	Switzerland	Unknown
N6	16	Mbord628	Horse	1	1-3a / 2-7b	1	USA	Unknown
N7		ASI-I	Pig	2	1-3a / 2-7a	1	UK	Unknown
N8		Mbord959	Seal	1	1-4a / 2-6a	1	Denmark	Unknown
N9		3E44	Rabbit	1	1-4a / 2-6a	1	Oklahoma, USA	Unknown
N10	3	Mbord673	Guinea Pig	2	1-3a / 2-7a	1	Germany	Unknown
N11		RB630	Rabbit	1	1-4a / 2-6a	1	Hungary	1987
N12		Mbord901	Turkey	2	1-8b / 2-8c	4	Germany	Unknown
N13		F4563	Human	2	1-5b / 2-8b	4	Louisiana, USA	2004
N14		ISU-CA 90 BB 02	Turkey	2	1-5b / 2-8d	4	California, USA	1990
N15		P50164	Pig	2	1-3a / 2-8e	1	Scotland	1999
N16		St.Louis	Human	3	1-8a / 2-7d	4	Missouri, USA	Unknown
N17		OSU O54	Turkey	3	1-8a / 2-7d	4	Minnesota, USA	1981
N18	1	Mbord762	Guinea Pig	2	1-3a / 2-6a	1	Ireland	Unknown
N19		F2	Turkey	2	1-4b / 2-6d	4	Iowa, USA	2004
N20		OSU O86	Turkey	3	1-8a / 2-7d	4	Iowa, USA	1982
N21		4609	Pig	2	1-3a / 2-7a	1	Iowa, USA	1960
N22	8	Mbord788	Dog	2	1-3a / 2-8a	1	Netherlands	Unknown
N23		Fosterα	Dog	2	1-3a / 2-9a	1	Scotland	1998
N24	1	Mbord745	Cat	2	1-3a / 2-7a	1	Denmark	Unknown
N25		99-R-0433	Human	3	no gene	Unknown	Massachusetts, USA	Unknown

N26	803	Guinea Pig	2	Unknown	Unknown	Unknown	Unknown

Table 3. Strains chosen from each known ST.

ST of MLST, ET of MLEE, isolate name, host, O-antigen type, *prn* type, complex, isolation location and isolation year are summarized in this table.

ET	Isolates	Host	Location
1	586	Pig	
2	ET2		
3	654	Pig	
4	595	Dog	USA
5	ET5		
6	590	Dog	
7	ET7		
8	732	Dog	
9	ET9		
10	ET10		
11	ET11		
12	ET12		
13	ET13		
14	591	Dog	United States
15	ET15		
16	730	Rabbit	
17	ET17		
18	ET18		
19	ET19		
20	ET20		
21	ET21		
22	NZ929	Sheep	New Zealand
23	SC2209	Sheep	Scotland
24	SC6	Sheep	Scotland
25	NZ585	Sheep	New Zealand
26	SC11	Sheep	Scotland
27	545	Pig	USSR
28	B24	Human	Unknown
29	707	Turkey	USA
30	705	Rabbit	USA
31	671	Rabbit	USA
32	824	Rabbit	Switzerland
33	902	Monkey	Germany
34	731	Horse	Denmark
35	B6	Human	Netherlands
36	B3	Human	
37			
38			

Table 4. Strains chosen from each electrophoretic type.

ET designation from each MLEE, isolate name, host, and isolation location are summarized in this table.

Priority	# of Strains	Sequence Type	
1	20		5 *B. pertussis* 2010 Californic epidemic strains
			2 strains from each of the six non-classical species of Bordetellae
		ST	20, 41, N25
2	20	ST	6, 7, 8, 9, 10, 11, 14, 21, 23, 26
		ST	28, 31, 33, 35, 37, 42, 43, N4, N23, N24
3	20	ST	2, 3, 13, 22, 29, 30, 34, N1, N5, N7
		ST	N8, N9, N10, N11, N12, N15, N17, N18, N20, N21
4	20	ST	1, 4, 5, 12, 15, 25, 32, 36, 38, 39
		ST	40, N2, N3, N6, N13, N14, N16, N19, N22, N26
5	20	ST	16, 17, 18, 19, 24, 27, ET1, ET2, ET3, ET4
		ET	5, 9, 12, 16, 21, 22, 25, 33, 35, 36
6	20	ET	6, 7, 8, 10, 11, 13, 14, 15, 17, 18
		ET	19, 20, 23, 24, 26, 27, 28, 29, 30, 31

Table 5. Priority bins for sequencing

2.) DNA Extraction

Isolates will be grown to mid-log phase in Stainer Scholte (SS) broth at 37°C, and total genomic DNA will be extracted using a QIAamp DNA Extraction kit (Qiagen). Five µg of total DNA will be collected and resuspended for each strain, resulting in a final concentration of 500ng/µl. DNA will then be sent to JCVI for library preparation and sequencing.

3.) DNA Sequencing to be Completed by JCVI

Because the genome comparisons we propose will not require closing the genome, we should be able to accomplish all of these analyses with 20-fold coverage of the 120 genomes (expecting to add several more, as additional STs are discovered) that are approximately 5Mb in size. This effort will require approximately 12,000 Mega bases of sequence. With only 20-fold coverage, there will be some gaps in the genome, and it is possible that these could include an entire gene. For this reason, we would like to increase coverage where practical. If there are any gaps that interfere in the analyses below, we will use the high sequence homology and low rearrangement rate between these strains to apply a PCR-based approach to complete a key region or define a break point. This will allow our highly collaborative group to accomplish the following goals during the genome analysis portion of this proposal.

4.) Genome Annotation

Since the *Bordetellae* are very closely related to each other, we will use an automated annotation transfer tool to annotate most genes in the genomes. Novel genes that are not automatically transferred from the reference genomes are of particular interest, and will be predicted using Glimmer3 (Delcher AL et al., 2007) and curated with BLAST (Altschul SF et al., 1990) and FASTA (Pearson WR and Lipman DJ, 1988) results. We will also use other bioinformatics databases such as Pfam (Sonnhammer EL et al., 1997) and Prosite (De Castro E et al., 2006), TMHMM (Krogh A et al., 2001), SignalP (Dyrlov Bendtsen J et al., 2004), and ISFinder (Siguier P et al., 2005), to accurately annotate genes new to this genus. Manual curation will be done with these novel regions using Artemis (Rutherford K et al., 2000) and Artemis Comparison Tool (ACT) (Carver TJ, et al., 2005).

5.) Genome Analysis

Goal 1. Define Pan-Genome (Core Genome, Dispensable Genome, Unique Genes).

The sequenced genomes of several *Bordetella* strains provide an outstanding opportunity to define the pan-genome of the classical *Bordetella* and identify genes/genomic sequences unique to individual *Bordetella* lineages. We expect to be able to assemble a core genome of genes present in all classical *Bordetella*. Proper pan-genome identification requires selecting a diverse set of strains; a gene lost in one strain of one lineage automatically identifies that gene as a non-core gene; sequencing gaps will be treated as missing data and not counted. We are confident that we have selected strains that accurately represent the diversity of the classical *Bordetella*; strains from this large collection have been isolated from various host types on multiple continents over the past 100 years and exhibit varied disease severities and/or different *in vitro* phenotypes. Although a strain from each ST or electrophoretic type has been selected, each of these strains is also diverse in multiple other categories, such as host range, O antigen types, not limiting our definition of diversity to SNPs found in housekeeping genes.

In order to define the pan-genome, we will use the genome sequence generated by JCVI and the annotation done at Penn State to extract the coding sequences from all newly annotated genomes and determine orthologs with OrthoMCL (Li L et al., 2003; Chen F et al., 2006) that will use all-against-all reciprocal BLASTP search. Genes that are shared by all the strains will be identified as core genes, while genes that are shared by some but not all strains will be identified as dispensable genes. Genes that have no BLAST hit with any other protein queries will be identified as strain-specific genes. From these diverse isolates, numerous unique strain- and lineage-specific genes will be identified.

Definitions of core genes, dispensable genes, and unique genes will open numerous doors and lead to new areas of investigation. Defining the core genes of these diverse strains provides an opportunity to experimentally address how gene loss or inactivation shapes the genome structure and benefits these bacteria by expanding the phenotypic differences among them. For example, a comparison between the core genome of *B. bronchiseptica* strains that can survive outside of the host verses the core genome of *B. pertussis* strains, which have a closed life cycle and exist only within the host (Bjornstad ON and Harvill ET, 2005), should define the sets of genes required in these two distinct niches. Dispensable genes should give insight into lineage specificity, i.e. identifying genes that are required by Complex IV *B. bronchiseptica* strains to infect humans instead of other mammals. Additionally, unique genes are inherently interesting, providing insights into unique strain phenotypes, as was seen with the association of hypovirulence with the loss of adenlyate cyclase toxin (Buboltz, 2008). This analysis allows researchers to begin to make associations between genome content and phenotypic differences. Through these pan-genome analyses, we will not only be able to understand the diversity of the classical *Bordetella*, but also define gene sets associated with phenotypic differences such as host range, virulence, types of pathology and acute versus chronic/persistent infections.

Goal 2. Identify any Mobile Genetic Elements (MGEs; insertion sequences, plasmids or phage) in each strain.

Each of the first three classical *Bordetella* species genomes sequenced contains a different set of IS and phage (Parkhill et al., 2003). It is expected that there will be substantial variation in the presence of the MGEs between the strains that we are proposing to sequence. ISFinder will be used to detect the IS elements in the genomes (Siguier P et al., 2005). We will look for phage regions that typically contain an integrase and several phage-associated genes as well as the presence of a tRNA or a flanking direct repeat to confirm the phage regions (Xu and Gogarten, 2008). These elements are associated with the recombinations that are believed to cause the observable changes within the genomes of *B. pertussis* and *B. parapertussis*, but not *B. bronchiseptica* (Parkhill etl al., 2003). Thus, this analysis will contribute to an understanding of the niche adaptation of *Bordetella* species by inferring MGEs' association with genomic recombination in different lineages and their association with any horizontally transferred genes. Identifying MGEs in various lineages of classical *Bordetella* species will inform our views on how they may have contributed to genome rearrangements, gene exchange and evolution in these organisms.

Goal 3. Identify Single Nucleotide Polymorphisms (SNPs) and create a detailed SNP-based phylogenetic tree of all strains.

We will use JCVI-generated sequence in our whole genome SNP analysis, as shown in Figure 2, to compare these 120 newly sequenced genomes. The complete or draft genome sequences will be shredded and remapped onto the reference genome (pan-genome of the classical *Bordetella* that will be identified from Goal 1), and candidate SNPs will be identified based on alignments. The phylogenetic tree will be estimated with RAxML v7.0.4 for all SNP sites, using a General Time Reversible (GTR) model with a gamma correction for among site rate variation, and 100 random bootstrap replicates (Harris S et al., 2010). Importantly, because this effort will include nearly every SNP in the classical *Bordetella* pan-genome, we will construct the most accurate classical *Bordetella* phylogeny to date. We will identify all the SNPs present in each genome, examine their distribution across the genome and among various genes, and identify evidence for/against positive/negative selection by dN/dS analysis. dN/dS ratios of genes of interest will be calculated through pairwise comparisons to determine the average value for each gene by JCoDA (Steinway SN et al., 2010), which provides pairwise and site-based selection pressure on coding sequences. Studying the selection pressure, such as dN/dS ratio, on the *Bordetella* species core genome will be informative and provide new targets for investigating how nucleotide differences contribute to understanding the differences between host ranges, virulence phenotypes, and vaccine-induced evolution in the *Bordetella* strains. For example, recent research in the Netherlands and France suggests strong

selection pressures exist against *B. pertussis* strains expressing proteins present in current vaccines, such as pertussis toxin or different fimbrial serotypes. Our SNP analysis would allow us to identify other selection pressures throughout the entire genome, most likely in genes encoding proteins present in whole cell or acellular vaccines. As additional strains are sequenced in the future, these strains can easily be added to this phylogenetic analysis. SNP analysis will allow researchers to more accurately discriminate between lineages and more closely related strains. This analysis will also distinguish genes that are under positive and negative selection pressures. Importantly, we predict that different lineages having specialized ecological niches will vary in the subsets of genes that are under positive and negative selection.

Goal 4. Identify HGT events.

HGT is a process in which an organism may acquire new genes/functions from other organisms. To identify possible HGT events on a genome-wide basis we will use DarkHorse (Podell S and Gaasterland T, 2007), which is a bioinformatics tool for identification and ranking of phylogenetically atypical proteins. This will select potential ortholog matches from GenBank nr reference database and calculate a lineage probability index (LPI) score for each protein. As LPI scores are inversely proportional to the phylogenetic distance between database match sequences and the query genome, candidates with higher LPI scores are unlikely to have been horizontally transferred. Then, HGTs will be identified by comparing a gene or set of contiguous genes that show homology with an out-group lineage/organism (without any affinity) and have a GC content different from the average GC content (~68%) of *B. bronchiseptica*. Although there is no published evidence of HGT to classical *Bordetella* species from bacteria outside of the genus, there is some evidence of HGT among *Bordetella* species. For example, there is evidence of *B. holmesii*, a relatively distant species, gaining virulence genes from *B. pertussis* (Diavatopoulos DA et al., 2006). Additionally, we have observed evidence that O-antigen loci were horizontally transferred between *B. bronchiseptica* strains (Buboltz AM et al., 2009). Therefore, it is possible that some HGT event(s) may be observed within the different lineages of *Bordetella*. They are also likely to be associated with MGEs. Since HGT can allow for the very rapid adaptation of organisms to new niches and/or the appearance of new pathologies or antibiotic resistance, determining the presence or absence of gene transfer in the classical *Bordetella* will advance our understanding of the recent evolution and the likelihood of sudden changes in these characteristics in the near future.

Goal 5. Associate gene loss with changes in a long and growing list of virulence characteristics (Forward genetics).

Combining a relevant and efficient infection model with genomic data allows for a powerful forward genetic approach that can make genomic data more biologically meaningful and thus valuable. *In vivo* characteristics that are both easily measured and relevant to pathogenesis can be the starting place, allowing comparative genomics to then identify the genes involved. In our analysis of the biological characteristics of various lineages of the *B. bronchiseptica* tree, we have found that particular virulence characteristics are associated with discreet branches, which are often associated with specific genomic changes. For example, we identified specific *B. bronchiseptica* strains that are hyper-virulent or hypo-virulent in the mouse model (Buboltz AM et al., 2008 and 2009), and we used this information to guide the choice of strains to be sequenced by the Sanger Institute. We observed that strains of the same ST as a hypo-virulent strain all share a deletion of the prominent toxin, Adenylate Cyclase Toxin (Buboltz AM et al., 2008). Similarly, a hyper-virulence phenotype is shared by strains that over-express a Type III Secretion System that is known to contribute to virulence (Buboltz AM et al., 2009). Our ability to define specific *in vivo* phenotypes in the powerful *B. bronchiseptica* animal infection model gives meaning to the relatively small genomic differences between particular lineages. The proposed sequencing will provide data that will allow association between observable phenotypic differences and genotypic differences between lineages or species.

IV.) Summary

The classical *Bordetella* serve as an excellent model to study the genetic basis for evolution of some of the most interesting and important characteristics of bacterial pathogens: host range, virulence and duration (acute/chronic-lifelong). The singularly powerful animal model, and the substantial and robust differences between each lineage in that model, presents a unique system in which the specific genomic differences between closely related strains/lineages can be related to their genomic differences. In this system, genes can be related to real in vivo phenotypes in a natural infection setting, as we have shown previously with the small number of genomes sequenced to date (Buboltz AM, 2008, Buboltz AM 2009). The moderate rate of gene loss (less "reversible" than point mutations) amongst *B. bronchiseptica* strains can be used to define branches of the phylogenetic tree unambiguously, verifying the SNP-based phylogeny. Thus, the data generated by this proposed project provides a unique paradigm for investigating how gene gain/loss impacts bacterial genome evolution as these organisms shift between commensals of a range of animals to become important human pathogens.

V.) References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, "Basic local alignment search tool," *J Mol Biol*, 215:403-410 (1990).
2. Bjornstad ON and Harvill ET, "Evolution and emergence of *Bordetella* in humans," *Trends in Microbiology*, vol. 13: 355-359 (2005).
3. Buboltz AM, Nicholson TL, Parette MR, Hester SE, Parkhill J, Harvill ET, "Replacement of Adenylate Cyclase Toxin in a Lineage of *Bordetella bronchiseptica*," *Journal of Bacteriology*, 190:5502-5511 (2008).
4. Buboltz AM, Nicholson TL, Weyrich LS, and Harvill ET, "Role of the Type III Secretion System in a Hypervirulent Lineage of *Bordetella bronchiseptica", Infection and Immunity,* vol. 77:3969-3977 (2009).
5. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J," ACT: the Artemis comparison tool," *Bioinformatics*, 21:3422-3423 (2005).
6. Celentano LP, Massari M, Paramatti D, Salmaso S, Tozzi AE, and the EUVAC-NET Group, "Resurgence of pertussis is Europe," *Pediatr Infect Dis J*, vol. 24: 761-765 (2005).
7. Centers for Disease Control and Prevention, "Pertussis (Whooping Cough) – What You Need To Know", (2002).
8. Chen F, Mackey AJ, Stoeckert CJ, JR, Roos DS, "OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups," *Nucleic Acids Res*, 34:D363-D368 (2006).
9. Daszak P, Cunningham AA, and Hyatt AD, "Emerging Infectious Diseases of Wildlife- Threats to Biodiversity and Human Health," *Science*, vol. 287: 443-449 (2000).
10. De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N, "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and struactural residues in proteins," *Nucleic Acids Res*, 34:W362-365 (2006).
11. De Melker HE, Schellekens JFP, Neppelenbroek SE, Mooi FR, Rumke HC, and Conyn-van-spaendonck MAE, "Reemergence of pertussis in the highly vaccinated population of The Netherlands: Observations on surveillance data," *Emerging Infectious Diseases*, 6: 348-357 (2000).

12. Delcher AL, Bratke KA, Powers EC, and Salzberg SL, "Identifying bacterial genes and endosymbiont DNA with Glimmer," *Bioinformatics,* 23:673-679 (2007).
13. Diavatopoulos DA, Cummings CA, Schouls LM, Brinig MM, Relman DA, and Mooi FR, " Bordetella pertussis, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of B. bronchiseptica.", *PLoS pathogens,* (2005).
14. Diavatopoulos DA, Cummings CA, van der Heide HGJ, van Gent M, Liew S, Relman DA, and Mooi FR, "Characterization of a highly conserved island in the otherwise divergent *Bordetella holmesii* and *Bordetella pertussis* genomes," *Journal of Bacteriology* vol. 188: 8385-8394 (2006).
15. Dyrlov Bendtsen J, Nielsen Hm von Heijne G, Brunak S, "Improved prediction of signal peptides: SignalP 3.0.," *J Mol Biol*, 340:783-795 (2004).
16. Goodnow RA, "Biology of Bordetella bronchiseptica," *Microbiol. Rev.* vol. 44:722-738 (1980).
17. Harris S, Feil EJ, Holden MTG, Quail MA, et al., "Evolution of MRSA during hospital transmission and intercontinental spread", *Science*, vol. 327:469-474 (2010).
18. Harvill ET, Cotter PA, and Miller JF, "Pregenomic Comparative Analysis between *Bordetella bronchiseptica* RB50 and *Bordetella pertussis* Tohama I in Murine Models of Respiratory Tract Infection," *Infection and Immunity*, vol. 67: 6109-6118 (1999).
19. Heininger U, Stehr K, Schmitt-Grohe S, "Clinical characteristics of illness caused by *Bordetella parapertussis* compared with illness caused by *Bordetella pertussis*," *Pediatr Infect Dis J.*, 13: 306-309 (1994).
20. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J Mol Biol* 305:567-580 (2001).
21. Li L, Stoeckert CJ, and Roos DS, "OrthoMCL: Identification of ortholog groups for eukaryotic genomes," *Genome Res,* 13:2178-2189 (2003).
22. Long GH, Karanikas AT, Harvill ET, Read AF, Hudson PJ (2010) Acellular pertussis vaccination facilitates Bordetella parapertussis infection in a rodent model of bordetellosis. Proc Biol Sci. 2010 Jul 7;277(1690):2017-25.
23. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R, "The microbial pan-genome," *Current Opinion in Genetics & Development,* 15:589-594 (2005).
24. Parkhill J *et al.*, "Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*," *Nat Genet.,* vol. 35:32-40 (2003).
25. Pearson WR and Lipman DJ, "Improved tools for biological sequence comparison," *PNAS*, 85:2444-2448 (1988).
26. Podell S and Gaasterland T, "DarkHorse: a method for genome-wide prediction of horizontal gene transfer", *Genome Biology*, vol. 8: R16 (2007).
27. Porter JF, Connor K, and Donachie W, "Isolation and characteristion of *Bordetella parapertussis*-like bacteria from ovine lungs," *Microbiology*, 140: 255-261 (1994).
28. Preston A, Parkhill J, and Maskell DJ, "The bordetellae: lessons from genomics", *Nature Review*, vol. 2:379-390 (2004).
29. Register KB and Harvill ET. Bordetella. In Pathogenesis of Bacterial Infections in Animals, 4th ed. C.L. Gyles, J.F. Prescott, J.G. Songer and C. O Thoen, eds. Wiley-Blackwell, Ames, IA. p. 411-427 (2010).
30. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B, "Artemis: Sequence visualization and annotation," *Bioinformatics*, 16:944-945 (2000).
31. Shashidhar BY et al., "Serologic survey for *Bordetella bronchiseptica* in Nebraska specific-pathogen-free pigs," *Am J Vet Res* 44:1123-1125 (1983).
32. Siguier P, Perochon J, Lestrade L, Mahillon J, and Chandler M, "ISfinder: the reference centre for bacterial insertion sequences," *Nucleic Acids Res,* 34:D32-36 (2005).
33. Sonnhammer EL, Eddy SR, Durbin R, "Pfam: a comprehensive database of protein domain families based on seed alignments," *Proteins*, 28:405-420 (1997).

34. Steinway SN, Dannenfelser R, Laucisu CD, Hayws JE, Nayak S, "JCoDA: a tool for detecting evolutionary selection," *BMC Bioinformatics*, 11:284 (2010).
35. Taylor LH, Latham SM, and Woolhouse ME, "Risk factors for human disease emergence," *Philos Trans R Soc Lond B Biol Sci* vol. 356:983-989 (2001).
36. Van der Zee A, Mooi F, Embden JV, and Musser J, "Molecular Evolution and Host Adaptation of *Bordetella* spp.: Phylogenetic Analysis Using Multilocus Enzyme Electrophoresis and Typing with Three Insertion Sequences," *Journal of Bacteriology*, vol. 179: 6609-6617 (1997).
37. Watanabe M, Nagai M, "Whooping cough due to *Bordetella parapertussis*: an unresolved problem," *Expert Rev Anti Infect Ther.*, 2: 447-54 (2004).
38. World Health Organization, "Pertussis – the disease", (2008).
39. Wernli D, Emonet S, Schrenzel J, and Harbarth S, "Evaluation of eight cases of confirmed Bordetella bronchiseptica infection and colonization over a 15-year period," *Clin. Microbiol. Infect.* vol. 17:201–203 (2011).
40. Xu Y and Gogarten JP, "Computational methods for understanding bacterial and archaeal genomes", vol. 7 (2008).