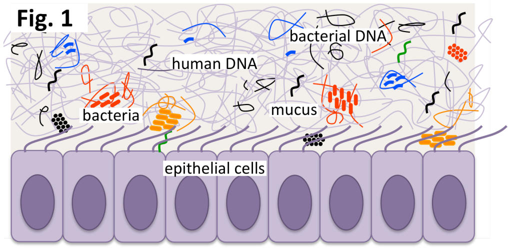
RATIONALE

Researchers and clinicians are losing the battle against bacterial pathogens, ultimately because we are unable to prevent or control horizontal gene transfer. I propose to remedy this by developing the first predictive model of gene transfer in the human respiratory tract.

Attempts at infection control commonly fail because pathogens have acquired new genes and alleles that confer antibiotic resistance or escape from immune surveillance and vaccine immunity[1, 2]. These alleles are acquired from both close relatives and other species, by transformation, conjugation or transduction. Gene transfer is of particular concern in respiratory communities, since many important pathogens are highly competent for DNA uptake and natural transformation. **Fig. 1**. illustrates a typical respiratory tract community. Opportunistic pathogens intermingle with true commensals, all embedded in a matrix of mucins and DNA that flows like a river across the ciliated mucosal surface. Diverse bacteria use their competence systems to take up the DNA, obtaining both nucleotides and the genes that will subsequently prevent control of infection. Genetic exchange in this complex evolving ecosystem has never been systematically investigated. Without any way to predict which new genotypes might arise, researchers and clinicians can only hope that new strains will not emerge, and then struggle to control the damage when they inevitably do. The greatest need is for knowledge of the key factors affecting DNA uptake and homologous recombination in the respiratory tract. Although laboratory studies have produced a wealth of information about the regulation of competence and the mechanism of DNA uptake[3], these cannot easily be applied to real infections, and population genetics studies of past recombination events are confounded by the unknown effects of genetic drift and natural selection[4].

Fig. 1



> **Hypothesis:** Identifying the constraints on natural transformation will allow recombination events in the respiratory tract to be predicted and treatment failures to be prevented.

Significance: The ability to anticipate gene transfer patterns will help researchers prepare for new strains of bacteria and guide the design of therapies that are less vulnerable to these changes, especially when applied to data from personalized and community studies of respiratory microbiota. Identification of the causal factors underlying these predictions may also allow interventions to prevent exchange.

BACKGROUND

PDFs of papers **[A-E]** *are provided in the Appendix; citations to other work from my group are in* ***bold***.

Gene transfer is the ultimate cause of most infection-control failures. The spread of antibiotic resistance is now officially one of the three greatest threats to human health[5]. The impact on respiratory tract (RT) pathogens is well documented. The chromosomal *penA* allele that now prevents treatment of *Neisseria gonorrhoeae* infections with cefixime is likely to spread to *N. meningitidis*[6]. A series of recent recombination events in *Streptococcus pneumoniae* have produced the multi-drug resistant vaccine-escape serotype 19A strain now spreading through Canada and the rest of the world[7]. The spread of antibiotic resistance not only results in increased disease severity, morbidity and mortality, but also costs an additional $6000-$30,000 per hospitalized patient[8]. Gene transfer also spreads such clinically important traits as serum resistance, capsulation, and intracellular invasion and survival[9, 10]. For example, the serotype-specific vaccine that currently prevents *Haemophilus influenzae* meningitis is threatened by exchange of capsule-determining genes[11].

How gene transfer occurs. DNA transfer in bacteria is unidirectional, whether by conjugation, transduction or transformation (the focus of this proposal). In addition to replacing one chromosomal allele with another, transformation can introduce foreign genes and large genetic islands if they are flanked on one or both sides by regions homologous to the recipient chromosome[12, 13]. Integrons and other resistance elements can also move from foreign DNA fragments into the recipient chromosome, and from there spread through the population by transformation[14].

Microbiota in healthy and diseased RT: A handful of pathogens cause most acute RT infections (pharyngitis, bronchitis, pneumonia). However, much more diverse populations cause morbidity and mortality in chronic infections in patients with asthma, chronic obstructive pulmonary disease (COPD) and cystic fibrosis (CF). The bacteria responsible for these infections also occur in normal RT, where they share an ecological niche with an enormous diversity of other bacteria. Most clinically important gene transfer likely occurs in the healthy RT rather than in infections, because of its diversity and the long residence times of these populations. Bacteria in the RT colonize the mucus layer that separates respiratory epithelial cells from the airway, where they form microcolonies and multispecies biofilms, binding to mucus glycoproteins, to each other and to host cell surfaces[15]. A study of 6 healthy adults identified 3431 distinct 16S rRNA sequences in the oral cavity, upper RT and lower RT[16]. Although other studies[17, 18] have reported different proportions of species in different parts of the RT, these environments are typically interconnected by coughing, sneezing, swallowing and the 'bronchial escalator' (cilia-driven upward flow of respiratory mucus)[16].

Haemophilus influenzae**:** This project will focus on the small Gram-negative bacterium *H. influenzae* (Pasteurellaceae)*,* a frequent cause of serious RT infections[19]. Its small well-studied genome, many sequenced isolates, and well-defined competence system facilitate studies of recombination. Its abundance and diversity in respiratory microbiomes create frequent opportunities for gene transfer, and the severity of the diseases it causes makes the results medically valuable.

***H. influenzae* is an important pathogen:** *H. influenzae* only infects humans. A normal component of upper RT microbiota and a frequent cause of RT diseases, it can also invade normally sterile tissues, leading to bacteremia, septic arthritis, cellulitis and especially meningitis in infants and small children; the latter has a 6% mortality rate and residual damage to hearing or intellect in about 50% of cases[20]. Meningitis rates have plummeted since the introduction of an effective vaccine against the type b serotype[21], but other serotypes and 'non-typeable' strains continue to be major causes of childhood ear infections (otitis media), conjunctivitis and sinusitis, and of pneumonia in the elderly and people with CF, COPD and AIDS[22, 23]. First Nations populations are especially vulnerable, as are children in developing countries where the vaccine is not available[24, 25].

***H. influenzae* diversity and recombination:** The RT offers many opportunities for recombination among and between commensal *H. influenzae* strains and closely related Pasteurellacean species. Analysis of samples from multiple sites in the mouth and throat found that Pasteurellaceae represent 6-18% of the human oropharynx microbiome[26]. The microbiomes of 6 healthy people included 145 distinguishable Pasturellaceae strains (≤97% 16S identity), and another study found 179 different strains of just *H. influenzae* in 127 healthy children[26, 27]. Analysis of 25 otitis media isolates found evidence of extensive recombination in both housekeeping and LPS genes[28].

Tools for *H. influenzae* genetics and genomics: *H. influenzae* is a tractable organism for genomic work; its genome is small (~1.9 Mb) and contains almost no repetitive DNA apart from the 6 copies of the rRNA operon. Genome sequences are available for 20 strains and for five strains of the close relative *H. haemolyticus*. As with other bacterial species, different strains show extensive genetic variation: single-nucleotide polymorphisms (SNPs) and insertion/deletions (indels), inversions, and more complex rearrangements. Any two sequenced isolates typically differ at ~2-3% of alignable positions, providing ~40,000 markers along the chromosome, and in several hundred genes **[A]**. Transformation is of course a powerful tool for genetic analysis, and we have used it with recombineering to fully characterize the 25-gene competence regulon and generate useful mutants[29-31].

Natural competence and transformation in *H. influenzae*: The different stages of transformation in *H. influenzae* and other Gram-negative bacteria are illustrated and numbered in **Fig. 2** (also see http://bit.ly/9abvEL). Cells become competent to take up DNA when the competence regulon is induced by nutritional signals[30, 32]. DNA uptake then proceeds in two steps: **(1)** double-stranded DNA is taken up across the outer membrane into the periplasm, and **(2)** a single DNA strand is translocated through a *rec2*-encoded pore in the inner membrane into the cytoplasm (the other strand is degraded **(3)**). Once in the cytosol, single-stranded DNA can recombine into the recipient chromosome by RecA-catalyzed strand invasion **(4)**; the resulting base-pair mismatches may be 'corrected' by mismatch repair **(5)**. Individual competent cells can take up several large fragments of genomic DNA and the frequency of transformation by single-nucleotide changes is typically about 0.2-1.0% **[A]**. The following paragraphs describe processes affecting transformation that will be characterized and incorporated into a quantitative model that calculates the probability of uptake and transformation for any DNA sequence. (A schematic of the planned model is shown in Appendix **Fig. A-1**.)

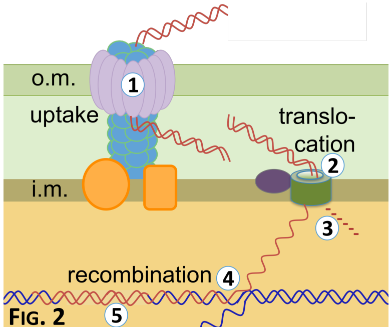


Fig. 2

Competence development: The type 4 pilus genes required for DNA uptake are also required for colonization of RT epithelia, and are strongly expressed *in vivo*[33]. However, we have found dramatic difference between strains in their ability to take up DNA and become transformed **[B]**. In induced cultures of the standard lab strain Rd, competence appears to be all-or-nothing at the cell level, with 10-50% of cells highly competent and the rest not transformable cells (these can be eliminated by selection).

DNA uptake (Fig. 2, step 1): This is highly sequence-dependent, preferentially beginning at 'uptake signal sequences' (USSs) (**Fig. 3**) highly overrepresented in genomes of all Pasteurellaceae, ~1/kb **[C]**[34-36]. Although USSs are often assumed to be absolutely required for uptake, they are better viewed as a sequence bias that helps the uptake machinery overcome the physical constraints imposed by stiff highly charged DNA molecules **[D]**. The importance of both transformation and uptake bias in the RT is confirmed by the reduced sequence divergence at USS positions seen when genomes of 20 *H. influenzae* strains are compared (unpublished data, **Fig. A-2**). **Fig. 3** shows sequence logos of **a.** the USS motif derived from genomic sequences and **b.** the uptake bias we have directly measured using Illumina sequencing of a degenerate test fragment after uptake and recovery from competent cells **[D]**. The unexpected discrepancy between **a.** and **b.** suggests that unknown effects also make important contributions to uptake or transformation biases. These will be investigated in Aims 1 and 2.

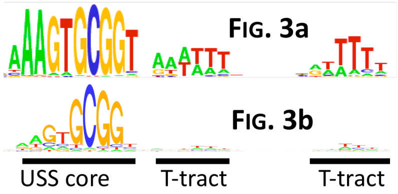


Fig. 3a

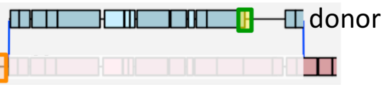
Fig. 3b

Translocation and strand degradation (Fig. 2, steps 2 and 3): DNA that does not recombine is fully degraded by cytoplasmic nucleases, but little is known about other influences on translocation and strand degradation. Early studies reported degradation of about 1.5 kb at the 3' end of the strand that enters the cytoplasm, and little or no degradation of its 5' end[37, 38]. The responsible nucleases are not known; knockouts of *recBC, recJ, recG* and *exoI* do not affect transformation[39, 40].

Recombination (Fig. 2, step 4): Recombination between homologous sequences depends on the degree of identity between them. RecA coats ssDNA and efficiently synapses it with homologous sequences in the chromosome[41, 42]. This synapsis can initiate at either end of the coated filament or internally; if the latter, the region of synapsis can spread in both directions[43]. This extension happens in steps of ~80 bp (important in Aim 2) that are aborted if sequence homology is insufficient[43]. Mismatched bases strongly inhibit initiation but have much less effect on subsequent strand incorporation[44]. Chi sites are not expected to contribute to transformational recombination since one DNA is already single-stranded. Much less is known about the effects of indels (insertion/deletions) and other 'structural' variation. In transformable bacteria, large insertions are recombined into recipient DNA about 10-fold less efficiently than single-nucleotide differences.

To begin clarifying how sequence variation affects transformation, we have used high-coverage DNA sequencing to analyze the recombination tracts created when DNA of a divergent strain transformed competent cells of the standard Rd strain **[A]**. These four transformant genomes contained several long recombination tracts, recognized as regions of contiguous donor-specific alleles in the background of recipient alleles (also see preliminary studies). **FIG. 4** shows a typical recombination tract, where blue donor DNA has replaced recipient sequences. Some donor-specific indels were acquired as parts of larger donor segments (green box), but others were located at tract breakpoints (orange box). (**FIG. A-3** aligns **FIG. 4** to **FIGS. 5 & 6**, with a detailed legend.) This was the first genome-wide sequence analysis of bacterial recombination tracts; studies in *S. pneumoniae* and *Helicobacter pylori* have similarly found several long tracts, often gapped or including structural variation[45, 46]. Generating more quantitative information about how and why transformation varies across the chromosome is the focus of Aim 2.

Fig. 4



Strand segregation and mismatch repair (FIG. 2, STEP 5): Sequence differences also affect mismatch repair (MMR), with biases that have been only partially characterized[47]. The direct product of strand invasion is heteroduplex DNA, with one strand from the donor and one from the recipient; in the absence of repair, segregation of these strands at DNA replication will give each genotype to one of the two daughter cells. Although in principle repair could remove either strand (both have normal Dam methylation), MMR enzymes preferentially remove mismatched donor strands both during and after strand invasion[48]. Since strains defective in MMR are common in natural RT populations[49], in Aim 2 we propose to investigate the contribution of MMR to transformation events.

Free DNA in the RT: Transformation in the RT is of course dependent on the available DNAs. Even in healthy people, RT mucus contains very large amounts of DNA (~300 µg/ml), and in CF and COPD DNA becomes so concentrated that its viscosity hinders the flow of mucus[50, 51]. Although this DNA is primarily of human origin, the high concentrations of bacterial DNAs typical of biofilms are also expected in RT communities[16].

Summary: My laboratory is poised to develop a predictive model of *H. influenzae* transformational recombination in the RT (**FIG. A-1**), using our expertise in the molecular biology, bioinformatics, genomics, and evolution of natural competence and transformation. This model will take as input the genome sequences of DNAs and recipient strains, and use a detailed specification of the constraints on DNA uptake and recombination to calculate the probability of transformation at any point in the recipient genome. The result will allow researchers and clinicians to anticipate the spread of virulence genes through *H. influenzae* populations in respiratory tract communities.

SPECIFIC AIMS

1. **Predict DNA uptake.**
2. **Predict transformation.**
3. **Test the predictability of transformation in an *in vitro* RT system.**

RESEARCH PLAN

Overview: To further the long-term goal of anticipating gene transfer events in the human RT, we will characterize the critical sequence factors that affect the efficiency of transformation. These will be incorporated into a model that estimates the probability of transformation at any genome position and by any donor DNA. One known influence is the strong DNA uptake bias for the USS motif; Aim 1 will reevaluate this in a chromosomal context and test for other uptake biases that have been obscured by the strong signal from the USS. Uptake biases alone cannot be responsible for the equally strong short-scale transformation biases described below (PS-iii; **FIG. 6**), so Aim 2 will investigate the sequence biases of homologous recombination and mismatch repair, and if necessary those of translocation and degradation. The final step (Aim 3) will be a formal test of the model's predictions against the transformation frequencies seen in a simulated upper RT environment, aiming for a model that explains at least 80% of the observed variance in transformation frequency (TF). Below the general methods are described before our preliminary studies, since similar experiments and analyses are used throughout.

Uptake and transformation methods: Aims 1 and 2 will use various combinations of (1) simple DNA uptake and transformation experiments using standard laboratory methods and (2) deep sequencing analysis of complex DNA pools recovered from experiments. Competence and transformation: Except in Aim 3, cells will be grown in supplemented brain heart infusion broth (sBHI), made competent by transfer to the standard starvation medium M-IV, and incubated with transforming DNA for 15 minutes before DNA recovery or plating on selective agar **[E]**. Aim 3 transformations will use mixtures of mucin and DNAs like those in the RT. Uptake: Our periplasmic recovery process cleanly isolates DNA after uptake but before translocation **[D]**. Uptake assays will use defined-sequence DNA fragments labeled with ^{32}P or ^{33}P, and cells will be extensively washed with DNase I before radioactive counting.

Strains and DNAs: The recipient strain will usually be the standard Rd lab strain, sometimes carrying a *rec1, rec2 or mutS* knockout to block recombination, translocation and mismatch repair respectively; additional mutants can easily be constructed as needed by recombineering[31]. Some experiments will use the clinical strain PittDD[52]. Donor DNAs will usually come from clinical strains 86-028NP and PittGG **[B]**. Each donor genome will contain point mutations giving resistance to novobiocin (Nov^R) and nalidixic acid (Nal^R); selection for transformation at these loci eliminates non-competent cells without affecting transformation elsewhere on the chromosome **[A]**. Most transformation and DNA uptake assays will use either defined genetically marked or radiolabeled fragments, or *H. influenzae* chromosomal DNA. Chromosomal DNA preps have fragment sizes of 50-200 kb and before use will usually be restriction-digested or sheared (by BioRuptor, HydroShear or Covaris G-tubes).

Illumina sequencing methods: DNA will be processed using standard library construction protocols for paired-end Illumina sequencing of short (~200bp) DNA fragments[53]. Briefly, DNA pools will be sheared by sonication, and multiplexed libraries will be constructed using the Illumina TruSeq kit. Sequencing will be performed by the UBC Biodiversity Centre or the Michael Smith Genome Sciences Centre, depending on cost and turn-around time. Yields currently provide nearly 20,000-fold coverage of the small *H. influenzae* genome and are expected to rise steadily over time. Where the yield from a single Illumina lane exceeds the needs of the experiment, libraries will be multiplexed for more economical sequencing. The proposed sequencing is tabulated in the Budget Justification.

Analytical methods. Initial data processing will be done in Unix using standard reference-guided assembly methods (BWA, SamTools, GATK, and IGV) to map short reads to the recipient and donor reference sequences and to identify SNPs and SNP frequencies. Further analysis will use custom tools in the R statistical programming language; we have already developed some of these for **[A]** and **[D]**, and are developing others with our bioinformatics collaborator Ira Hall. In uptake experiments, the depth-of-coverage of each position will be compared between the input and recovered datasets; the high SNP density will allow correction for any contamination by the recipient chromosome. In transformation experiments, we will exclude sequencing errors that would otherwise be read as isolated donor bases by using only reads containing the 2 or more donor SNP alleles seen in true recombination tracts. This analysis will also identify reads containing putative recombination breakpoints. Other tools will exclude sequence alignment artifacts by comparing reciprocal alignment of individual read pairs to both donor and recipient references. Generalized linear models implemented in R will be used to disentangle the contributions of different sequence factors to uptake or transformation. (These models are essentially multivariate regressions, fitting potential explanatory variables to an observed dependent variable and measuring their fit and significance.) The results will be combined in the full predictive model, again implemented in R.

PRELIMINARY STUDIES (referred to below as 'PS')

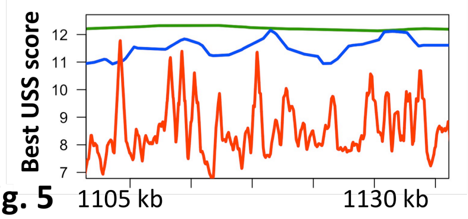
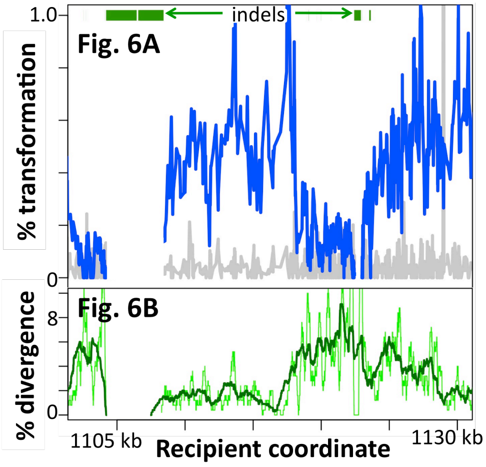


Fig. 5

i. **A model of DNA uptake:** One product of **[C]** was a USS-scoring matrix that evaluates how well a sequence fits the experimentally measured uptake bias. We used this matrix to score every position in the 86-028NP genome, and then used the scores to calculate uptake for fragments of various sizes. **FIG. 5** shows the same 30 kb interval shown in **FIG. 4**; the sharp peaks seen for 250 bp fragments (red) are smoothed for 2.5 kb and 6.5 kb fragments (blue and green, respectively), because these usually contain at least one high-scoring USS. This pilot model is the starting point for Aim 1.

ii. **Quality control:** The analyses below require meticulous tracking of the sequence variants that distinguish donor and recipient. We have improved the method used in **[A],** which combines whole-genome alignment of reference sequences with control sequencing of donor and recipient strains to generate 'gold-standard' sets of variants containing about 90% of all apparent SNPs and indels.

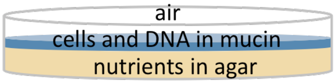
iii. **Expanded analysis of recombination tracts:** We have extended the work in **[A]** by identifying all the recombination tracts in 72 newly sequenced transformants. On average, clones had replaced 1% of their genomes, in 1-6 independent segments averaging 7.2 kb in length. Analysis of these tracts confirmed that they encompass regions of both high and low sequence identity and often end at indels. Some colonies consisted of two genotypes, one with and one without an unselected recombination tract, suggesting that heteroduplexes may be segregated without mismatch repair; this will be tested in Aim 2.



iv. **First genome-wide map of TF:** Sequencing enough clones to measure how transformation frequency varies across the chromosome would be prohibitively costly. Instead we sequenced a pool of 10,000 transformed Nov^R colonies and scored allele frequencies at the 37,601 gold-standard donor SNPs. The ~20,000-fold coverage resolved transformation events from sequencing errors (blue and grey lines, respectively, in **FIG. 6A**; same interval as above) and revealed dramatic and reproducible differences in TF over distances as short as 100 bp. This variation is partially correlated with % sequence divergence (compare **FIGS. 6A&B**), but there are many discrepancies. These are unlikely to be due to the USSs, since input fragments had a mean length of 6.5 kb; Aim 2 will seek the causes of this variation.

v. **Efficient transformation in a simulated RT environment:** We are developing growth and transformation conditions that simulate those in the RT, using mucin purified by extensive dialysis. Cells grow well, develop competence normally, and transform efficiently in liquid media containing 1% mucin. We are now transforming and selecting cells in a thin layer of mucin medium separating agar-solidified medium from air (**FIG. 7**), a much better model for RT conditions than standard submerged biofilms. **FIG. A-4** shows the efficient transformation obtained under these conditions.

Fig. 7



Aim 1 Predict DNA uptake.

Overview: The biases of DNA uptake determine what DNA is available for transformation, and thus the effects of other strains and species on the genetic potential of competent bacteria. Characterizing these biases is also a prerequisite for disentangling the effects of uptake and recombination. The uptake prediction map in **FIG. 5** is the first in a series of increasingly sophisticated models we will develop for DNA uptake. In **Stage I** this map will be compared with actual maps of DNA uptake efficiency, generated using Illumina sequencing of chromosomal DNA fragments recovered from the periplasm of competent cells. Differences between the predicted and observed uptake maps will be identified and the causes investigated. In **Stage II** more complex variables will be investigated with new experimental maps generated using different donor strains and species, different recipient strains, and complex mixtures of DNAs. This data will give a detailed robust uptake-prediction model for any *H. influenzae* strain and any input DNA.

Stage I. Investigating the effects of USS and other sequences: Uptake of a given SNP depends on the presence of USSs on the fragments it occurs in. The scoring matrix used to generate the uptake map in **FIG. 5** is based on data for a single 200 bp fragment with only 32 variable bp, but little is known about how fragment length and USS number affect uptake. We will therefore generate new uptake maps using chromosomal DNA preparations of four fragment sizes (0.25, 2.5, 6.5 and >20 kb). This will provide the uptake efficiency of each donor DNA position, but individual fragment ends will only be known for the 0.25 kb dataset, since larger fragments will have been further sheared before sequencing. Thus we will also test populations of fragments with precisely known ends generated by predigesting donor DNA with one of six restriction enzymes. Since these assays will need only modest sequence coverage, 12 input and recovered samples can be tested in parallel and combined in one lane of sequencing.

The *H. influenzae* genome contains about 1 USS per kb, so most 0.25 kb fragments will have 0 or 1 USS, while larger fragments will usually have several. Thus we expect to see large differences in uptake between different 0.25 kb fragments, but much less dramatic differences in the larger fragment datasets (**FIG. 5).** The 0.25 kb data will reveal (1) how uptake efficiency differs between single USS variants, (2) effects of any non-USS sequences that promote uptake, and (3) effects of any sequences that inhibit uptake (see also caveat). Differences between the predicted and observed maps for the large-fragment and restricted-DNA datasets will reveal effects of interactions between USS. The sequence and fragment-size factors identified in these whole-genome experiments will then be specifically tested in standard uptake experiments with defined fragments, allowing the effect of each factor to be measured without interference from others (as in Fig. 8 of **[C]** and Fig. 2 of **[D]**).

Example: The initial model assumes that a fragment's highest-scoring USS determines its uptake. If uptake peaks are found to be higher than predicted whenever two USSs are on the same fragment (restricted-DNA dataset) or close enough to be on the same fragment (non-restricted datasets), both USSs must contribute to uptake, so the model would be modified to include this interaction effect. The size of this effect would then be evaluated in uptake assays using radiolabeled fragments that differed only in how well their two USSs matched the consensus. The results would provide the values used by calculations in a more refined version of the model.

Stage II. Investigating uptake effects in RT communities: *H. influenzae* populations in the RT are typically diverse, so the model must be applicable to many different recipient and donor strains.

Uptake by a different *H influenzae* recipient strain: To test the critical expectation that strains do not differ in uptake specificity, the above experiments will be repeated using as recipient a *rec2* derivative of strain PittDD (chosen because of its high DNA uptake). Any deviations from the model's predictions can be further investigated using other recipients. If the level of DNA uptake does not limit the experiment's sensitivity, we will also test as recipient another strain with low uptake **[B]**.

Uptake of DNA from different donors: If the uptake model is robust, its predictions should apply to any donor DNA. We will test DNAs from each of 5 other *H. influenzae* strains (PittGG, PittDD. R2866, PittAA and Eagan) **[B]**, multiplex sequencing making this economical. If any differences are found, DNAs from other strains can be tested. Because other Pasteurellaceaen genomes are also enriched for the *H. influenzae* USS[54, 55], we will test DNAs of the respiratory pathogens *H. haemolyticus* and *H.*

parainfluenzae and of the oral pathogen *Actinobacillus actinomycetemcomitans*, as these are good surrogates for the diverse Pasteurellacean species in the RT. Finally, because RT environments contain abundant human DNA and other DNAs not enriched for any USS, we will test uptake of commercially available human DNA.

Uptake of DNA from complex mixtures: In RT microenvironments, *H. influenzae* DNA must compete with abundant human DNA as well as DNAs of other bacteria. USSs occur in these DNAs, but at the much lower densities predicted by their base compositions (*e.g.* the human genome contains ~1 USS per 30 Mbp). The final tests will evaluate three levels of competition (i) intraspecific, using a mixture of the five previously tested *H. influenzae* strains (ii) interspecific, using a mixture of these with DNAs from the three other Pasteurellaceae species, and (iii) interdomain, using a mixture of all these plus 90% human DNA. These will, respectively, test for (i) competition between variants present in different strains, (ii) the potential for horizontal transfer of virulence genes evolved in other Pasteurellacean species, and (iii) interference by host DNA.

Caveats: (1) If the observed uptake maps are substantially different from the predictions, unbiased motif-searches will be used to find unanticipated uptake motifs. (2) DNA sequences that inhibit DNA uptake may have been depleted from the *H. influenzae* genome over its many millions of years of biased DNA uptake **[D]**. They can also be sought in fragments of 'foreign' chromosomal DNA that have been ligated to a perfect-consensus USS. (3) If ethical issues preclude sequencing human DNA, we will instead use calf thymus DNA. (4) No analytical methods exist to identify the sources of DNA taken up from mixtures; these will need to be developed.

Outcomes: This work will give a robust predictive model of the DNA uptake pathway's biases, a prerequisite for inferring recombination biases from the transformation data analyzed in Aim 2. The results will have been validated with multiple donors and recipients. The immediate utility of this model is its ability to give the probability that any sequence variant from any pool, allowing it to be used to predict the spread of antibiotic resistances and other virulence determinants through *H. influenzae* populations. In the future, this could lead to clinical applications using specific DNAs that block genetic exchange or genetically modify the RT microbiota.

Aim 2. Predict transformation

Overview: The dramatic differences in transformation frequency seen in **FIG. 6A** indicate that alleles at different loci may spread through *H. influenzae* populations at very different rates, with obvious implications for the evolution of virulence determinants. This Aim will produce a quantitative model that predicts transformation frequency of any donor sequence into any recipient chromosome (**FIG. A-1**). This model will incorporate the uptake biases elucidated in Aim 1 as well as the effects of sequence divergence and other chromosomal properties measured here. This process will be ongoing - new quantitative relationships will be incorporated into the transformation model, whose predictions will then be re-tested against the various TF datasets. (Note that this is iterative but not circular.) **Stage I** will analyze the TF data of PS-iii and -iv and of a parallel experiment with unselected cells. This will refine the analytical methods and guide design of subsequent experiments. As in Aim 1, **Stage II** will then use different donor strains and species and different recipient strains to disentangle the many variables affecting transformation.

Stage I. Preliminary analyses: Analysis will begin with the relationship between TF and % nucleotide identity, quantifying the correlations seen between **FIGS 6A & 6B**. Since identity is sensitive to the length of the interval being compared, it will be scored over a wide range of window sizes, but especially the 80 bp suggested by RecA studies[43]. At this stage we will also begin the ongoing development of the more complex analytical methods that will be used below. Generalized linear models will analyze the correspondence between the observed TFs and other features of the 86-028NP donor and Rd recipient chromosomes, such as correlations of TF with proximity to indel/rearrangement differences of different sizes. Gibbs motif sampling will look for short sequence motifs strongly associated with high and low TFs; this will be especially valuable for chromosomal features whose locations are not conserved between strains.

At the same time a parallel TF dataset will be generated from unselected competent cells pooled just 2 hr after DNA uptake. This data will be free of artefacts due to unequal growth and survival of different 86-028NP-Rd recombinants, although less sensitive due to the inclusion of cells that did not take up any DNA. A *recA*$^-$ recipient control will detect unrecombined donor dsDNAs that persist in the periplasm; this control may also identify sequence features that block translocation (see below). If significant differences are seen in TFs measured before and after growth into colonies, the experiments described below will be done with (or replicated with) cells pooled without selection.

Stage II. Characterizing the sequence biases of transformation: Different donor DNAs: Factors affecting the efficiency of recombination fall into two classes. In the first are the many sequence differences between donor and recipient that affect the efficiency of strand invasion and the extent of MMR (densities and types of SNPs, indels, complex insertion/deletions and rearrangements). In the second are sequences shared by donor and recipient but varying along the chromosome (local base composition and nucleoid structure, sequence motifs, proximity to the origin of replication, and direction and frequency of transcription). A powerful way to separate these factors is comparison of donor DNAs from independently diverged strains. We will generate TF sequencing data (from pooled cells or from 100,000 selected colonies) for strains from 5 branches of the *H. influenzae* phylogeny (the same strains used in Aim 1-II) **[B]**. Generalized linear models will be applied to the combined data, measuring the amount of TF variance explained by each factor, and its significance. We will also obtain or create antibiotic-resistant strains of *H. haemolyticus, H. parainfluenzae* and *A. actinomycetemcomitans* and transform their DNA into *H. influenzae* strains, using selection and sequencing to investigate the constraints on recombination. These species' greater sequence divergence with *H. influenzae* will directly measure the probability of horizontal gene transfer from these pathogens to *H. influenzae*. Factors making significant contributions to the variation in TF will be validated by experiments using standard defined-fragment transformation assays.

Example: Preliminary analysis of the sequencing data might indicate that segments with 5 or more mismatches were associated with low TFs. A systematic reanalysis of these SNP-cluster segments would then be used to clarify the relationship, perhaps finding that TFs were particularly low when mismatches were tightly clustered rather than separated by strings of matches. Transformation assays with cloned *nov*^{R} fragments carrying dispersed or clustered mismatches (in the noncoding sequence just upstream of the *nov*^{R} point mutation) could then be used to measure the effect as differences in NovR TF.

Different recipients: The generality of the above results will be tested using different recipients, focusing on strains with different levels of DNA uptake or transformation (*e.g.* Eagan, PittAA **[B]**). An experiment that simply reverses the roles of donor and recipient will be especially powerful, since any differences in local TF would point to (i) different recombinational constraints for the RecA-coated donor strand and the base-paired recipient strands, and/or (ii) biochemical differences in the recombination machinery of the two strains. Finding no significant differences would eliminate these factors from consideration, focusing attention more strongly on the roles of sequence divergence and chromosomal features.

Mismatch repair: Strains defective in MMR can comprise up to 15% of *H. influenzae* isolates in cystic fibrosis patients[56]. Prevalence of such 'mutator' strains is usually attributed to selection for new mutations, but for competent bacteria they may also act to increase lateral gene transfer between divergent strains and species (in *E. coli*, MMR mutants have this phenotype). TF analysis in a *mutS* recipient will identify the extent to which the mismatch repair machinery constrains exchange to between highly similar sequences, and any sequence biases of this. While our *mutS* mutant transforms normally with singly marked DNA, heterologous DNAs have not been tested. Observed sequence biases in the activity of MMR will be incorporated into the model, discounted for the expected frequency of MMR-defective strains in natural populations[49].

Other biases: If the model's predictive power remains below the goal of explaining 80% of the variance in TF, we will look for biases in translocation and strand degradation. The *recA* control above will have identified any DNA sequences that remain double-stranded because they have not been translocated. This can be supplemented by a time course of translocation, exposing cells to DNA for only 5 min (stop by adding DNase I), and following the disappearance of sequences from the periplasmic fraction over the next hour. Direct analysis of translocated DNA strands will be more difficult since this requires the recovery of single-stranded DNAs from the cytoplasm before they can be degraded or recombined. We will optimize this procedure using an 8 kb fragment carrying the *nov*^{R} allele and a *recA*$^-$ recipient. Once optimized, the method could be applied to restricted 86-028NP chromosomal DNA fragments.

Caveats: Functional constrains on sequence divergence may complicate interpretation of the strain comparisons; new analytical methods will be needed to control for these. Because TFs are expected to be lower with *H. haemolyticus* and *A. actinomycetemcomitans* DNAs, transformants may need to be selected for Nov^R and Nal^R, with analysis focusing on these sections of the genome. We can also use recombineering[31] to insert selectable markers in any potentially informative chromosomal context in Rd or other *H. influenzae* strains.

Outcomes: Understanding why some genes recombine more than others will shed light on the different evolutionary histories of virulence genes. Identification of recombination hotspots and coldspots will both reveal evolutionary mechanisms and suggest new strategies for limiting harmful recombination.

Aim 3. Test the predictability of transformation in an *in vitro* RT system.

Rationale: To be useful, the transformation-prediction model developed in Aims 1 and 2 should reliably give the probability of clinically relevant recombination events in the RT. In **Stage I** we will refine our *in vitro* transformation system, adjusting culture conditions to closely mimic those of the RT. In **Stage II** the robustness of our model's predictions will be tested in this system.

Stage I. An *in vitro* system of RT transformation: Standard conditions for *H. influenzae* transformation are very unlike those in the RT, where *H. influenzae* lives in the mucus layer that separates epithelial cells from inspired air. We have developed a simple simulated RT system with efficient transformation of cells cultured in a layer of 1% mucin and DNA at the interface between air and agar-solidified brain-heart infusion (PS-v). The first step will be to modify the defined *H. influenzae* culture medium described by Coleman *et al.*[57] to more closely resemble RT conditions identified by analysis of bronchoalveolar lavage fluid[58]. We will use our detailed understanding of the regulation of competence to ensure that these conditions permit the development of competence and frequent transformation known to occur in the RT, testing for transformation by antibiotic-resistance alleles **[E]**. Depending on the results of Aims 1 and 2, the mucin layer will be supplemented with appropriate concentrations and mixtures of DNAs. Since high DNA concentrations can have chemical and biochemical effects independent of their information content, we will collaborate on this with Shawn Lewenza, who has extensive experience with culturing cells in the presence of free DNA[59, 60].

Stage II. Testing the model: The predictive model developed in Aims 1 and 2 will provide a probability that each sequence variant in the donor DNA will transform a recipient chromosome.The final step of this work will be to test the model's ability to predict genome-wide transformation frequencies under the RT conditions established above. Cells on the RT plates developed above will incubated with the same mixture of potential donor and human DNAs used in Aim 1-II. The RT system gives lower TFs than the standard starvation protocol; these are likely more representative of real-world conditions, but they lower the sensitivity of the sequencing analysis. If sequencing has become sufficiently sensitive, all the cells on the plate will be pooled after 6 hr incubation and their DNA sequenced. Otherwise transformants will be first selected for antibiotic resistance alleles provided in the genomes of potential donors.

Caveats: As in Aim 1, the analysis will be complicated by the mixture of donor DNAs. The expected low TFs make this experiment especially sensitive to contamination by unrecombined donor DNA. Recovered cells will be exhaustively washed with DNase I, and levels of contamination independently checked with tracer DNAs that contain USSs but cannot recombine.

Outcomes: This is the culmination of the whole project. The ability to culture and transform cells under simulated RT conditions will provide an economical and convenient alternative for some experiments currently using tissue culture or animal models.

SIGNIFICANCE AND FUTURE DIRECTIONS:

It is time to begin anticipating genetic exchange rather than just reacting to it.

My group has the unique combination of experimental, modeling and bioinformatics skills needed for this work. By designing the uptake and transformation experiments to provide data for the model, we avoid the difficulty of consolidating data from different laboratories that used different methods for different goals. The graduate students trained in this integrated approach will be valued additions to the pool of highly qualified personnel. The work is also highly feasible. Our publications and preliminary studies demonstrate our ability to analyze, and interpret the high volumes of genomic data they generate. Furthermore, the experiments are independent; problems with one would not affect success of the others.

In addition to publishing the results in appropriate journals, we will facilitate use of the model by creating an on-line resource where researchers and clinicians can input information about donor strains and potential recipients, obtaining a genome-wide map of predicted transformation frequencies.

Many clinically relevant applications are possible. *Anticipating the rise of resistant strains during antibiotic therapy:* Given a choice between two antibiotics, clinicians could choose to avoid those with resistance likely to arise by recombination between strains common in the healthy population. *Anticipating long-term consequences of vaccination programs:* Current vaccines target capsule antigens, but these are known to change by recombination. A model of recombination can be used to focus clinical and epidemiological studies to the variants most likely to arise. *Manipulating DNA uptake:* Uptake of beneficial genes could be promoted by introducing designed DNAs into the RT, and conversely, the ability to block DNA uptake in pathogens could prevent the spread of harmful alleles. *Controlling exchange events in personal microbiomes:* The plummeting costs of genome sequencing may soon allow sequencing of the microbiomes of individuals at risk of serious respiratory infections. The model would then permit genetic changes to be predicted and perhaps manipulated.

This work will provide a great deal of information about the biology of DNA uptake and transformation. The analyses we propose can readily be extended to other competent bacteria, first respiratory pathogens (*N. meningitidis, S. pneumoniae*), then others (*Vibro cholerae*, *Helicobacter pylori, Campylobacter jejuni, Staphylococcus aureus*). Most of the recombinational results are expected to apply to other species, with minor modifications, and differences in uptake effects can be investigated using the methods we have pioneered.

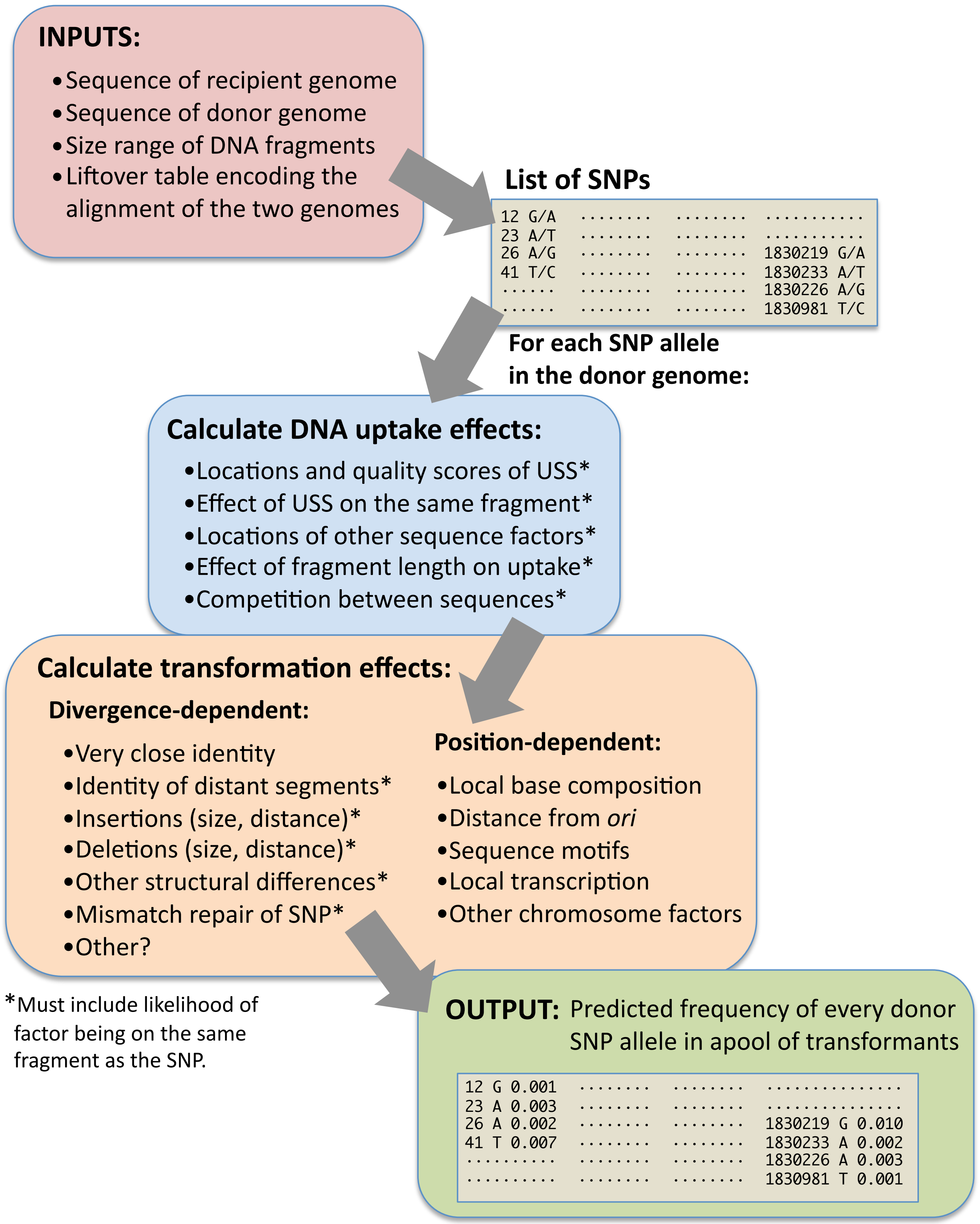
Genetic exchange is the biggest challenge to current prevention and therapeutic interventions, and our laboratory is the best prepared to begin the assault on it.

Appendix contents

FIGURES

REFERENCES

Figure A-1. Schematic of the predictive model



INPUTS:

- Sequence of recipient genome
- Sequence of donor genome
- Size range of DNA fragments
- Liftover table encoding the alignment of the two genomes

List of SNPs

```
12 G/A   ........   ........   ...........
23 A/T   ........   ........   ...........
26 A/G   ........   ........   1830219 G/A
41 T/C   ........   ........   1830233 A/T
......   ........   ........   1830226 A/G
......   ........   ........   1830981 T/C
```

For each SNP allele in the donor genome:

Calculate DNA uptake effects:

- Locations and quality scores of USS*
- Effect of USS on the same fragment*
- Locations of other sequence factors*
- Effect of fragment length on uptake*
- Competition between sequences*

Calculate transformation effects:

Divergence-dependent:

- Very close identity
- Identity of distant segments*
- Insertions (size, distance)*
- Deletions (size, distance)*
- Other structural differences*
- Mismatch repair of SNP*
- Other?

Position-dependent:

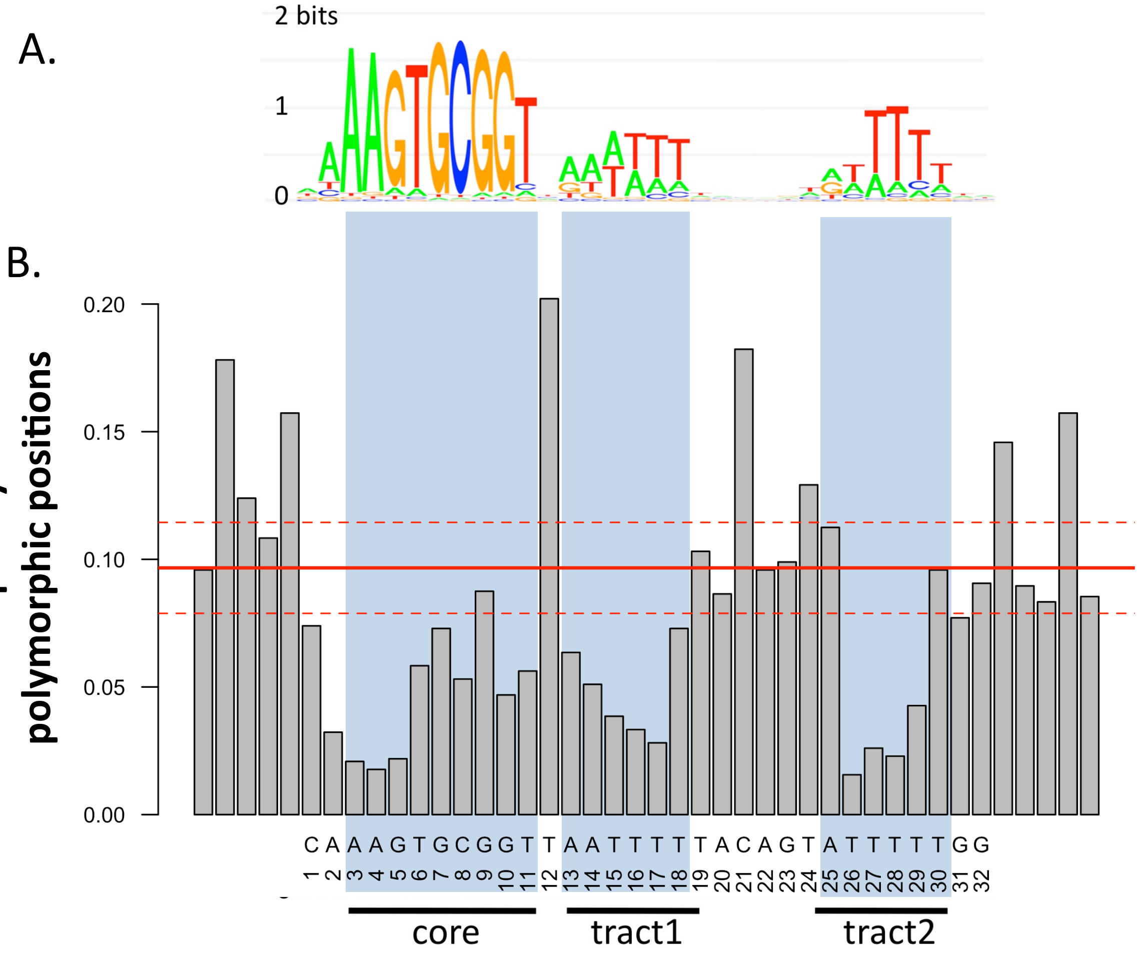
- Local base composition
- Distance from *ori*
- Sequence motifs
- Local transcription
- Other chromosome factors

OUTPUT: Predicted frequency of every donor SNP allele in apool of transformants

```
12 G 0.001   ........   ........   ...............
23 A 0.003   ........   ........   ...............
26 A 0.002   ........   ........   1830219 G 0.010
41 T 0.007   ........   ........   1830233 A 0.002
..........   ........   ........   1830226 A 0.003
..........   ........   ........   1830981 T 0.001
```

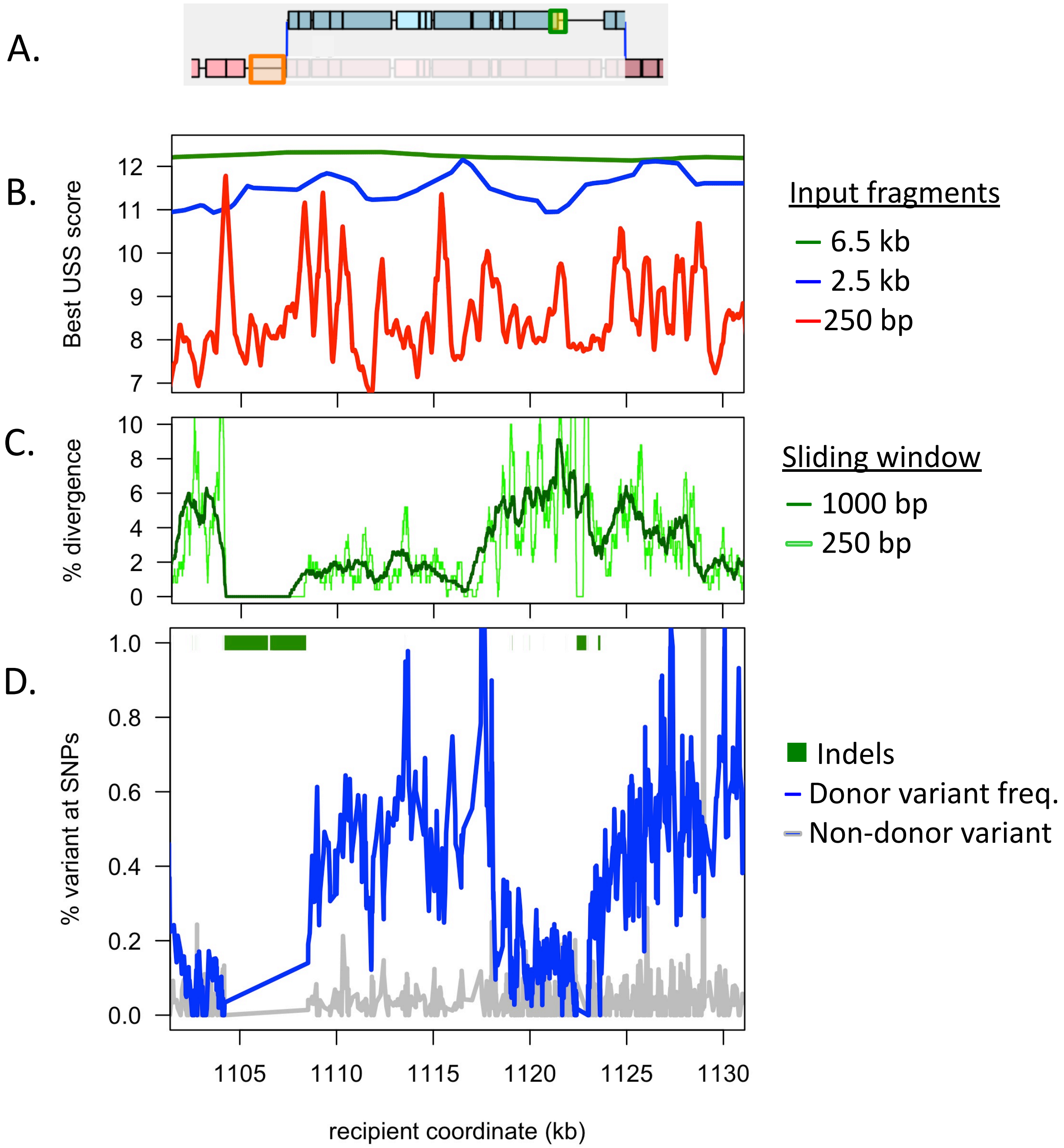
*Must include likelihood of factor being on the same fragment as the SNP.

Figure A-2: Decreased genetic divergence within USSs shows the significance of transformation in the RT environment



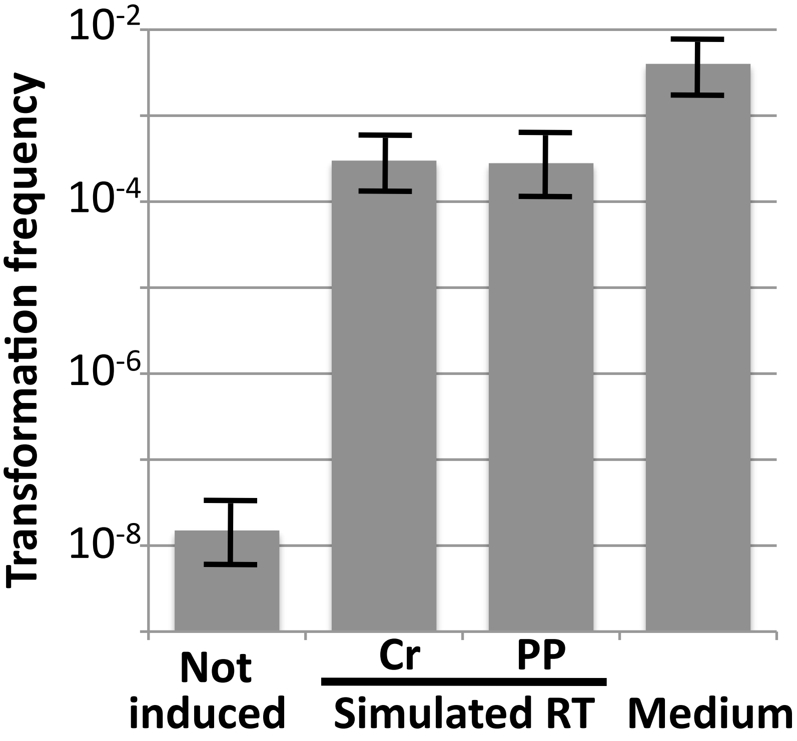
A. *Genomic USS motif.* Sequence logo of the 2205 genomic USSs in the *H. influenzae* KW20 genome. Other isolates have indistinguishable genomic USS motifs. **B.** *Sequence divergence is constrained within the USS core and flanking T-rich tracts.* Data was derived from the multiple alignment of 20 *H. influenzae* genome sequences and include 1020 ungapped 20 taxon alignments. First, anywhere the USS scoring matrix derived in **[B]** 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 A-3: Chromosomal variation in uptake and transformation



All parts of the figure show the same 30 kb interval of the chromosome. **A.** *Recombination tract in a single transformed clone*. Pink shows recipient backbone; blue shows 16.6 kb donor segment with 452 donor-specific SNPs. Orange and yellow boxes indicate recipient-specific and donor-specific insertions, respectively. **B.** *Predicted DNA uptake for 3 fragments sizes.* The sequence of 86-028NP wwas scored with the USS matrix from **[B]**. The scoring assumes that the highest scoring USS on a fragment dictates its probability of being taken up, hence the increasingly uniform uptake of larger fragments. **C.** *Divergence between donor and recipient.* Sliding windows of 250 and 1000 bp are used*.* **D.** *Transformation frequency*. Data from 10,000 pooled Rd colonies transformed with Nov^R 86-028NP DNA andsequenced to 20,000-fold coverage. Blue: frequency of donor-specific SNP alleles; grey: frequency at SNP positions of bases not in Rd or 86-028NP.

Figure A-4: Efficient transformation in a mucin layer



Efficient transformation in a simulated respiratory tract mucin layer. Competent cells were spread on 60 mm plates containing 10 ml of sBHI agar that had been pre-spread with 200 µl of 1% dialysed mucin (crude (Cr) or partially purified (PP)) and 1µg Nov^{R} donor DNA. Control plates with no added DNA had zero colonies, giving a limit of detection of <10^{-8} CFU/ml. The medium control is competent cells transformed in the liquid starvation medium used to induce competence.
)

REFERENCES

Papers A-E are provided in the Appendix. Other papers from my laboratory are indicated below by boldface.

[A] 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.

[B] Maughan, H., and Redfield, R.J. (2009). Tracing the evolution of competence in *Haemophilus influenzae*. *PLoS One* **4**, e5854.

[C] 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.

[D] 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*.

[E] Poje, G., and Redfield, R.J. (2003). Transformation of *Haemophilus influenzae*. *Methods Mol Med* **71**, 57-70.

Reference

1. *Marri, P.R., Paniscus, M., Weyand, N.J., Rendon, M.A., Calton, C.M., Hernandez, D.R., Higashi, D.L., Sodergren, E., Weinstock, G.M., Rounsley, S.D., et al. (2010). Genome sequencing reveals widespread virulence gene exchange among human Neisseria species. PLoS One **5**, e11835.*
2. *van Hoek, A.H., Mevius, D., Guerra, B., Mullany, P., Roberts, A.P., and Aarts, H.J. (2011). Acquired antibiotic resistance genes: an overview. Front Microbiol **2**, 203.*
3. *Kruger, N.J., and Stingl, K. (2011). Two steps away from novelty--principles of bacterial DNA uptake. Mol Microbiol **80**, 860-7.*
4. *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.*
5. *World Health Organization, (2012). The evolving threat of antimicrobial resistance - Options for action. In World Health Organization. (Geneva, Switzerland: World Health Organization), pp. 1-120.*
6. *Liao, M., Gu, W.M., Yang, Y., and Dillon, J.A. (2011). Analysis of mutations in multiple loci of Neisseria gonorrhoeae isolates reveals effects of PIB, PBP2 and MtrR on reduced susceptibility to ceftriaxone. J Antimicrob Chemother **66**, 1016-23.*
7. *Croucher, N.J., Harris, S.R., Fraser, C., Quail, M.A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J.H., Ko, K.S., et al. (2011). Rapid pneumococcal evolution in response to clinical interventions. Science **331**, 430-4.*
8. *Maragakis, L.L., Perencevich, E.N., and Cosgrove, S.E. (2008). Clinical and economic burden of antimicrobial resistance. Expert Rev Anti Infect Ther **6**, 751-63.*
9. *Nakamura, S., Shchepetov, M., Dalia, A.B., Clark, S.E., Murphy, T.F., Sethi, S., Gilsdorf, J.R., Smith, A.L., and Weiser, J.N. (2011). Molecular basis of increased serum resistance among pulmonary isolates of non-typeable Haemophilus influenzae. PLoS Pathog **7**, e1001247.*
10. *Clementi, C.F., and Murphy, T.F. (2011). Non-Typeable Haemophilus influenzae Invasion and Persistence in the Human Respiratory Tract. Front Cell Infect Microbiol **1**, 1.*
11. *Rijkers, G.T., Vermeer-de Bondt, P.E., Spanjaard, L., Breukels, M.A., and Sanders, E.A. (2003). Return of Haemophilus influenzae type b infections. Lancet **361**, 1563-4.*
12. *Schubert, S., Darlu, P., Clermont, O., Wieser, A., Magistro, G., Hoffmann, C., Weinert, K., Tenaillon, O., Matic, I., and Denamur, E. (2009). Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog **5**, e1000257.*
13. *Meier, P., and Wackernagel, W. (2003). Mechanisms of homology-facilitated illegitimate recombination for foreign DNA acquisition in transformable Pseudomonas stutzeri. Mol Microbiol **48**, 1107-18.*
14. *Domingues, S., Harms, K., Fricke, W.F., Johnsen, P.J., da Silva, G.J., and Nielsen, K.M. (2012). Natural Transformation Facilitates Transfer of Transposons, Integrons and Gene Cassettes between Bacterial Species. PLoS Pathog **8**, e1002837.*
15. *Murphy, T.F., Bakaletz, L.O., and Smeesters, P.R. (2009). Microbial interactions in the respiratory tract. Pediatr Infect Dis J **28**, S121-6.*

16. *Charlson, E.S., Bittinger, K., Haas, A.R., Fitzgerald, A.S., Frank, I., Yadav, A., Bushman, F.D., and Collman, R.G. (2011). Topographical continuity of bacterial populations in the healthy human respiratory tract. Am J Respir Crit Care Med **184**, 957-63.*
17. *Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. Science **326**, 1694-7.*
18. *Erb-Downward, J.R., Thompson, D.L., Han, M.K., Freeman, C.M., McCloskey, L., Schmidt, L.A., Young, V.B., Toews, G.B., Curtis, J.L., Sundaram, B., et al. (2011). Analysis of the lung microbiome in the "healthy" smoker and in COPD. PLoS One **6**, e16384.*
19. *Ulanova, M., and Tsang, R.S. (2009). Invasive Haemophilus influenzae disease: changing epidemiology and host-parasite interactions in the 21st century. Infect Genet Evol **9**, 594-605.*
20. *Hallstrom, T., and Riesbeck, K. (2010). Haemophilus influenzae and the complement system. Trends Microbiol **18**, 258-65.*
21. *Peltola, H. (2000). Worldwide Haemophilus influenzae type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates. Clin Microbiol Rev **13**, 302-17.*
22. *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.*
23. *Thanavala, Y., and Lugade, A.A. (2011). Role of nontypeable Haemophilus influenzae in otitis media and chronic obstructive pulmonary disease. Adv Otorhinolaryngol **72**, 170-5.*
24. *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.*
25. *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.*
26. *Farjo, R.S., Foxman, B., Patel, M.J., Zhang, L., Pettigrew, M.M., McCoy, S.I., Marrs, C.F., and Gilsdorf, J.R. (2004). Diversity and sharing of Haemophilus influenzae strains colonizing healthy children attending day-care centers. Pediatr Infect Dis J **23**, 41-6.*
27. *Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol **13**, R42.*
28. *Cody, A.J., Field, D., Feil, E.J., Stringer, S., Deadman, M.E., Tsolaki, A.G., Gratz, B., Bouchet, V., Goldstein, R., Hood, D.W., et al. (2003). High rates of recombination in otitis media isolates of non-typeable Haemophilus influenzae. Infect Genet Evol **3**, 57-66.*
29. ***Redfield, R.J. (1991). sxy-1, a Haemophilus influenzae mutation causing greatly enhanced spontaneous competence. J Bacteriol 173, 5612-8.***
30. ***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.***
31. ***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.***
32. ***Maughan, H., Sinha, S., Wilson, L., and Redfield, R.J. (2008). Pasteurellaceae: Biology, Genomics and Molecular Aspects, (Caister Academic Press).***
33. *Jurcisek, J.A., Bookwalter, J.E., Baker, B.D., Fernandez, S., Novotny, L.A., Munson, R.S., Jr., and Bakaletz, L.O. (2007). The PilA protein of non-typeable Haemophilus influenzae plays a role in biofilm formation, adherence to epithelial cells and colonization of the mammalian upper respiratory tract. Mol Microbiol **65**, 1288-99.*
34. ***Kristensen, B.M., Sinha, S., Boyce, J.D., Mell, J.C., and Redfield, R.J. Natural transformation in Gallibacterium anatis. submitted to Applied and Environmental Microbiology.***
35. *Smith, H.O., Gwinn, M.L., and Salzberg, S.L. (1999). DNA uptake signal sequences in naturally transformable bacteria. Res Microbiol **150**, 603-16.*
36. ***Maughan, H., Wilson, L.A., and Redfield, R.J. (2010). Bacterial DNA uptake sequences can accumulate by molecular drive alone. Genetics 186, 613-27.***
37. *Barany, F., Kahn, M.E., and Smith, H.O. (1983). Directional transport and integration of donor DNA in Haemophilus influenzae transformation. Proc Natl Acad Sci U S A **80**, 7274-8.*
38. *Pifer, M.L., and Smith, H.O. (1985). Processing of donor DNA during Haemophilus influenzae transformation: analysis using a model plasmid system. Proc Natl Acad Sci U S A **82**, 3731-5.*
39. *Kumar, G.A., Woodhall, M.R., Hood, D.W., Moxon, E.R., and Bayliss, C.D. (2008). RecJ, ExoI and RecG are required for genome maintenance but not for generation of genetic diversity by repeat-mediated phase variation in Haemophilus influenzae. Mutat Res **640**, 46-53.*

40. *Wilcox, K.W., and Smith, H.O. (1976). Mechanism of DNA degradation by the ATP-dependent DNase from Hemophilus influenzae Rd. J Biol Chem* ***251****, 6127-34.*

41. *Kowalczykowski, S.C., and Eggleston, A.K. (1994). Homologous pairing and DNA strand-exchange proteins. Annu Rev Biochem* ***63****, 991-1043.*

42. *Mani, P., Yadav, V.K., Das, S.K., and Chowdhury, S. (2009). Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. PLoS One* ***4****, e4399.*

43. *Mani, A., Braslavsky, I., Arbel-Goren, R., and Stavans, J. (2010). Caught in the act: the lifetime of synaptic intermediates during the search for homology on DNA. Nucleic Acids Res* ***38****, 2036-43.*

44. *Rajan, R., Wisler, J.W., and Bell, C.E. (2006). Probing the DNA sequence specificity of Escherichia coli RECA protein. Nucleic Acids Res* ***34****, 2463-71.*

45. *Hiller, N.L., Ahmed, A., Powell, E., Martin, D.P., Eutsey, R., Earl, J., Janto, B., Boissy, R.J., Hogg, J., Barbadora, K., et al. (2010). Generation of genic diversity among Streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog* ***6****, e1001108.*

46. *Dorer, M.S., Sessler, T.H., and Salama, N.R. (2011). Recombination and DNA repair in Helicobacter pylori. Annu Rev Microbiol* ***65****, 329-48.*

47. *Hawk, J.D., Stefanovic, L., Boyer, J.C., Petes, T.D., and Farber, R.A. (2005). Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. Proc Natl Acad Sci U S A* ***102****, 8639-43.*

48. *Radman, M., and Wagner, R. (1986). Mismatch repair in Escherichia coli. Annu Rev Genet* ***20****, 523-38.*

49. *Watson, M.E., Jr., Burns, J.L., and Smith, A.L. (2004). Hypermutable Haemophilus influenzae with mutations in mutS are found in cystic fibrosis sputum. Microbiology* ***150****, 2947-58.*

50. *Ratjen, F., Paul, K., van Koningsbruggen, S., Breitenstein, S., Rietschel, E., and Nikolaizik, W. (2005). DNA concentrations in BAL fluid of cystic fibrosis patients with early lung disease: influence of treatment with dornase alpha. Pediatr Pulmonol* ***39****, 1-4.*

51. *Han, M.K., Huang, Y.J., Lipuma, J.J., Boushey, H.A., Boucher, R.C., Cookson, W.O., Curtis, J.L., Erb-Downward, J., Lynch, S.V., Sethi, S., et al. (2012). Significance of the microbiome in obstructive lung disease. Thorax* ***67****, 456-63.*

52. Maughan, H., and Redfield, R.J. (2009). Extensive variation in natural competence in Haemophilus influenzae. Evolution 63, 1852-66.

53. *Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature* ***456****, 53-9.*

54. Findlay, W.A., and Redfield, R.J. (2009). Coevolution of DNA uptake sequences and bacterial proteomes. Genome Biol Evol 1, 45-55.

55. *Wang, Y., Orvis, J., Dyer, D., and Chen, C. (2006). Genomic distribution and functions of uptake signal sequences in Actinobacillus actinomycetemcomitans. Microbiology* ***152****, 3319-25.*

56. *Roman, F., Canton, R., Perez-Vazquez, M., Baquero, F., and 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****, 1450-9.*

57. *Coleman, H.N., Daines, D.A., Jarisch, J., and Smith, A.L. (2003). Chemically defined media for growth of Haemophilus influenzae strains. J Clin Microbiol* ***41****, 4408-10.*

58. *Davis, G.S., Giancola, M.S., Costanza, M.C., and Low, R.B. (1982). Analyses of sequential bronchoalveolar lavage samples from healthy human volunteers. Am Rev Respir Dis* ***126****, 611-6.*

59. *Mulcahy, H., Charron-Mazenod, L., and Lewenza, S. (2008). Extracellular DNA chelates cations and induces antibiotic resistance in Pseudomonas aeruginosa biofilms. PLoS Pathog* ***4****, e1000213.*

60. *Mulcahy, H., Charron-Mazenod, L., and Lewenza, S. (2010). Pseudomonas aeruginosa produces an extracellular deoxyribonuclease that is required for utilization of DNA as a nutrient source. Environ Microbiol* ***12****, 1621-9.*

My laboratory studies natural competence and transformation, mainly in the bacterium *Haemophilus influenzae*. We have achieved our current grant's objectives on the regulation of competence, and are now focusing on its consequences and their clinical impact. By dissecting the different steps that affect the probability of transformation, we aim to better predict and therefore prevent its outcomes.

PROGRESS TOWARDS THIS PROPOSAL'S AIMS

The following published studies provide a foundation for the work we propose (see CV publication list):

1. *Genomic uptake sequences:* Analysis of genomic uptake sequences identified two Pasteurellacean clades and found no obvious differences between *H. influenzae* and its close relatives (Redfield 2006). Additional analysis of uptake sequences showed selection for minimal interference with protein coding, and evidence of ongoing impact of biased DNA uptake on genome and proteome evolution (Findlay 2009). A model of uptake sequence evolution (Maughan 2010) confirmed that they accumulate passively as an indirect result of uptake bias, and identified the parameters underlying this accumulation.
2. *Uptake bias:* Maughan (2010) also measured actual uptake specificity by comparing uptake of individual synthetic DNA constructs. We have now developed a powerful deep-sequencing method to characterise uptake bias at high resolution (Mell 2012); this will be extensively used in the new work.
3. *Recombination:* We have used deep sequencing to fully characterize recombination tracts, thus measuring the genome-wide consequences of competence (Mell 2011).
4. *Population biology of natural competence:* We characterized extensive variation in competence development, DNA uptake and transformability in 25 disease-causing strains of *H. influenzae* (Maughan 2009), and in two other pathogenic Pasteurellaceae (Bosse 2009, Kristensen 2012).
5. *Mutant construction:* We have improved a 'recombineering' protocol for mutant construction, allowing us to place selectable markers at any point in the chromosome for studies of recombination (Sinha 2012b) and to make clean knockout mutants for this new work (*rec1, rec2, mutS*).
6. *Tools for sequence analysis:* We have developed powerful new analytical tools and strategies for deep-sequencing analysis of DNA uptake and transformation, detailed in Mell 2011 and Mell 2012.

Together these studies show the feasibility of our current proposal. We have also generated extensive unpublished data pertinent to this proposal (described in more detail there): chacterization of a genome-wide transformation frequency map, analysis of 92 additional transformants following the work of Mell 2011, and the development of a transformation system that more closely mimics *in vivo* conditions.

PROGRESS UNDER THE CURRENT OPERATING GRANT: Regulation of CRP-S promoters in *H. influenzae* and *E. coli* (5 yr, $127,567, expires Sept. 2012)

Five published papers together address all three of this project's questions:

1. *How is sxy regulated in H. influenzae?* We have characterized both transcriptional and translational regulation (Cameron 2007, 2008a, 2008b). In addition, we are now finishing work showing that purine depletion triggers Sxy translation (Sinha and Redfield, in prep).
2. *How is sxy regulated in E. coli?* Initial characterization of competence and its regulation in *E. coli* showed that, like *H. influenzae*, *E. coli* has a functional CRP-S regulon controlled by Sxy and CRP (Sinha 2009), and that its genes encode a weakly functional DNA uptake machinery (Sinha 2012a).
3. *How does Sxy activate transcription in H. influenzae and E. coli?* Complementation tests showed that Sxy's function is broadly conserved between these divergent species (Cameron 2006, 2008b, Sinha 2009), and work in both *H. influenzae* and *E. coli* confirmed that physical interactions between Sxy and CRP are essential for transcriptional activation of CRP-S genes (Sinha 2009).

Two additional papers extend our results to other species: Bosse (2009) showed that Sxy also regulates competence in *Actinobacillus pleuropneumoniae*, and Kristensen (2012) characterized the competence of *Gallibacterium anatis* and developed a transformation protocol for this previously untransformable species.

Responses to Reviews

The proposal we are submitting is not a revision of our unsuccessful proposal to investigate the mechanism of DNA uptake, but a completely new proposal to identify the factors that determine which DNA sequences are taken up and recombined into recipient genomes. Although many of the reviewers' comments on the previous proposal are no longer applicable, we have taken advantage of as much of their advice as possible.

- The new proposal includes an explicitly stated and highlighted hypothesis.

- The Progress Report now separately describes results relevant to the aims of this new proposal and results that directly address the goals of our previously funded proposal on the regulation of competence in *H. influenzae* and *E. coli*.

- All components of the new proposal are areas for which we have demonstrated expertise and preliminary data. (We had less experience with investigating the macromolecular transport mechanisms that were the focus of the previous proposal.)

- This new proposal more directly addresses clinical goals, whereas the previous proposal had limited clinical relevance.