Introduction and Objectives:

My research program asks why bacteria are naturally competent (able to take up DNA fragments from their surroundings). Answering this question will determine whether competence is a valid model for the evolution of sex and, if it is not, will direct evolution-of-sex research to such eukaryote-specific factors as genome size, ploidy and endosymbionts.

Like sex, natural competence often causes genetic recombination between relatives ('transformation'), and it is thought to have evolved for the same function. Over the past 20 years my laboratory has been testing this assumption, mainly by using studies of gene regulation to reveal how selection has acted on competence. DNA uptake is tightly regulated in most bacteria, and we reasoned that studies of competence regulation in the laboratory would reveal selection's actions in the natural environment. Working with the human commensal and pathogen *Haemophilus influenzae*, we have shown that DNA uptake genes are turned on when cells lack their preferred sugars and nucleotides. **These results suggest that competence has evolved not for recombination but because cells obtain nutrients from DNA** (nucleotides, sugar, P, N, C). Further support comes from studies of regulation in other species and from our simulations, which showed that, because transformation recombines DNA from dead cells, it is selected only under conditions even more restrictive than those favouring meiotic sex[1, 2].

Here we focus on the other aspect of competence that may reflect selection for recombination - the 'self-specific' DNA uptake biases seen in the genus *Neisseria* and family Pasteurellaceae (including *H. influenzae*). The DNA uptake machinery of these bacteria preferentially binds to short 'uptake sequences' common in the species' own genomes and those of close relatives. By causing conspecific DNA to be preferentially taken up from DNA mixtures, it acts like a mate-recognition mechanism. We will evaluate the alternative explanation that self-specificity evolves due to intrinsic physical constraints of the uptake machinery, using Illumina sequencing to directly characterize DNA uptake biases.

Specifically, for each of 6 diverse species we propose to:

1. **Sensitively detect any self-specificity.**
2. **Thoroughly characterize any sequence biases of the DNA uptake machinery.**
3. **Identify evolutionary forces leading to self-specificity.**

Recent Progress: I emphasize our recent work on uptake bias and specificity, also citing some of our earlier work. Citations to my group's papers are in boldface; PDFs of papers **F**, **I**, **K** and **Q** are provided.

Properties of uptake sequences: Our sequence comparisons showed that uptake sequences are not mobile elements; they have evolved *in situ* as motifs shared by many related species [**Q**]. We also showed that the canonical Pasteurellacean and Neisserial uptake sequences are not preferentially found at sites where recombination might be beneficial but in genomic sites where they least interfere with coding and other genomic functions [**K**]. A later paper used simulation modeling to show that biased uptake of homologous DNA causes uptake sequences to accumulate without selection for recombination, and introduced more nuanced analysis using the Gibbs motif sampler [**I**].

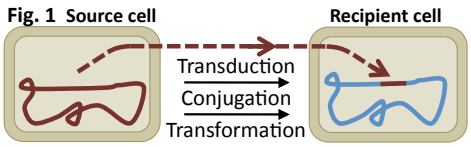
Evolutionary and phylogenetic context of competence and uptake bias: The distributions of competence genes, self-specificity and uptake sequences in 8 Pasteurellacean species indicate that their common ancestor was not only competent but had self-specificity like that of *H. influenzae* [**Q**]. This shows that uptake sequences are not species tags. A separate analysis of the basal Pasteurellacean *Gallibacterium anatis* found strong uptake specificity despite few recognizable uptake sequences [**E**].

3. Uptake bias of *H. influenzae*: We have developed a novel deep sequencing approach that assays uptake specificity by comparing datasets of degenerate uptake sequences from very large pools of DNA fragments before and after uptake by competent cells [**F**]. For this we developed novel computational methods, including an analytical solution that accounts for sequencing errors and input biases, generates correct motif models of uptake specificity, and measures positional dependencies between different parts of the uptake motif. This new method makes possible the uptake-bias experiments proposed below.

Literature Review:

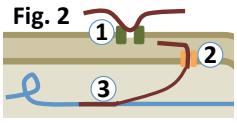
The function of meiotic sex is still not understood. Most eukaryotic life cycles include 2 components that together constitute sexual reproduction: syngamy, the fusion of two haploid cells, and meiosis, a cell division that generates haploid cells from a diploid cell[3, 4]. For the past 50 years evolutionary biologists have struggled to show how this meiotic sex can be adaptive; we know that it randomizes combinations of alleles over the generations, but not why this is beneficial. The many competing hypotheses include more efficiently eliminating harmful mutations, separating beneficial mutations from harmful ones, saving small populations from mutation catastrophe and moving populations between unstable fitness peaks in an adaptive landscape[5, 6]. Each benefit overcomes the cost of sex under some conditions, but none applies over the very wide range of conditions where sex has been successful: facultative or obligate sexual reproduction in haploid or diploid single-celled or multicellular organisms[3].

What can be learned from bacteria? Although meiotic sex occurs only in eukaryotes, alleles in bacteria can be recombined by three DNA-transfer processes: transduction, conjugation and transformation (**FIG. 1**). These parasexual processes are widely viewed as functional analogs of meiotic sex, since most microbiologists are unaware of the evolutionary problems posed by sex (see [6] for an exception). The success of strains that have acquired beneficial new alleles is often cited as evidence for this view, as are demonstrations that recombination by parasexual processes can increase fitness in laboratory cultures[7, 8]. However laboratory experiments must use conditions unlike those in natural bacterial environments, and they cannot tell us how selection did in fact act. Fortunately, insight into the past action of selection can come from close examination of the relevant molecular mechanisms, particularly by showing how specific genes benefit from the events they promote. Below I first describe evidence for non-sexual functions of the physical recombination machinery and of transduction and conjugation, and then consider transformation in more detail.



Bacterial *rec* genes are genes for DNA replication and repair, not genetic recombination. All cells contain proteins that break, anneal and rejoin complementary DNA strands[9], but this physical recombination becomes genetic recombination only if one of the DNA strands carries a variant allele from another cell. These *rec* proteins have vital functions in DNA replication and repair, especially restoring stalled replication forks and providing templates for repair of otherwise-lethal DNA lesions. However they show no evidence of selection for genetic recombination, and molecular biologists now agree that their recombination effects are fully explained by selection for efficient replication and repair[10, 11]. Thus evidence of selection for recombination must be sought in the processes that transfer DNA between cells.

Transduction and conjugational recombination are caused by phage and plasmid infection. The evolutionary success of plasmids and phages depends on transferring their DNA into new hosts[12], but the DNA transfer processes they encode can also move fragments of chromosomal DNA from one cell to another[13]. The ensuing recombination sometimes increases the cell's fitness, but the genes responsible for DNA transfer show no evidence of selection for chromosomal transfer[14-16].



This leaves only transformation - does it exist for genetic recombination? Naturally competent species encode protein machinery that brings DNA from the environment into the cell (**FIG. 2**, steps 1 and 2), causing transformation if recombination with this DNA changes the genotype (**FIG. 2**, step 3)[17]. This DNA uptake is commonly assumed to have evolved and be maintained by selection for transformation, although our modeling work has shown that the recombinational benefits of DNA uptake are even smaller and more elusive than those of meiotic sex[1, 2]. Other theoretical and experimental work has found conditions allowing competence genes to be maintained, for example if selection is episodic and the expression of competence fluctuates[18-23].

DNA uptake is not a byproduct of infection or another cellular process. Two more direct uses of incoming DNA have been proposed, as templates for repair of DNA lesions[20], and as a source of nutrients (free DNA is abundant in most natural environments)[24]. Our comprehensive characterization of the *H. influenzae* genes and their regulatory mechanisms has shown that competence is induced not by DNA damage but by molecular signals of depleted energy resources and nucleotide pools [**O-P**, **R**][25]. Competence genes are regulated by similar signals in related groups [**B**, **E**, **J**, **N**][26, 27], and by less well characterized nutritional signals in more distant groups[28]. Although these findings support the nutrient hypothesis, most reviews of natural competence reject it, citing the self-specificity of DNA uptake as a mate-recognition mechanism to enhance recombination[6, 28].

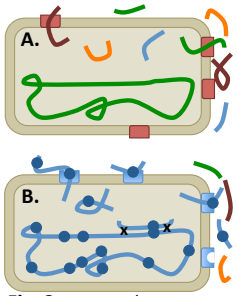


Fig. 3 ● = uptake sequence

Self-specificity. Most bacteria will take up DNA from any species (**FIG. 3A**), but as mentioned above, bacteria in the family Pasteurellaceae or the genus *Neisseria* preferentially take up DNA fragments from their own and closely-related species (**FIG. 3B**). This self-specificity occurs because they favour fragments containing short uptake sequences (blue dots in **FIG. 3B**) that are very abundant in their own genomes (*e.g.* AAGTGCGGT in *H. influenzae* and GCCGTCTGAA in *Neisseria sp* [28, 29, 33, 34]. Evolution by natural selection for recombination would require both selection for biased uptake proteins and an extreme form of altruism, since uptake sequences only promote recombination after the cell carrying them is dead; factors addressed in complex models by Chu *et al* [32]. This problem becomes much easier to address when the two components are considered separately.

Evolution of uptake sequences: I chose *H. influenzae* as a research organism partly because of its uptake sequences, since these show the importance of DNA uptake in its natural environment, the human respiratory tract. The uptake sequences of Pasteurellaceae and *Neisseria sp*. are different, but other common features suggest that the same evolutionary forces are responsible **[I,Q]**[29]. The model described in **[I]** shows that uptake sequences can be an unselected consequence of uptake bias, provided DNA of close relatives is available and sometimes recombines with chromosomal alleles. These conditions will usually be met, since homologous DNA is common in most microbial environments and proteins causing recombination are ubiquitous[35]. This process can fully explain accumulation of uptake sequences in species with biased DNA uptake, independent of any selection for recombination.

Molecular biology of uptake bias: Until recently uptake bias could only be studied by laborious assays comparing the uptake of defined fragments containing canonical, singly-mismatched or randomized uptake sequences [**I**][30, 31]. These showed uptake biases of >100-fold for fragments containing the canonical sequences, with a single sequence being sufficient for uptake. In contrast, a single experiment using our new uptake-recover-sequence (U-R-S) method (see Recent Progress) identified the effect of each of the 4 bases at each of 32 positions in *H. influenzae*'s full uptake motif. It also revealed unsuspected effects of interactions between positions, and differences between uptake bias and the genomic uptake sequences that need further investigation (Figs. 4 and 6 in **[I]**).

Evolution of uptake bias: All DNA-binding proteins have sequence biases. These inevitably arise from the need for close contacts between amino acid residues and the double helix. Strong biases are expected whenever tight binding or force transduction are needed, as is the case for the proteins that initiate uptake since DNA must be deformed to fit through the outer membrane pore (step 1 in **FIG. 2**)[36].

We hypothesize that uptake bias is due to physical interactions between incoming DNA and proteins of the uptake machinery, not to selection for recombination.

This hypothesis predicts that uptake biases will the norm for naturally competent bacteria. Tests for uptake bias have only been done for the few species where competition assays have found strong self-specificity; thus biases with no accompanying genomic enrichment may remain undetected. In addition, some bacteria show strong self-specificity but lack any genomic repeats resembling uptake sequences[28]. This may be due to weaker bias for simple motifs that would rarely cause conspecific uptake. Below we test this hypothesis by a broad survey of uptake bias and self-specificity.

Methodology

Project	Species	Family, Class	Collaborator	Upt-seq.?	Self-spec?	%G+C	Refs
Both	*H. influenzae* (control)	Pasteurellaceae, γ-proteobacteria	none needed	~1/kb	Yes	38	37
A.	*Acinetobacter baylyi*	Moraxellaceae, γ-proteobacteria	Averhoff (Germany)	No	No	40	41
	Thermus thermophilus	Thermales, Deinococci		No	No	68	41
	Vibrio cholerae	Vibrionaceae, γ-proteobacteria	Blokesch (Switzerland)	No	No	48	26
	Pseudomonas stutzeri	Pseudomonadaceae, γ-proteobacteria	Baltrus (USA)	No	No	64	42
B.	*Gallibacterium anatis*	Pasteurellaceae, γ-proteobacteria	none needed	~1/25 kb	Yes	40	[E]
	Campylobacter jejuni	Campylobacteriaceae, ε-proteobacteria	Gaynor (UBC)	No	Yes	30	40

Overview: We will test four species reported to have no self-specificity in competition assays and no genomic uptake sequences (PhD project A), and two with self-specificity not explained by uptake sequences (PhD project B) (see Table). None of these species have been directly tested for uptake bias. Both projects will begin by using U-R-S experiments to characterize even weak uptake bias and self-specificity in each species, and then use these rich datasets as the foundations for further investigations. **If our hypothesis is correct, all species will have uptake bias even in the absence of self-specificity.** (Timelines for both projects are provided with the Budget Justification, and their value to training is detailed in the HQP plan.)

Standard methods:

Preparation: For all species, assays using radiolabeled DNA will optimize DNA concentrations and incubation times for DNA uptake and recovery [**F**]. To prevent DNA degradation, cells will carry *rec2* mutations (available for most species; we will construct the *G. anatis* mutant [**B**]), and will be made competent by standard species-specific procedures. Both students will spend time in collaborating labs to learn species-specific details of culture, competence and DNA uptake. Input DNA and cells: Self-specificity assays will use an equal mixture of chromosomal DNAs (~10 kb fragments) from each of the 7 species in Table 1. Uptake bias assays will use a synthetic 200 bp test fragment with a fully randomized 60 bp central segment flanked by Illumina sequencing tags [**F**]; every fragment in this input pool is expected to be unique. DNA uptake and recovery: DNA remaining outside the cells after incubation will be degraded using DNase I, and fragments that have been taken up will be recovered intact by organic extractions [**F**]. Sequencing input and taken-up DNA: Chromosomal DNAs will be fragmented to ~250 bp before ligation of Illumina tags and sequencing. We expect $\geq 2 \times 10^7$ reads/sample (one MiSeq lane or 10 multiplexed HiSeq samples).

Initial analyses Self-specificity experiments: Reads of chromosomal sequences will be aligned to reference genomes of each species in the mixture; the species are not closely related so assignment to the correct source will be straightforward. Self-specificity will be measured by overrepresentation of conspecific sequences in the recovered-DNA dataset relative to the input dataset. Defining a biologically significant threshold for self-specificity is difficult, so we will set a conservative threshold of 10% over-representation of conspecific DNA (1000-fold below that expected for *H. influenzae*). The distribution of reads around each genome will then be analyzed to detect preferential uptake of specific sequences, as this would be informative in its own right and also provide guidance for analysis of the uptake-bias. Uptake-bias experiments: For the degenerate-fragment pools, each recovered dataset will initially be checked against the input for differences in base composition and in the frequencies of all 3-mers and higher sequence strings, and any patterns will be used to guide the subsequent analyses. For example, if DNAs recovered from *V. cholerae* are enriched for AT-tracts like those that flank the *H. influenzae* uptake sequence[37], later motif searches will be given these as a 'prior'.

Species with little or no confirmed self-specificity: (likely the four species of project A). Any uptake bias strong enough to be readily detected in the chromosomal-distribution or n-mer analyses will be confirmed by simple uptake experiments using defined radiolabeled fragments, and further refined by a U-R-S experiment using a degenerate segment biased towards the preferred sequence/motif (as in [**F**]).

Detecting weaker biases is likely to require development of new analytical strategies and methods (past experience predicts that this will be non-trivial). Motif searches will begin with the Gibbs recursive sampler[38], and will be guided by methods used to find transcription-factor binding motifs[39]. The sensitivity of the analysis will greatly exceed the biologically significant level—to be conservative we will ask for at least a two-fold enrichment of some sequence or motif. The genomic distributions of any uptake bias motifs will be mapped and compared to random expectations. (Components of this analysis will provide suitable projects for undergrads with computer skills.)

Species with confirmed self-specificity: This will likely be the two species of Project B, but similar analyses will be done for any Project A species that show clear self-specificity. The first step will be to repeat the U-R-S experiments using short fragments (250-500 bp) of each species' chromosomal DNA (separately, not a mixture), to detect with higher resolution whether some chromosome positions are taken up better than others.

G. anatis exhibits strong self-specificity but few of the expected Pasteurellacean uptake sequences [**E**]. If its uptake bias for the fully-degenerate fragment is found to be indistinguishable from that of the *H. influenzae* control, we will also test with the 24% degenerate *H. influenzae*-based fragment used in [**F**].

The molecular cause of *C. jejuni*'s documented self-specificity is not known[40]. One possibility is a sequence bias and genome enrichment of a motif too simple or complex to be detected as a typical uptake sequence. We may or may not see preferential uptake of some segments of the genome, depending on how widely separated the recognition elements are.

Caveats: If recovery of DNA is very low, this increases the risk of contamination by randomized input DNA, obscuring biases, but our control will provide a rigorous test of this, since *H. influenzae* is expected to have only very low levels of uptake from randomized fragments. Some biases may be overlooked if the preliminary data analyses do not provide any guidance for the motif searches; for example without prior knowledge that AT-tracts are enriched, motif searches might not be designed to look for phasing of these. If the *G. anatis* and/or *C. jejuni* preference is just for simple AT-rich motifs, we would see correlation of uptake with base composition in both U-R-S experiments. Finding self-specificity with no detectable uptake bias (perhaps in *C. jejuni*) would prompt an analysis of possible biases favouring DNA with specific methylation patterns or other modification.

Outcomes: (i) Finding uptake bias in the absence of self-specificity for most or all Project A species would support our hypothesis, and also our explanation for the self-specificity of the Pasturellaceae and Neisserias. It would indicate that most bacteria could promote self-uptake but do not, strengthening the more general hypothesis that recombination is an unselected consequence of DNA uptake. (ii) On the other hand, finding that uptake can occur with no detectable bias would reject our hypothesis that biases are inevitable but would not reject the broader hypothesis that bacteria have no functional equivalent to meiotic sex. (iii) Finding that uptake bias is always accompanied by some degree of self-specificity would confirm the predictions of the simulations in [**I**] but would not otherwise test either hypothesis.

Impact

For our research program: Outcome (i) would encourage us to set aside further work on uptake specificity and focus instead on non-genetic consequences of DNA uptake, especially its role in providing cells with nucleotides. Outcomes (ii) and (iii) would focus our attention on the features that distinguish clades with and without bias. We would also broaden our investigations by developing a model that integrates the evolutionary forces acting on DNA uptake (recombination, repair and nutrition).

For evolution-of-sex research: Outcome (i) would direct evolution-of-sex research to eukaryote-specific factors such as small population sizes and alternation of generations. Outcomes (ii) and (iii) would raise the possibility that some (but not all) groups of bacteria use DNA uptake for recombination.