F. PROJECT SUMMARY

Applicants must provide, in the space below, a 200-word summary in lay language, of their research proposal. It is critical that the summary be in lay language as it will be reviewed and evaluated, by two stakeholder members of Cystic Fibrosis Canada's Medical/Scientific Advisory Committee.

The immense complexity of microbial communities living in our lungs, throats, and sinuses presents a challenge to understanding declining lung function in patients with cystic fibrosis (CF). Novel ways to study these communities using new DNA sequencing technology have identified the relevant bacteria, but many questions remain about how each contributes to health and disease and how individual bacterial populations change over time during infection.

The bacterium *Haemophilus influenzae* is an important cause of disease in children with CF, and children are often infected with several types at once. We will analyze the DNA of this species and its close relatives by purifying it directly from patient samples collected from children with CF. To do this, we will exploit a peculiarity of *H. influenzae* cellular physiology, in which these bacteria actively take up DNA from their environment, but only from their own or closely related species. By purifying, sequencing, and analyzing these preferred DNA fragments from patient samples, we will track genetic changes in the species within individual patients over time. This study will pinpoint specific bacterial genes involved in adapting to new conditions, like changes in treatment and disease severity, and thus offer new insights into patient care.

If this application is for a research grant renewal, the applicant must provide a 200-word summary, in lay language, of progress achieved during the term of the previous grant. It is critical that the summary be in lay language as it will be reviewed and be evaluated, by two stakeholder members of Cystic Fibrosis Canada's Medical/Scientific Advisory Committee.

Not applicable.

G. PROGRESS REPORT

Applicants for renewal grants must provide details of progress achieved during the term of the previous grant. Productivity/accomplishments relating to specific aims outlined in the previous application should be addressed. One additional page may be used, and should be inserted following this page.

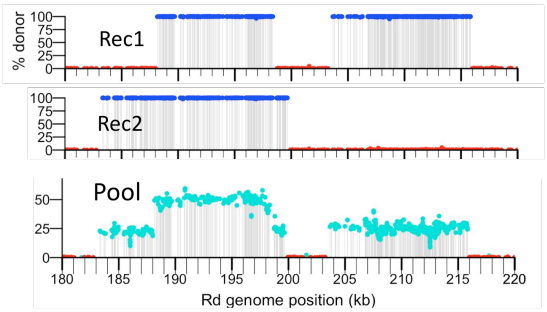
Applicants for initial grants are encouraged to summarize their previous studies and areas of expertise which are of relevance to the current proposal. One additional page may be used, and should be inserted following this page.

My laboratory focuses on the basic biology of natural competence, the ability of many bacteria to take up DNA from their environment and incorporate it into their chromosomes. Natural competence allows these bacteria to exchange genes between otherwise clonal lineages. Thus competent bacteria can share genes involved in antibiotic resistance and other pathogenesis traits, such as evasion of the immune system[1].

This proposal is a unique chance for us to directly help clinical researchers understand how bacterial genomes evolve during the chronic infections of cystic fibrosis. This combines our molecular tools for studying DNA uptake with our bioinformatics and genomics expertise, and applies them to longitudinal studies of pediatric CF being carried out by our collaborators.

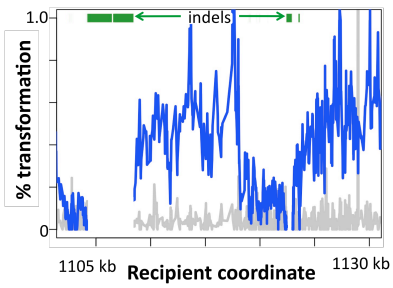
We have extensively characterized the natural competence pathway in *Haemophilus influenzae*, the subject of the proposed work, investigating its regulation, mechanism, and consequences to genetics and evolution[2-14]. We are the only group focused on natural competence in the Pasteurellaceae and have worked with *Haemophilus influenzae* for over 25 years.

We have established the protocols for growing, manipulating, and transforming *H. influenzae*[15, 16]; we have extensively characterized the regulation of natural competence[6, 8-10, 14]; we recently completed a study examining the effects of knocking out every competence-regulated genes, providing a large set of genetic resources[5]. We also have extensive experience with bioinformatics and evolutionary genetic analysis, having characterized the mechanism and evolution of DNA uptake specificity[2, 3, 17], the consequences of competence in natural populations[18], and natural variation in competence itself[12, 13].



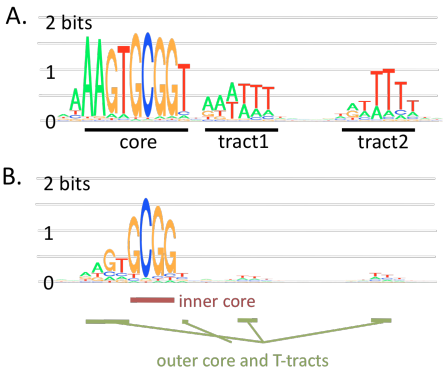
More recent work has focused on experimental genomics analyses of DNA uptake and genetic transformation, whose outcomes are directly relevant to the proposed project. We published the first genome-wide study of natural transformation, which not only showed the extent that individual competent cells are transformed by divergent DNA from a clinical isolate (0.2-3% of individual chromosomes are rapidly replaced), but also showed that we can accurately measure variant allele frequencies at nearly all positions of the genome in pools[18]. In the *example above* two transformed genomes with recombination tracts in the same 40 kb interval were pooled with two clones lacking recombination tracts in that interval. The top panels show the donor allele frequency in singly sequenced clones, and the bottom panel shows the allele frequency in the pool of clones.

Unpublished extensions of these transformation genomics studies show that we can also detect and quantify the frequencies of very rare alleles and use information within genome sequencing reads to assemble haplotypes from variant alleles in the same read pair. In the example to the *right*, total DNA from a transformed culture was sequenced to 20,000-fold genomic coverage, and the allele frequency at each polymorphic



position distinguishing donor from recipient was measured. Blue shows the donor-specific allele frequency at these positions, while grey shows the background frequency of non-donor non-reference bases at the same positions (a measure of sequencing error). The results above not only show that transformation of single chromosomes is extensive and its rate varies extensively over the chromosome, but they show that accurate allele frequency measurements can be obtained at low cost using Illumina sequencing and our computation pipelines

We have also developed a method to sequence DNA fragments taken up by *H. influenzae* cells by trapping preferred fragments in competent cells' periplasm using a translocation-deficient mutant and purifying them by organic extraction (described further in SECTION K: DETAILED PROGRAM PROPOSAL). Subsequent comparision of input DNAs and taken up DNAs shows a strong enrichment for fragments with good matches to the USS motif, highlighting the importance of positional dependencies (*i.e.* interaction effect) between USS bases (*figure to the right*). The work also suggested that this combination of molecular biology and genomic analysis could be a potent—though admittedly unusual—approach to metagenomics. We propose to highly enrich clinical DNA samples for those fragments with a high density of USSs, found only in *H. influenzae* and its Pasteurellaceae relatives[4, 19, 20], and then apply modern deep sequencing technology to study population dynamics during chronic infection.



Given our long experience with *H. influenzae* molecular genetics and evolution, with natural competence and uptake specificity, and with our recent experimental genomics work, we are the only laboratory poised to carry out the proposed use of uptake specificity to enrich for Pasteurellaceae DNA fragments from clinical samples. Because of our clinical collaborations, we will generate new insight into how *Haemophilus* genomes change and evolve during chronic infections of children with CF.

I. STATEMENT OF RELEVANCE

Applicants must describe in specific terms the relevance to, and potential importance, of the proposed research to cystic fibrosis. Relevance to CF is an important criterion for funding. In addition, please outline relevance to other diseases/disorders, i.e. possible spin-offs.

Progressive decline in lung function in CF patients is largely attributable to chronic bacterial infections. Although *Pseudomonas aeruginosa* and others dominate in adults, *Haemophilus influenzae* is often an important component in children. To better understand how bacterial populations evolve and change during long-term infection, we propose to analyze populations of *Haemophilus influenzae* on a genome-wide scale in pediatric CF patients, taking advantage of an ongoing longitudinal study of children with CF (through our collaborators).

The difficulties in studying bacterial populations at a genome-wide level are at least two-fold. Isolating and culturing independent clones from a sample allows for both genotyping and phenotyping of clones, but is laborious and expensive. Alternative culture-free methods that use DNA sequencing are often limited by human DNA in clinical samples, either targeting phylogenetic marker genes (like 16S rDNA), or describing microbial communities overall metabolim, due to the low concentration of DNA from any particular species.

To circumvent these limitations and focus analysis to a specific organism, we propose to use a recent innovation from our lab that exploits a quirk of cellular physiology—DNA uptake specificity. Like many airway pathogens and other bacteria, *H. influenzae* is naturally competent, able to take up environmental DNA and add it to its chromosomes by recombination. But *H. influenzae* only takes up DNA containing short uptake sequences, which are highly abundant in its own genome. We will use this to purify *Haemophilus* DNA from clinical samples, even in the presence of excess DNA from human and other bacteria. Using high yield DNA sequencing, we will characterize *H. influenzae* and related species' genome-wide population dynamics. This will estimate population genetic parameters (divergence, mutation, recombination, and clonal expansion) and correlate these with genomic regions, as well as with changes in *H. influenzae*'s relative abundance and in patient metadata.

Haemophilus influenzae is a commensal of the human nasopharynx, but it is also an opportunistic pathogen that causes both invasive and non-invasive disease. Carriage declines with age, with most people clearing their respiratory tracts of *H. influenzae* by adulthood. However it can make up a substantial proportion of the communities in the airways of children with CF, comprising up to 30% of the isolates from one study. Application of our method to *H.* influenzae populations in children with CF could also be extended to other diseases. H*. influenzae* is the most commonly isolated microbe from COPD disease exacerbations, and can also invade normally sterile tissues, leading to bacteremia, septic arthritis, cellulitis and meningitis in infants and small children[20]. Meningitis rates have plummeted since the introduction of a vaccine against serotype b strains[21], but other serotypes and 'non-typeable' strains continue to be causes of childhood ear infections (otitis media), conjunctivitis and sinusitis, and of pneumonia in the elderly and people with pulmonary disorders and immunodeficiency[22, 23]. First Nations populations are especially vulnerable, as are children in developing countries where the vaccine is not available[24, 25].

The immediate utility of the proposed experimental and analytical analyses are several-fold: (1) Allow rapid inexpensive genome-scale monitoring of Pasterellaeceae bacterial populations in CF patients, especially with respect to recombination and horizontal gene transfer; (2) facilitate identifying genes with clinical relevance directly from clinical DNA samples, such as antibiotic resistance and intracellular invasion; and (3) extend to study genome-wide population dynamics of other organisms with uptake specificity, *i.e.* members of the *Neisseria* genus. Eventually, we expect our approach will extend to other members of the CF microbiome, and potentially provide insights into how horizontal gene transfer between bacteria might be blocked.

J. SUMMARY OF PROPOSED RESEARCH

Please provide a one-page summary of the rationale, general objectives, and specific goals of the proposed research.

New genomics approaches to clinical microbiology are yielding broad insights into the bacterial communities living on our bodies, including a deeper understanding of the bacteria living in the respiratory tracts of CF patients. However, we have substantially less insight into the population dynamics of individual bacterial species during chronic infection: How genetically diverse are cells of a single species within a patient? How does this diversity change during disease exacerbations and therapeutic interventions? How much horizontal gene transfer takes place between related bacteria? While several analytical tools exist for such studies, appropriate genome-wide datasets from human microbiomes are sorely lacking, especially when a species' abundance is low and because human DNA can make up a large proportion of the DNA in a clinical sample. Understanding genome-wide population dynamics will inform patient treatment by indicating how populations change in response to new therapies and what bacterial genes are important during changes in disease status.

We propose a novel approach to microbial population genomics, targeting populations of the naturally competent bacterium *H. influenzae* that reside in the airways of pediatric cystic fibrosis patients. To target this group, we will exploit a peculiarity of *H. influenzae*'s natural competence pathway—DNA uptake specificity—where cells actively and preferentially take up DNA from their own species and close relatives, due to the high density of "uptake signal sequences" (USSs) in these genomes. By applying deep sequencing analysis to DNA preferentially taken up from clinical samples, we will acquire extensive population-level sampling of all *H. influenzae* genes in single clinical specimens, while excluding 'off-target' DNAs from other bacteria and the human host.

AIM A: Uptake specificity for *Haemophilus* genomes. We will perform control uptake experiments to determine how variation in genomic USSs (both their sequence and density) affects (1) the recovery of DNA fragments of different sizes, (2) the measure of allele frequencies, and (3) the effect of competition by host and other bacterial DNAs.

AIM B: *Haemophilus* population dynamics in pediatric CF. We will enrich for Pasteurellaceae DNA directly from DNA of clinical samples, primarily sputum from children with CF, prior to high coverage sequencing. These metagenomic data will be used to estimate population genetic parameters along the chromosome. Analysis of longitudinal samples will determine how intra-specific genomic diversity changes in response to changes in disease status and therapy.

The outcome of these analyses will be a rapid, low-cost, and comprehensive characterization of Pasteurellaceae communities in pediatric CF respiratory tracts. The population genetic parameters we estimate will be used to detect targets of selection in the *H. influenzae* genome within the human host, and will point to novel genes involved in evasion of the immune system and antibiotic resistance. This will directly inform studies to subvert these bacterial defenses.

The experimental and analytical techniques we establish can later be applied to mapping and identifying genes conferring clinically important phenotypes, such as antibiotic resistance and invasion of airway epithelia. The approach can eventually be extended to other bacteria of the human airway, many of which are also naturally competent and regularly exchange alleles and loci by natural transformation.

K. DETAILED PROGRAM PROPOSAL

Please provide a detailed proposal that includes the following: background, general approach and hypothesis; specific aims; research proposal, including experimental procedures and any other information germane to the proposal; significant results obtained to date; personnel requirements; and short- and long-term goals, and the hurdles you expect to encounter in achieving your goals.

Excluding references and appendices that contain relevant data presented in charts, figures, diagrams, gene maps, etc., this proposal may not exceed 10 pages in the paper copy of the application, including this one (single-sided, single-spaced, 1-inch margins on all sides, in 12 point font size). Nine additional pages may be inserted following this page. *PLEASE NOTE THAT PAGES IN EXCESS OF THE MAXIMUM WILL BE REMOVED FROM THE APPLICATION*.

The list of references and appendices of relevant data must be inserted following the Detailed Program Proposal. While the reference list and appendices of data are excluded from the 10-page limit, applicants are asked to exercise discretion and limit data to that most relevant to the proposal.

Please include a table-of-contents for Section K (Detailed Research Proposal) in both the paper copy and PDF to help guide reviewers through this section. The table-of-contents is excluded from the 10-page limit.

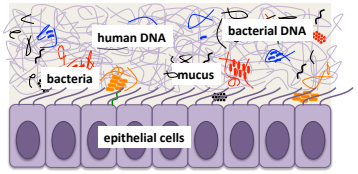
To make full use of the space in this section, you should delete the instructions above by using your delete or backspace key.

TABLE OF CONTENTS

BACTERIAL POPULATION DYNAMICS IN PEDIATRIC CYSTIC FIBROSIS AIRWAYS: TARGETED METAGENOMICS OF *HAEMOPHILUS INFLUENZAE*

I. BACKGROUND

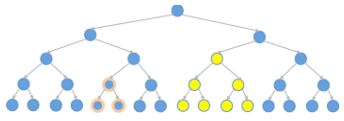
The airway microbiome and cystic fibrosis: Complex microbial communities inhabit the human airway. Because progressive decline in CF patients' lung function is due in large part to chronic bacterial infections, it is critical to understand how bacterial communities change over time: across disease exacerbations, antibiotic treatments, and indicators of the immune response. Many of the bacteria responsible for these infections also reside in normal airways, where they share an ecological niche with a wide diversity of other microbes [1]. Bacteria in the respiratory tract colonize the mucus layer that separates respiratory epithelial cells from the airway (*figure to the right*), where they form microcolonies and multispecies biofilms, binding to mucus glycoproteins, to each other, and to host cell surfaces [2].



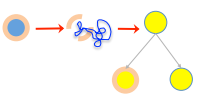
In the context of chronic infections, it is crucial to understand how individual pathogens change and evolve in response to changes in their environment. Although DNA sequencing methods can broadly profile the abundance of bacterial taxa in clinical samples, understanding genome-wide population dynamics within bacterial species cannot be easily investigated with current approaches. The ability to investigate intra-species population dynamics at a genome-wide scale would permit identification of bacterial genes under selection within individual patients, and link these to changes in treatment or disease status, especially due to recombination between related bacteria.

The background focuses attention on the bacterium *Haemophilus influenzae*, since this is the species we target in the culture-free metagenomics approach proposed below. The significance of *H. influenzae* and the proposed work to CF and other pulmonary disorders are detailed in Section I: STATEMENT OF RELEVANCE.

Population dynamics and natural competence: Populations of bacteria, even within-species, are extremely diverse (see below); new strains are acquired from other people, mutations accumulate, and recombination between strains occurs [3-4]. Bacteria reproduce clonally (*figure to the right*), but new mutations can increase in frequency if they carry a selective advantage, carrying genetic variation in the background along with them as abundance changes. In CF, *H. influenzae* strains are often mutators, at least sometimes due to mutations in the mismatch repair gene *mutS* [5-7]. This increases the mutation rate of these strains, and may facilitate adaptation.



Recombination is another important force in genome evolution [3,8-9]. Many bacteria, and many airway pathogens, are naturally competent, able to exchange genes between otherwise clonal lineages of related bacteria. Naturally competent bacterial cells actively transport DNA from their environment across the cell envelope, and this donor DNA can be added to chromosomes by homologous recombination (transformation), when sufficient sequence identity between donor and recipient molecules exists [10,11]. Transformation spreads pathogenesis-related traits through populations (*figure to the right*), both by transfer of allelic



variation and integration of whole loci [12]. Spread of antibiotic resistance is particularly important, but other traits, such as cell surface markers and the ability to invade airway epithelial cells (to evade the immune system), can also spread by transformation[13-16].

Natural competence has made major contributions to the evolution of *H. influenzae* and other competent species. For example, analyses of *H. influenzae* clinical isolates using MLST (multilocus sequence typing) found evidence of extensive recombination in housekeeping genes [16]. But different loci show varying levels of population-level recombination, potentially indicating selective forces or variation in recombination rate.

Unfortunately, most population studies have been restricted to very few loci, or to whole genome sequences from clinical isolates derived from different patients and in different contexts, so understanding how bacterial genomes change during chronic infection remains guesswork. Current approaches to studying microbiomes are described below.

(1) Culture-based methods: Culturing bacterial clones from clinical samples allows surveys of the genotypic and phenotypic diversity within particular species or groups. This has identified and characterized the dominant pathogenic species that colonize the lungs of patients with CF. These include the focus of this study, *Haemophilus influenzae* (particularly important for CF in children), along with *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Burkholderia* spp. and emerging pathogens *Achromobacter xylosoxidans* and *Stenotrophomonas maltophilia* [1,17].

Clinical isolates of *H. influenzae* display extensive phenotypic and genotypic variation. Phenotypically, clinical isolates differ in their levels of antibiotic resistance, serum resistance, intracellular invasion of epithelial cells, biofilm formation, establishment of otitis media, natural competence, mutability, and potentially even vaccine escape [6,15,18-23]. Among the 20 sequenced *H. influenzae* strains, pairs typically differ by ~2-3% substitutions per nucleotide, in addition to hundreds of “accessory” loci contained in indel polymorphisms [3,24]. How genetic variants controlling virulence traits change in frequency during chronic infection remain unknown. Another MLST study found 179 distinct *H. influenzae* strains in 127 healthy children [25].

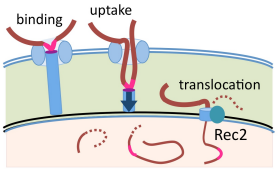
(2) Culture-free taxonomic profiling: Culture-free methods for profiling taxonomic composition in airway-associated microbiomes are increasingly common, applying inexpensive DNA sequencing directly to clinical samples. This typically focuses on profiling the relative abundance of bacterial families using PCR-based sequencing of 16S rDNA, so does not address within-group genomic variation. Such studies have indicated roles in CF disease progression for declining overall taxonomic diversity and also for difficult-to-culture anaerobes (*e.g. Prevotella* and *Veillonella* spp.)[17,26-28].

For example, a study of 6 healthy adults identified 3431 distinct 16S rDNA sequences (*i.e.* distinct taxa) in the oral cavity, and upper and lower respiratory tracts [29]. Other studies have reported different proportions of species in different parts of the airway [30-31], but these are interconnected by coughing, sneezing, swallowing, and the ‘bronchial escalator’ (cilia-driven upward flow of respiratory mucus). The microbiomes of 6 healthy people found that *H. inflenzae* and its relatives (family Pasteurellaceae) represent 6-18% of the human oropharynx microbiome, with 145 distinct Pasteurellaceae (≥97% 16S identity)[32]. This indicates that they are diverse, but tells us little about genomic variation.

(3) Metagenomic analysis: An alternative culture-free method is to simply sequence total DNA from clinical samples to high depth and use this data to characterize the microbiome as a whole. This is becoming more practical with decreasing sequencing costs, but such metagenomic information still typically says very little about genome dynamics within species. Instead, the current focus is on measuring the abundance of genes encoding different metabolic pathways (as in the human microbiome project) [33]. A major limitation, especially for airway specimens, is the high abundance of human DNA (mostly from neutrophils), which can comprise >98% of DNA from sputum, especially in CF patients [34-36], thus sufficient sequencing to obtain high coverage of bacterial metagenomes in these mixtures would be prohibitively expensive.

A need for population-specific metagenomics: Because we currently lack good metagenomic datasets of individual species, well-conceived algorithms for understanding bacterial population dynamics gather dust waiting for appropriate datasets. In particular, the framework of Johnson and Slatkin allows key population genetic parameters to be estimated directly from short-read metagenomic sequence data [37-39]. These include population-scale rates of growth, mutation, and recombination. This could identify loci under selection and mutation and recombination "hotspots", clarifying how populations evolve in chronic infection to inform treatment.

We propose to initiate genome-scale population studies right away, using as a model system the population dynamics of *H. influenzae* in a pediatric cystic fibrosis context. To enrich for *Haemophilus* DNA before sequencing, we will exploit the DNA uptake specificity of its natural competence pathway (described below). The outcome will be a low-cost comprehensive analysis of *H. influenzae* and related species' genomes from clinical samples of children with CF. This study will provide dense sampling of *H. influenzae* genetic diversity across the genome, and directly measure the potential for horizontal gene transfer into *H. influenzae* from respiratory tract DNA. We will analyze samples from ongoing longitudinal studies of CF airway bacteria, allowing changes in genomic variation to be correlated changes to pulmonary function, antibiotic regime, and other patient metadata.



The mechanism of DNA uptake: Naturally competent bacteria actively use a pseudopilus fiber to pull environmental DNA through an outer membrane pore (*figure to the right*) [11,40-41]. Once double-stranded DNA is in the periplasm, one strand is translocated through the inner membrane to the cytoplasm by a second pore (which includes the product of the *rec-2* gene [42]), while the other is degraded. Recombination of cytoplasmic DNA into chromosomes depends on the RecA recombinase and similarity between the DNAs [43-44].

The Uptake Signal Sequence: Competent *Haemophilus* cells strongly prefer to take up DNA from other *Haemophilus* cells. Cells take up *H. influenzae* DNA over unrelated DNAs by >1000-fold in competition experiments [45]. The cause of this self-specificity is the extreme over-representation in the *H. influenzae* genome of a preferred sequence motif, the uptake signal sequence (USS) [46-47]. The 9-bp core occurs nearly 1000 times per megabase of genome (Mb), dramatically more than that expected by chance. In contrast, the human genome has only ~1 core USS/Mb, and other respiratory tract

bacteria have ~1-10 core USS/Mb (TABLE 1). An extended USS motif (*figure above*) has been characterized by both bioinformatics analysis of Pasteurellaceae genomes and direct experimentation, showing that the preferred sequence can be defined by a longer motif with flanking T-rich segments and interaction effects between USS bases [45,48-49]. **USSs dramatically increase the efficiency that *Haemophilus* DNA is taken up, even in the presence of excess DNA from other sources**.

The USS acts at uptake initiation upon specific binding to the pseudopilus tip, with subsequent progression of DNA uptake acting independently of sequence. No upper size limit on uptake of USS-containing fragments has been established, but fragments from 50 bp to 50 kb are efficiently taken up. The location and density of USS sites in Pasteurellaceae genomes has remained stable over time, and the best model for their origin and maintenance is that they accumulate when random mutations that confer increased uptake efficiency spread through populations by biased DNA uptake and transformation [47]. Supporting this model, nucleotide divergence within USS sites is reduced compared to the genome-wide average (unpublished, FIG. 1).

Conclusion: We propose to first characterize this remarkable self-specificity and then and then exploit it to investigate the genome-scale population dynamics of *H. influenzae* during chronic infections of children with CF. The experimental and analytical techniques can later be applied to identifying genes conferring clinically important phenotypes and to other bacteria in the human airway, many of which are also naturally competent and regularly exchange alleles and loci by transformation.

II. GENERAL APPROACH AND HYPOTHESIS

We hypothesize that tracking *Haemophilus* population dynamics during chronic infection of children with CF will reveal how these bacteria evolve in response to therapeutic interventions and disease exacerbations. We expect that understanding these dynamics and their genomic consequences for *H. influenzae* and other Pasteurellaceae will provide critical insights into the complex ecology of the CF airway, including the identification of genes actively under selection within individual patients and the potential for horizontal gene transfer. For example, culture-based studies have found variation in antibiotic resistance in clinical isolates, and we expect that resistant strains and/or resistance alleles would increase in frequency in response to new antibiotic therapy.

To exclude most off-target DNA from humans and other bacteria before metagenomic sequencing, we will use the DNA uptake specificity of *H. influenzae* to enrich for *Haemophilus* genomes directly from DNA extracts of clinical samples. We will initially carry out a series of control uptake experiments, using pools of genomes with known composition and sequence as donor DNA. This will refine the experimental and analytical methods and provide tests of our current model of uptake specificity necessary for interpreting the data obtained from clinical samples.

Next, we will apply our method to study *Haemophilus* populations within a longitudinal study of pediatric CF, and use uptake specificity to purify and sequence to high genomic depth DNA from the *Haemophilus* cells residing in clinical samples. The resulting datasets will be analyzed using adaptations of existing algorithms that estimate population genetic parameters. **Analysis of these data will reveal how mutation,**

recombination, and selective forces act on these genomes over time within individual patients and identify genetic targets of ongoing selection.

III. Specific Aims

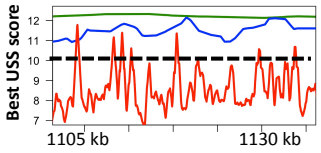
Aim A: Uptake specificity for *Haemophilus* genomes. We will perform control experiments to determine how variation in genomic USSs (both their sequence and density) affects (1) the recovery of DNA fragments of different sizes, (2) measuring of allele frequencies, and (3) effects of competition by host and other bacterial DNAs.

Aim B: *Haemophilus* population dynamics in pediatric CF. We will enrich for Pasteurellaceae DNA directly from DNA of clinical samples, primarily sputum from children with CF, prior to high coverage sequencing. These metagenomic data will be used to estimate population genetic parameters along the chromosome. Analysis of longitudinal samples will determine how intra-specific genomic diversity changes in response to changes in disease status and therapy.

IV. Research Proposal

Preliminary studies

We recently developed a method to recover and sequence DNA fragments preferentially taken up by *Haemophilus influenzae* competent cells (*figure to right*)[48]. The method exploits (1) a mutation in the recipient that blocks translocation of dsDNA out of the periplasm (the *rec-2* mutant), and (2) an organic extraction that selectively purifies periplasmic DNA and excludes bulk chromosomal DNA (Fig. 2). The purified DNA is then sequenced on the Illumina platform, giving sequences from 10^7-10^8 fragments that cells had taken up. Our first study, characterizing *H. influenzae*'s specificity for the USS, found that the genomic consensus (Fig. 1A) is the optimal uptake sequence and revealed interactions between bases in the motif [48].



We have used this new understanding to simulate an "uptake efficiency map" across the *H. influenzae* genome. In this simulation, all sequences in the genome are assigned a score based on their fit to the USS motif, and each fragment's highest scoring site predicts its uptake. The *figure to the right* shows the result for a 30 kb stretch of *H. influenzae*'s genome for 3 sizes of DNA: 0.25 kb (RED), 2.5 kb (BLUE), and 6.5 kb (GREEN). Positions above the cutoff score (10.2; dotted line) are expected to make up >98% of periplasmic DNA. This suggests that using fragments of ~6.5 kb would give appreciable coverage of the whole *H. influenzae* genome (~4-6 USS/fragment), and exclude most non-Pasteurellaeceae DNA. We expect ~1000-fold enrichment of Pasteurellaceae DNA over other sources, based on genomic USS densities (Table 1), but real enrichment may be substantially higher. We conservatively assumed that >1 USS/fragment would not increase efficiency, but this might be expected with more potential binding sites.

Because our initial study used synthetic 200 bp fragments with degenerate USSs, we confirmed our ability to purify genomic DNA from the periplasm (Fig. 3). As expected, short fragments (~250 bp) were taken up less efficiently than longer fragments (ranging from ~250 bp to 10 kb), since a smaller fraction of fragments contained USS. Yields

were sufficient for sequencing library preparation (~100-200 ng), indicating that our planned experiments are realistic.

General Methods

Cells and DNA: DNA from lab and clinical samples will be trapped by a *rec-2* mutant derivative of the lab strain KW20 [42,50]. Competent cultures will be prepared using the standard protocol of transferring exponential cultures to a starvation medium for 100 minutes before uptake experiments [51]. Donor DNA will usually be sheared to ~6.5 kb using Covaris g-tubes; other sizes will be generated by sonication (<500 bp) or HydroShear (2.5 kb). For AIM A, genomic DNA of *H. influenzae* and other bacteria will be extracted from type strains in our lab or obtained from colleagues, and human DNA will be purchased commercially (Sigma). For AIM B, total DNA from clinical samples will be obtained from our clinical collaborators (details below).

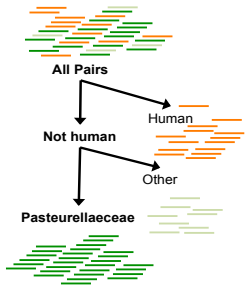
Uptake assays: All experiments proposed below will be accompanied by radiolabeled uptake assays, as described. These will include saturation curves of DNA uptake and the effects of competition between donors [51]. This will allow for optimizing experimental conditions (concentration of cells and DNA, scale needed for high yield, *etc.*). Re-uptake of periplasm-purified pools will demonstrate that the selection was successful.

DNA purifications and sequencing: Periplasm-trapped DNA fragments will be purified by organic extraction, modified from Kahn *et al.* [52-53] (*aqueous*: 1.M CsCl in TE; *organic*: 1:1 phenol:acetone), and converted to indexed Illumina libraries by standard methods (long fragments will be resheared to ~250 bp). Depending on the experiment, libraries will be sequenced on either a MiSeq or HiSeq instrument, either individually or as pools of libraries, depending on desired yield. Current yields for 2x100 bp paired-end reads are 20 million (MiSeq) or 200 million (HiSeq) (4 or 40 Gigabases of sequence) per lane. Especially for AIM A, multiplexing several libraries will reduce costs (one HiSeq lane yields ~20,000-fold coverage of *H. influenzae*).

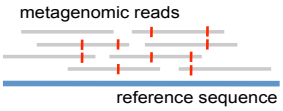
Data processing: Raw sequence reads will be aligned to all the reference genomes relevant to the sample. To account for sequence reads with ambiguous taxonomic assignment, we will use an aligner (NovoAlign or YAHA [54]) that retains reads that map to multiple references and all their associated alignments. Alignment parameters used in AIM B will be chosen based on the optimal parameters determined from the control experiments of AIM A.

For AIM A, the reference genomes will include all those in the defined sample. The ratio of recovered and input read depths at each genomic position will generate 'uptake efficiency maps', and variant allele frequencies will be calculated for positions with >10 read depth (we aim for at least 100-fold coverage of *Haemophilus*).

For AIM B, alignment to the human genome and non-Pasteurellaceae airway genomes will filter out most off-target sequences. Remaining reads will be aligned to all sequenced *H. influenzae* isolates, as well as additional Pasteurellaceae genomes (*figure to the right*). DNA from un-sequenced organisms may also be present, so we will use *de novo* assembly (SOAP [55]) to build contigs from unaligned and BLAST will attempt to identify their origin.



Subsequent processing and analysis will primarily focus on reads aligned to *H. influenzae* and other Pasteurellaceae, measuring variant allele frequencies across the chromosome, and using tools developed for phasing haplotypes to build up linkages between closely-spaced variants (*to the right*). Our collaborators Ira Hall and Paul Planet will assist us in developing computational pipelines, especially assignment of metagenomic reads to specific bacterial taxa.



AIM A: Uptake specificity for *Haemophilus* genomes

While the work of this aim will be interesting in its own right, in the context of this proposal, these experiments serve as a set of controls for the analysis of pediatric CF clinical samples in AIM B. We will be able to validate the computational methods and evaluate artifacts, since the donor DNAs will be defined.

The main products will be three sets of "uptake efficiency maps": (1) uptake maps of *H. influenzae* chromosomal DNA of three fragment sizes to test how the presence of multiple USSs on the same fragments influences uptake efficiency; (2) uptake maps from a pool of 5 sequenced *H. infuenzae* and other Pasteurellaceae to test our ability to measure variant allele frequencies and correctly recover haplotypes from population samples; (3) uptake maps of a 'reconstructed' clinical sample, consisting of genomes from *H. influenzae*, additional CF-associated microbes, and human, to test the effects of competition by chance USSs in off-target genomes. All sequencing experiments in this aim will have matched control libraries sequenced—derived from the input donor DNA—to normalize read depth from periplasm-purified sequence data.

(1) Effect of DNA fragment size: We will begin by profiling DNA uptake of *H. influenzae* genomic DNA randomly sheared to 3 different sizes, 250 bp, 2.5 kb, and 6.5 kb. We predict that the profile will be spiky when fragments are short, while more uniform with longer DNAs (described above). Comparison of the experimental and predicted maps from short fragments will identify novel sequences that promote or inhibit uptake. Maps with longer fragments will show how multiple USSs on a fragment affect uptake efficiency, allowing any synergy between USSs to be quantified.

(2) Measuring allele frequencies: The same experiment will be replicated (using 6.5 kb fragments), but using a pool of 5 DNAs from sequenced *H. influenzae* isolates pooled into 3 different proportions in order to determine the accuracy with which we can measure allele frequencies. Alignment parameters and computational pipelines will be modified to optimize the accuracy of allele frequency measurements.

(3) Effect of competition: We will finally evaluate how well *H. influenzae* DNA uptake specificity discriminates Pasteurellaceae DNA from host and other bacterial DNAs. This will repeat the uptake profiling experiment above (using long DNAs of ~6.5 kb), but with DNA composed of varying levels of foreign DNAs, particularly excess of human DNA as a competitor.

Caveats: If multiple USSs do not increase efficiency, higher sequence depth will be needed in AIM B. Measuring variant allele frequencies will be straightforward, but building these into larger haplotypes that distinguish strains will require adapting phasing tools designed for diploids [56-58]. Sequencing of off-target genomes and high

background from non-USS containing fragments may require additional optimization of experimental protocols before AIM B becomes realistic.

Outcomes: These experiments will be completed before the end of **YEAR 1** and provide crucial information about sequencing and data analysis requirements for AIM B. They also test molecular models of DNA uptake and measure the potential for horizontal gene transfer from related bacteria. High coverage sequencing of rare USS-containing loci in off-target bacterial genomes would indicate our ability to acquire limited population genetic data from non-Pasteurellaceae in AIM B.

AIM B: *Haemophilus* population dynamics in pediatric cystic fibrosis

We will use *H. influenzae*'s uptake specificity to purify USS-containing DNA fragments directly from clinical DNA extracts, allowing *Haemophilus* and related DNA to be detected over the background of DNA from the human host and other bacteria. Sequencing will then provide a high coverage Pasteurellaceae metagenome for population genetic inferences of clonal expansions, mutation rate, and recombination rate. These estimates in turn will inform our understanding of population dynamics of Pasteurellaceae species during chronic infections, across disease exacerbations and therapeutic interventions.

Clinical samples: We will take advantage of ongoing studies of pediatric CF being carried out by our clinical collaborators Drs. Paul Planet and Arnie Smith.

Dr. Planet's group is halfway into a 6-year prospective study of microbial composition in CF airways for pediatric patients (up to 20 years of age), with samples collected every three months and during exacerbations and hospitalizations. They use culture-independent methods to characterize microbial composition in sputum, oropharyngeal swabs, and brocheoalveolar lavage (BAL) samples, applying 16S rDNA sequencing to assign individual PCR fragments from these samples to microbial families. He will soon have results from 650 samples from ~150 patients. Since his study is already profiling taxonomic abundance of families, our analysis can be restricted to samples containing Pasteurellaceae; subsequent quantitative PCR (qPCR) screening by us will test which of these samples are positive for *Haemophilus influenzae*.

Dr. Smith's group runs a program that continuously monitors children with CF and obtains direct culture data from sputum samples, including CFU of *H. influenzae* per gram of sputum. This may offer the opportunity to complement metagenomic data with genome sequencing from cloned clinical isolates. Dr. Smith's group is also characterizing mutator strains frequently found in CF patients, and this will connect directly with the allele frequency measurements and estimates of population-level mutation rate we will obtain from our metagenomic survey of *Haemophilus* populations.

Timeline: In **YEAR 1**, we will apply our approach to a small cross-sectional study, selecting 15 samples from different individuals that have a high proportion of Pasteurellaceae (>5% of the bacterial 16S sequence reads in the sample). This will focus on comparing the Pasteurellaceae genomic diversity between individuals and between sputum, swab and BAL samples. In **YEAR 2**, we will apply our approach to longitudinal samples from at least 3 patients with a high proportion of Pasteurellaceae in at least 5 samples. The number of samples we can analyze will be dictated by the results of the initial study and by the relationship between budget and changing costs of

DNA sequencing. Continuation in **YEAR 3** would analyze additional sets of longitudinal samples to test specific hypotheses, particularly related to antibiotic therapy.

Initial characterization CF sputum DNA extracts: Two crucial considerations will need to be accounted for before purifying Pasteurellaceae DNA from clinical samples: (1) the concentration of human and Pasteurellaceae DNA in the extracts, and (2) the size distribution of human and Pasteurellaceae DNA. The first of these will be determined by qPCR analysis of single-copy genes from human and *H. influenzae*, and extrapolating to the total DNA of each. Second, Southern blots of DNA extracts will be probed with *H. influenzae* and human DNAs (single-copy and genomic) to measure the size distributions. These analyses will determine the amount of sequence depth that will be needed to obtain high coverage (at least 100-fold) of Pasteurellaceae in the mixture.

Enriching for *Haemophilus* DNA: Total sputum (or other) DNA provided by our collaborators (and sheared to 6.5 kb) will be subjected to uptake by *H. influenzae*, purification from the periplasm, and sequencing to high depth. For the initial cross-sectional experiment, we will sequence to high depth (~$2x10^7$ read pairs; 4 Gb). Based on the ~1000-fold difference in USS density between *Haemophilus* and unrelated genomes, we predict that a sample with 1% bacterial DNA, of which 10% is *H. influenzae*, would yield ~1000-fold coverage of the *Haemophilus* metagenome.

While much of the off-target DNA recovered is expected to be of human origin, we also expect to have enriched for any loci in foreign genomes that happen to contain USSs, providing limited population genetic data of other resident bacteria. This will complement Dr. Planet's investigations on changes in taxonomic abundance by providing measurements of the changing diversity within individual taxa.

Metagenome analysis: Following the procedures outlined in **General Methods** above, we will calculate allele frequencies at all *H. influenzae* genomic positions. Special care will be taken to define cut-off thresholds for alignments to *H. influenzae* versus other Pasteurellaceae. Using information within read pairs, we will use this to build short haplotypes in short genomic intervals by highly stringent local *de novo* assembly using GATK. This will then provide a profile of pairwise nucleotide divergence and allelic diversity across the chromosome. As we expect much of the population dynamics in these samples to be clonal, we expect to see correlated changes in allele frequency at most genomic positions at different time points, but recombination events that spread critical traits to different clones will be detected by allele frequency changes restricted to specific loci (a simplified schematic illustrated in FIG. 4).

Population genetic estimates: For more robust parameter estimates, we will use tools from Johnson and Slatkin [37-39] to make estimates of population-scaled clonal expansion rate (*R*), mutation rate (theta), and recombination rate (rho), both for the metagenome as a whole and as point estimates on sliding windows across the genome; importantly, their method reports confidence intervals. While these parameter estimates are not direct, *i.e.* they are scaled by the effective population size N_e, they can be readily compared, so for example, the ratio rho/theta provides the ratio of recombination rate / mutation rate.

The approach was specifically designed to deal with the type of metagenomic data we will obtain, with individual sequenced fragments usually derived from different cells in the population. On the one hand, using metagenomic sequence reads loses the long-

distance linkage relationships obtained from sequencing individual clones; on the other hand, the sampling itself is substantially less biased and much less laborious. The method uses the allele frequency spectrum at all genomic positions to infer population growth and mutation, and simultaneously infers recombination breakpoints using an extension of the composite likelihood estimator of McVean *et al.* [59], allowing for computationally tractable estimates from metagenomic sequence reads. Importantly, these algorithms make realistic assumptions about the mechanism of bacterial gene transfer, rather than theoretical frameworks that use assumptions from sexually reproducing eukaryotes.

Changes in *Haemophilus* populations in longitudinal series: We will use the R statistical programming language, especially the 'zoo' add-on package, to correlate changes in population genetic parameter estimates, both genome-wide and as scans along the chromosome, with patient metadata and the relative abundance of Pasteurellaceae detected in Dr. Planet's taxonomic profiles. These will reveal how clonal genetic diversity changes over time and pinpoint specific regions with elevated mutation or that are spreading by recombination in response to specific changes.

We will specifically address the question of whether intra-specific genetic diversity reduces as a function of disease progression. Several studies point to overall declining taxonomic diversity, and this could be accompanied by increasingly clonal populations within individual species.

Mutators and population mutation rate: We will specifically examine populations for the frequency of putative mutations in *mutS* and other DNA repair genes, since such mutators are highly prevalent in CF. We will examine the frequency of repair mutants in populations as a function of population mutation rate. Because *mutS* has also been implicated in the fidelity of recombination, we will also examine how mutator frequency correlates with population-recombination rates.

Caveats: (1) If Pasteurellaceae DNA is extremely rare in a clinical sample, coverage may be lower than predicted and require more sequencing than is affordable for a specific sample (though we expect costs to continue to drop). This might be a particular problem during CF exacerbations, since the inflammatory response is correlated with increased neutrophil DNA in sputum samples. Alternatives include subtractive hybridization using commercial human DNA, or multiple rounds of uptake and recovery. We may need to focus on young children and infants, where total DNA in sputum is lower. (2) Some samples, particularly for oropharyngeal swabs, will contain very little total DNA, so whole-genome amplification may be required. (3) The ability to assign sequences to specific species is essentially determined by the stringency of our alignment, and parameter estimates may vary, depending on alignment parameters. (4) Recent foreign acquisitions in *H. influenzae* (*e.g.* prophage) contain few or no USSs, so we will miss these accessory loci.

Outcomes: We expect that this will be the most comprehensive analysis of natural bacterial populations ever conducted, and it will be directly relevant to understanding chronic infections in CF and provide insights into patient care. The approach will be exceptionally inexpensive compared to alternatives, and it offers a novel and powerful method for studying Pasteurellaceae in other contexts.

References for SECTIONS G, I and J

1. Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., et al. (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A* **98**, 182-7.
2. Findlay, W.A., and Redfield, R.J. (2009). Coevolution of DNA uptake sequences and bacterial proteomes. *Genome Biol Evol* **1**, 45-55.
3. Maughan, H., Wilson, L.A., and Redfield, R.J. (2010). Bacterial DNA uptake sequences can accumulate by molecular drive alone. *Genetics* **186**, 613-27.
4. Redfield, R.J., Findlay, W.A., Bosse, J., Kroll, J.S., Cameron, A.D., and Nash, J.H. (2006). Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol Biol* **6**, 82.
5. Sinha, S., Mell, J.C., and Redfield, R.J. (2012). 17 CRP-S-regulated genes are needed for natural transformation in Haemophilus influenzae. *J Bacteriol*.
6. Cameron, A.D., and Redfield, R.J. (2006). Non-canonical CRP sites control competence regulons in Escherichia coli and many other gamma-proteobacteria. *Nucleic Acids Res* **34**, 6001-14.
7. Cameron, A.D., and Redfield, R.J. (2008). CRP binding and transcription activation at CRP-S sites. *J Mol Biol* **383**, 313-23.
8. Cameron, A.D., Volar, M., Bannister, L.A., and Redfield, R.J. (2008). RNA secondary structure regulates the translation of sxy and competence development in Haemophilus influenzae. *Nucleic Acids Res* **36**, 10-20.
9. MacFadyen, L.P., Chen, D., Vo, H.C., Liao, D., Sinotte, R., and Redfield, R.J. (2001). Competence development by Haemophilus influenzae is regulated by the availability of nucleic acid precursors. *Mol Microbiol* **40**, 700-7.
10. Macfadyen, L.P., Dorocicz, I.R., Reizer, J., Saier, M.H., Jr., and Redfield, R.J. (1996). Regulation of competence development and sugar utilization in Haemophilus influenzae Rd by a phosphoenolpyruvate:fructose phosphotransferase system. *Mol Microbiol* **21**, 941-52.
11. Macfadyen, L.P., Ma, C., and Redfield, R.J. (1998). A 3',5' cyclic AMP (cAMP) phosphodiesterase modulates cAMP levels and optimizes competence in Haemophilus influenzae Rd. *J Bacteriol* **180**, 4401-5.
12. Maughan, H., and Redfield, R.J. (2009). Tracing the evolution of competence in Haemophilus influenzae. *PLoS One* **4**, e5854.
13. Maughan, H., and Redfield, R.J. (2009). Extensive variation in natural competence in Haemophilus influenzae. *Evolution* **63**, 1852-66.
14. Redfield, R.J., Cameron, A.D., Qian, Q., Hinds, J., Ali, T.R., Kroll, J.S., and Langford, P.R. (2005). A novel CRP-dependent regulon controls expression of competence genes in Haemophilus influenzae. *J Mol Biol* **347**, 735-47.
15. Poje, G., and Redfield, R.J. (2003). General methods for culturing Haemophilus influenzae. *Methods Mol Med* **71**, 51-6.
16. Poje, G., and Redfield, R.J. (2003). Transformation of Haemophilus influenzae. *Methods Mol Med* **71**, 57-70.
17. Mell, J.C., Hall, I.M., and Redfield, R.J. (2012). Defining the DNA uptake specificity of naturally competent Haemophilus influenzae cells. *Nucleic Acids Res*.
18. Mell, J.C., Shumilina, S., Hall, I.M., and Redfield, R.J. (2011). Transformation of natural genetic variation into Haemophilus influenzae genomes. *PLoS Pathog* **7**, e1002151.
19. Bosse, J.T., Sinha, S., Schippers, T., Kroll, J.S., Redfield, R.J., and Langford, P.R. (2009). Natural competence in strains of Actinobacillus pleuropneumoniae. *FEMS Microbiol Lett* **298**, 124-30.
20. Kristensen, B.M., Sinha, S., Boyce, J.D., Bojesen, A.M., Mell, J.C., and Redfield, R.J. (2012). Natural Transformation of Gallibacterium anatis. *Appl Environ Microbiol* **78**, 4914-22.
21. Cardines, R., Giufre, M., Pompilio, A., Fiscarelli, E., Ricciotti, G., Bonaventura, G.D., and Cerquetti, M. (2012). Haemophilus influenzae in children with cystic fibrosis: antimicrobial susceptibility, molecular epidemiology, distribution of adhesins and biofilm formation. *Int J Med Microbiol* **302**, 45-52.
22. Hallstrom, T., and Riesbeck, K. (2010). Haemophilus influenzae and the complement system. *Trends Microbiol* **18**, 258-65.
23. Heath, P.T. (1998). Haemophilus influenzae type b conjugate vaccines: a review of efficacy data. *Pediatr Infect Dis J* **17**, S117-22.
24. Agrawal, A., and Murphy, T.F. (2011). Haemophilus influenzae infections in the H. influenzae type b conjugate vaccine era. *J Clin Microbiol* **49**, 3728-32.
25. Dworkin, M.S., Park, L., and Borchardt, S.M. (2007). The changing epidemiology of invasive Haemophilus influenzae disease, especially in persons > or = 65 years old. *Clin Infect Dis* **44**, 810-6.
26. Saha, S.K., Baqui, A.H., Darmstadt, G.L., Ruhulamin, M., Hanif, M., El Arifeen, S., Oishi, K., Santosham, M., Nagatake, T., and Black, R.E. (2005). Invasive Haemophilus influenzae type B diseases in Bangladesh, with increased resistance to antibiotics. *J Pediatr* **146**, 227-33.

References for SECTION K: DETAILED PROGRAM PROPOSAL

1. Delhaes L, Monchy S, Frealle E, Hubans C, Salleron J, et al. (2012) The airway microbiota in cystic fibrosis: a complex fungal and bacterial community--implications for therapeutic management. PLoS One 7: e36313.
2. Murphy TF, Bakaletz LO, Smeesters PR (2009) Microbial interactions in the respiratory tract. Pediatr Infect Dis J 28: S121-126.
3. Feil EJ, Spratt BG (2001) Recombination and the population structures of bacterial pathogens. Ann Rev Micro 55: 561.
4. Spratt BG, Hanage WP, Feil EJ (2001) The relative contributions of recombination and point mutation to the diversification of bacterial clones. Curr Opin Microbiol 4: 602-606.
5. Watson ME, Jr., Burns JL, Smith AL (2004) Hypermutable Haemophilus influenzae with mutations in mutS are found in cystic fibrosis sputum. Microbiology 150: 2947-2958.
6. Roman F, Canton R, Perez-Vazquez M, Baquero F, Campos J (2004) Dynamics of long-term colonization of respiratory tract by Haemophilus influenzae in cystic fibrosis patients shows a marked increase in hypermutable strains. J Clin Microbiol 42: 1450-1459.
7. Perez-Vazquez M, Roman F, Garcia-Cobos S, Campos J (2007) Fluoroquinolone resistance in Haemophilus influenzae is associated with hypermutability. Antimicrob Agents Chemother 51: 1566-1569.
8. Maiden MC (1998) Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. Clin Infect Dis 27 Suppl 1: S12-20.
9. Didelot X, Maiden MC (2010) Impact of recombination on bacterial evolution. Trends Microbiol 18: 315-322.
10. Redfield RJ (2001) Do bacteria have sex? Nat Rev Genet 2: 634-639.
11. Maughan H, Sinha S, Wilson L, Redfield RJ (2008) Competence, DNA Uptake and Transformation in the Pasteurellaceae. In: Kuhnert P, Christensen H, editors. Pasteurellaceae: Biology, Genomics and Molecular Aspects: Caister Academic Press.
12. Kruger NJ, Stingl K (2011) Two steps away from novelty-principles of bacterial DNA uptake. Mol Micro 80: 860-867.
13. Clementi CF, Murphy TF (2011) Non-Typeable Haemophilus influenzae Invasion and Persistence in the Human Respiratory Tract. Front Cell Infect Microbiol 1: 1.
14. Maragakis LL, Perencevich EN, Cosgrove SE (2008) Clinical and economic burden of antimicrobial resistance. Expert Rev Anti Infect Ther 6: 751-763.
15. Nakamura S, Shchepetov M, Dalia AB, Clark SE, Murphy TF, et al. (2011) Molecular basis of increased serum resistance among pulmonary isolates of non-typeable Haemophilus influenzae. PLoS Pathog 7: e1001247.
16. Cody AJ, Field D, Feil EJ, Stringer S, Deadman ME, et al. (2003) High rates of recombination in otitis media isolates of non-typeable Haemophilus influenzae. Infect Genet Evol 3: 57-66.
17. Goddard AF, Staudinger BJ, Dowd SE, Joshi-Datar A, Wolcott RD, et al. (2012) Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. Proc Natl Acad Sci U S A 109: 13769-13774.
18. Morey P, Cano V, Marti-Lliteras P, Lopez-Gomez A, Regueiro V, et al. (2011) Evidence for a non-replicative intracellular stage of nontypable Haemophilus influenzae in epithelial cells. Microbiology 157: 234-250.
19. Look DC, Chin CL, Manzel LJ, Lehman EE, Humlicek AL, et al. (2006) Modulation of airway inflammation by Haemophilus influenzae isolates associated with chronic obstructive pulmonary disease exacerbation. Proc Am Thorac Soc 3: 482-483.
20. Starner TD, Zhang N, Kim G, Apicella MA, McCray PB, Jr. (2006) Haemophilus influenzae forms biofilms on airway epithelia: implications in cystic fibrosis. Am J Respir Crit Care Med 174: 213-220.
21. Leibovitz E, Jacobs MR, Dagan R (2004) Haemophilus influenzae: a significant pathogen in acute otitis media. Pediatr Infect Dis J 23: 1142-1152.
22. Maughan H, Redfield RJ (2009) Tracing the evolution of competence in Haemophilus influenzae. PLoS One 4: e5854.
23. Schouls L, van der Heide H, Witteveen S, Zomer B, van der Ende A, et al. (2008) Two variants among Haemophilus influenzae serotype b strains with distinct bcs4, hcsA and hcsB genes display differences in expression of the polysaccharide capsule. BMC Microbiol 8: 35.
24. Mell JC, Shumilina S, Hall IM, Redfield RJ (2011) Transformation of natural genetic variation into Haemophilus influenzae genomes. PLoS Pathog 7: e1002151.
25. Farjo RS, Foxman B, Patel MJ, Zhang L, Pettigrew MM, et al. (2004) Diversity and sharing of Haemophilus influenzae strains colonizing healthy children attending day-care centers. Pediatr Infect Dis J 23: 41-46.
26. Cox MJ, Allgaier M, Taylor B, Baek MS, Huang YJ, et al. (2010) Airway microbiota and pathogen abundance in age-stratified cystic fibrosis patients. PLoS One 5: e11044.
27. Rogers GB, Carroll MP, Serisier DJ, Hockey PM, Jones G, et al. (2004) characterization of bacterial community diversity in cystic fibrosis lung infections by use of 16s ribosomal DNA terminal restriction fragment length polymorphism profiling. J Clin Microbiol 42: 5176-5183.
28. Field TR, Sibley CD, Parkins MD, Rabin HR, Surette MG (2010) The genus Prevotella in cystic fibrosis airways. Anaerobe 16: 337-344.

29. Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, et al. (2011) Topographical continuity of bacterial populations in the healthy human respiratory tract. Am J Respir Crit Care Med 184: 957-963.
30. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, et al. (2009) Bacterial community variation in human body habitats across space and time. Science 326: 1694-1697.
31. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, et al. (2011) Analysis of the lung microbiome in the "healthy" smoker and in COPD. PLoS One 6: e16384.
32. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, et al. (2012) Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13: R42.
33. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8: e1002358.
34. Lethem MI, James SL, Marriott C, Burke JF (1990) The origin of DNA associated with mucus glycoproteins in cystic fibrosis sputum. Eur Respir J 3: 19-23.
35. Ratjen F, Paul K, van Koningsbruggen S, Breitenstein S, Rietschel E, et al. (2005) DNA concentrations in BAL fluid of cystic fibrosis patients with early lung disease: influence of treatment with dornase alpha. Pediatr Pulmon 39: 1-4.
36. Brandt T, Breitenstein S, von der Hardt H, Tummler B (1995) DNA concentration and length in sputum of patients with cystic fibrosis during inhalation with recombinant human DNase. Thorax 50: 880-882.
37. Johnson PL, Slatkin M (2006) Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res 16: 1320-1327.
38. Johnson PL, Slatkin M (2008) Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol 25: 199-206.
39. Johnson PL, Slatkin M (2009) Inference of microbial recombination rates from metagenomic data. PLoS Genet 5: e1000674.
40. Dubnau D (1999) DNA uptake in bacteria. Annu Rev Microbiol 53: 217-244.
41. Pelicic V (2008) Type IV pili: e pluribus unum? Mol Microbiol 68: 827-837.
42. Barouki R, Smith HO (1985) Reexamination of phenotypic defects in *rec-1* and *rec-2* mutants of *Haemophilus influenzae* Rd. J Bacteriol 163: 629-634.
43. Kowalczykowski SC, Eggleston AK (1994) Homologous pairing and DNA strand-exchange proteins. Annu Rev Biochem 63: 991-1043.
44. Cox MM (1991) The RecA protein as a recombinational repair system. Mol Microbiol 5: 1295-1299.
45. Redfield RJ, Findlay WA, Bosse J, Kroll JS, Cameron AD, et al. (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol Biol 6: 82.
46. Findlay WA, Redfield RJ (2009) Coevolution of DNA uptake sequences and bacterial proteomes. Gen Bio Evo 1: 45.
47. Maughan H, Wilson LA, Redfield RJ (2010) Bacterial DNA uptake sequences can accumulate by molecular drive alone. Genetics 186: 613-627.
48. Mell JC, Hall IM, Redfield RJ (2012) Defining the DNA uptake specificity of naturally competent Haemophilus influenzae cells. Nucleic Acids Res.
49. Smith HO, Tomb JF, Dougherty BA, Fleischmann RD, Venter JC (1995) Frequency and distribution of DNA uptake signal sequences in the Haemophilus influenzae Rd genome. Science 269: 538-540.
50. Sinha S, Mell JC, Redfield RJ (2012) 17 CRP-S-regulated genes are needed for natural transformation in Haemophilus influenzae. J Bacteriol.
51. Poje G, Redfield RJ (2003) Transformation of Haemophilus influenzae. Methods Mol Med 71: 57-70.
52. Kahn ME, Barany F, Smith HO (1983) Transformasomes: specialized membranous structures that protect DNA during Haemophilus transformation. Proc Natl Acad Sci U S A 80: 6927-6931.
53. Kahn M, Concino M, Gromkova R, Goodgal S (1979) DNA binding activity of vesicles produced by competence deficient mutants of Haemophilus. Biochem Biophys Res Commun 87: 764-772.
54. Faust GG, Hall IM (2012) YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics.
55. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.
56. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
58. Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76: 449-462.
59. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160: 1231-1241.

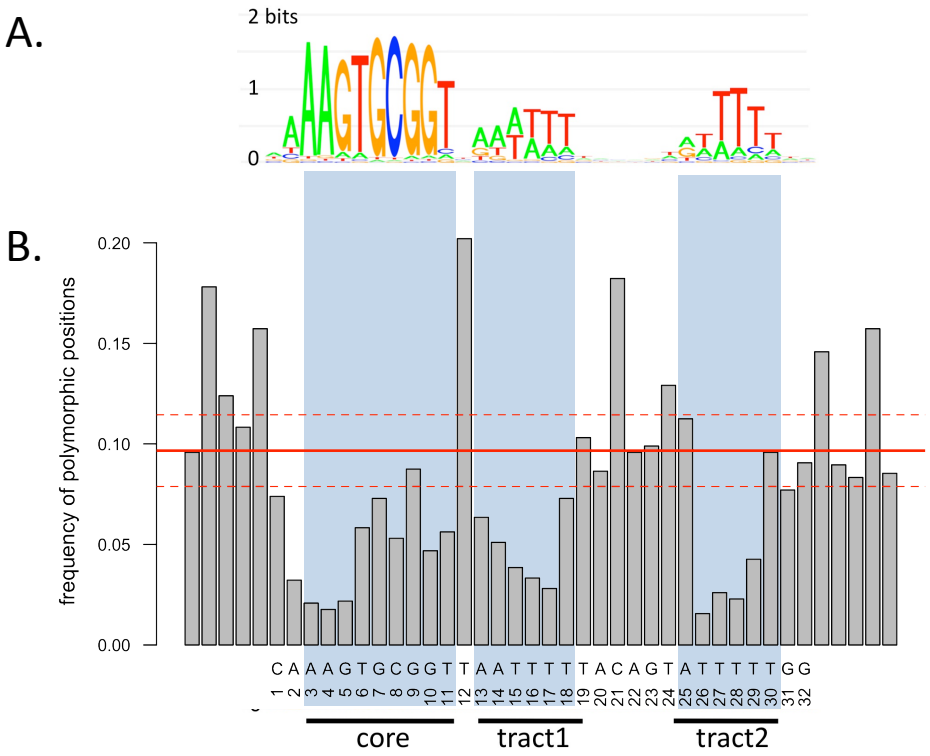
TABLE 1: USS density in representative genomes

Type genomes of representative taxa	Size (Mb)	%GC	USS[b]	USS/Mb
Typical CF pathogens[a]				
Haemophilus influenzae 86-028NP	1.9	38.2%	1516	791.9
Pseudomonas aeruginosa PAO1	6.3	66.6%	17	2.7
Staphylococcus aureus N315	2.8	32.8%	9	3.2
Burkholderia cepacia GG4	6.5	66.7%	19	2.9
Stenothrophomonas maltophilia K279a	4.9	66.3%	15	3.1
Achromobacter xylosoxidans A8	7.0	66.0%	14	2
Atypical CF pathogens[a]				
Prevotella melaninogenica ATCC 25845	3.2	41.0%	10	3.2
Fusobacterium nucleatum ATCC 25586	2.2	27.2%	2	0.9
Bacteroides fragilis YCH46	5.3	43.3%	20	3.8
Veillonella parvula DSM 2008	2.1	38.6%	4	1.9
Porphyromonas gingivalis W83	2.3	48.3%	7	3
Other respiratory tract microbes and Pasteurellaceae				
Aggregat. Actinomycetem. D115-1	2.1	44.6%	1760	835.8
Pasteurella multocida Pm70	2.3	40.4%	927	410.6
Streptococcus pneumoniae TIGR4	2.2	39.7%	8	3.7
Moraxella catarrhalis RH4	1.9	41.7%	19	10.2
Neisseria meningitidis MC58	2.3	51.5%	25	11
Neisseria lactamica 020-06	2.2	52.3%	23	10.4
Neisseria gonorrhoeae FA1090	2.2	52.7%	21	9.7
Host				
Homo sapiens hg19	3,137.2	41.6%	2972	0.9

a Taken from **Gaddard *et al.* 2012, PNAS**

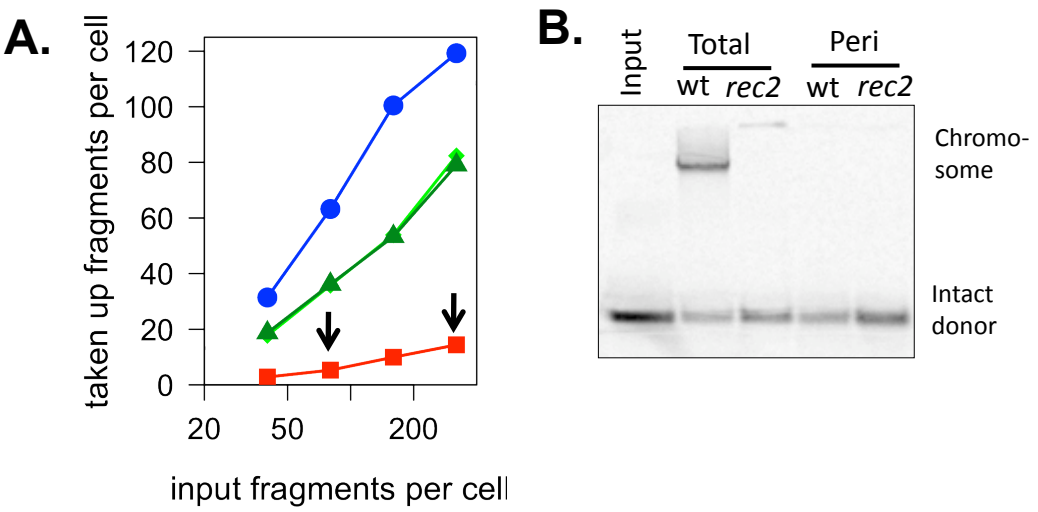
b Exact count of the 9mer AAGTGCGGT and its reverse complement (determined in UNIX)
Note, this is a simplistic view of the USS, and we later will use the full motif model.

FIGURE 1: Genetic divergence is decreased within USSs



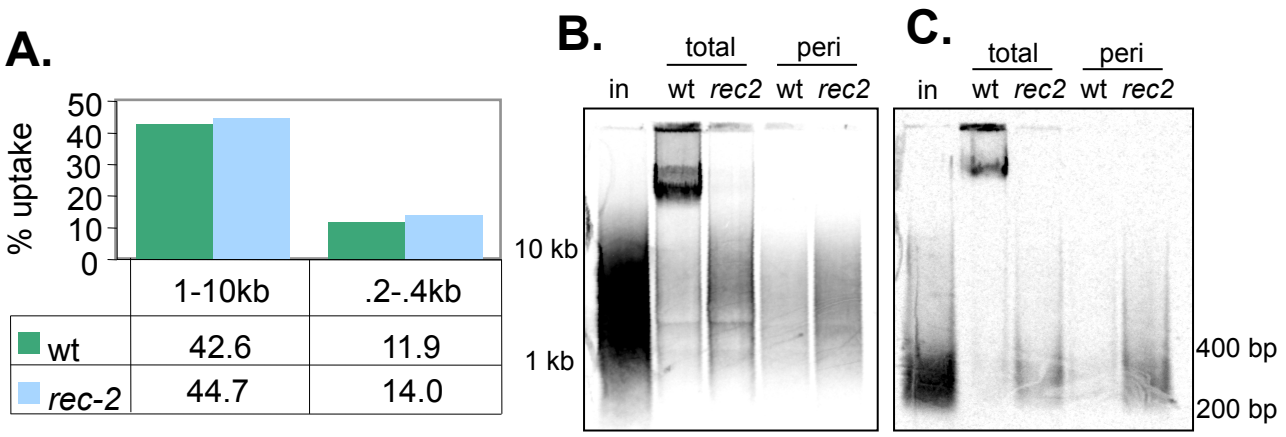
A. *Genomic USS motif.* Sequence logo of the 2205 genomic USSs in the *H. influenzae* KW20 genome. Other isolates have indistinguishable genomic USS motifs. **B.** *Sequence divergence is constrained within the USS core and flanking T-rich tracts.* Data was derived from a Mauve multiple alignment of all 20 *H. influenzae* genome sequences. First, anywhere the USS scoring matrix derived in **Mell *et al.* 2012** finds a high scoring USS, each position of the USS ± 5bp was scored for whether the position was polymorphic in the multiple alignment. The total frequency that the position was polymorphic across all USSs in the dataset is shown on the y-axis. Grey bars are for USS and flaking bases. The red line indicates the genome average, and the dotted red lines indicate 95% confidence intervals.

FIGURE 2: Uptake and re-uptake of synthetic DNA



A. Uptake by *rec-2* cultures after 30 min of DNA fragments with the consensus USS (blue circles), a 24% degenerate USS pool (red squares), or one of two recovered periplasmic pools (green triangles and diamonds, corresponding to purifications from cell incubated with 320 and 32 fragments/cell, respectively (black arrows)). **B.** Autoradiogram showing total and periplasmic organic extractions from wt and *rec-2* cultures after 5 minutes uptake (128 USS fragments/cell). The aqueous fraction of periplasmic extractions retains intact fragments but excludes labeled chromosomal DNA. Adapted from **Mell *et al.* 2012**.

FIGURE 3: Uptake and purification of sheared genomic DNA



A. % uptake for sonicated fragments of two size distributions: 1-10 kb and 200-400 bp (200 ng DNA / 10^9 cells after 30 minutes. **B.** and **C.** show gels of 1-10kb and 200-400 bp fragments taken up by cells, respectively. Cells incubated with DNA were split in two aliquots, where total DNA was extracted from one, and periplasmic DNA from the other. Notably, chromosome labeling is only observed in the total DNA of wild-type cells, but not in the periplasmic prep of wild-type cells. Furthermore, *rec-2* cells accumulate intact periplasmic DNA, and size distributions are comparable between input and recovered pools.

FIGURE 4: Simplified schematic of recombination *in vivo*