Title: Adaptation in a fibronectin binding autolysin of *Staphylococcus saprophyticus*

Authors: Tatum D. Mortimer[1,2], Douglas S. Annis[3], Mary B. O'Neill[1,4], Lindsey L. Bohr[1,2], Tracy M. Smith[1,5], Hendrik N. Poinar[6,7,8,9], Deane F. Mosher[3], Caitlin S. Pepperell[1,5]

Affiliations:

[1]Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, United States

[2]Microbiology Doctoral Training Program, University of Wisconsin-Madison, Madison, United States

[3]Department of Biomolecular Chemistry, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA

[4]Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin, United States

[5]Department of Medicine (Infectious Diseases), School of Medicine and Public Health, University of Wisconsin-Madison, Madison, United States

[6]McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, Hamilton, Canada

[7]Department of Biology, McMaster University, Hamilton, Canada

[8]Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton,
Canada

[9]Humans and the Microbiome Program, Canadian Institute for Advanced Research, Toronto, Canada

Abstract:

Human-pathogenic bacteria are found in a variety of niches, including free-living, zoonotic, and microbiome environments. Identifying bacterial adaptions that enable invasive disease is an important means of gaining insight into the molecular basis of pathogenesis and understanding pathogen emergence. *Staphylococcus saprophyticus*, a leading cause of urinary tract infections, can be found in the environment, food, animals, and the human microbiome. We identified a selective sweep in the gene encoding the Aas adhesin, a key virulence factor that binds host fibronectin. We hypothesize that the mutation under selection (*aas*_2206A>C) facilitates colonization of the urinary tract, an environment where bacteria are subject to strong shearing forces. The mutation appears to have enabled emergence and expansion of a human pathogenic lineage of *S. saprophyticus*. These results demonstrate the power of evolutionary genomic approaches in discovering the genetic basis of virulence and emphasize the pleiotropy and adaptability of bacteria occupying diverse niches.

Importance:

Staphylococcus saprophyticus is an important cause of urinary tract infections (UTI) in women, which are common, can be severe, and are associated with significant impacts to public health. In addition to being a cause of human UTI, *S. saprophyticus* can be found in the environment, in food, and associated with animals. After discovering that UTI strains of *S. saprophyticus* are for the most part closely related to each other, we sought to determine whether these strains are specially adapted to cause disease in humans. We found evidence suggesting that a mutation in the gene *aas* is advantageous in the context of human infection. We hypothesize that the mutation allows *S. saprophyticus* to survive better in the human urinary tract. These results show how bacteria found in the environment can evolve to cause disease.

Introduction.

Urinary tract infections (UTI) are a global health problem of major significance, with an estimated annual incidence of 150-250 million and a lifetime risk of 50% among women (1–3). The associated costs to individuals and health care systems are substantial, with recent estimates from the United States numbering in the billions per year (4). UTIs are also associated with severe complications such as pyelonephritis, sepsis and premature labor (4). *Staphylococcus saprophyticus* is second only to *Escherischia coli* as a cause of UTI in reproductive aged women (5, 6).

S. saprophyticus can be found in diverse niches including the environment, foods, livestock, and as a pathogen and commensal of humans. Several features of the epidemiology of *S. saprophyticus* suggest that infections leading to UTIs are acquired from the environment, rather than as a result of person-to-person transmission (7). This implies that adoption of the pathogenic niche by *S. saprophyticus* has not entailed a trade-off in its ability to live freely in the environment. A recent PCR-based survey of virulence factors in clinical and animal associated isolates showed that *dsdA*, a gene encoding D-serine deaminase that is important for survival in urine (8), and *uafA* and *aas*, adhesins that mediate binding to uroepithelium (9, 10), were present in all isolates surveyed (11), suggesting an underlying pleiotropy, with these virulence factors playing important roles in the diverse environments occupied by *S. saprophyticus*.

The human urinary tract could represent an evolutionary dead end for *S. saprophyticus* with “virulence factors” such as DsdA, UafA and Aas serving an essential function in the primary environmental niche and enabling invasion of the urinary tract as an accidental by-product of this unknown primary function. In this case, we would expect urinary isolates to be interspersed throughout a phylogeny of isolates from the primary niche(s). However, our previous research (7) indicated that human urinary tract infections are associated with a specific lineage of *S. saprophyticus*. Invasion of the human urinary tract enables *S. saprophyticus* to grow to large numbers in urine, isolated from competing bacterial species, before being re-deposited in the environment. This is analogous to *Vibrio cholerae*, which cycles through human and environmental niches and grows to high abundance in the human gut before being deposited in the

environment via stool (12, 13). Based on our previous observations and the example provided by other human pathogens that cycle through the environment, we hypothesized that the human urinary tract is an ecologically important niche for *S. saprophyticus* and sought to identify genetic signatures of adaptation to this niche.

The increased availability of sequencing data has enabled comparative genomic approaches that have led to identification of changes in gene content in association with pathogen emergence and shifts in host association. Several notable human pathogens, including *Mycobacterium tuberculosis*, *Yersinia pestis*, and *Francisella tulerensis*, are the product of a single emergence characterized by gene loss and horizontal acquisition of virulence factors (14–16). Similarly, genomic analysis of *Enterococcus faecium* revealed gene gains and losses affecting metabolism and antibiotic resistance in the emergence of a hypermutable hospital-adapted clade that coincided with the profound shift in hospital ecology caused by development of antibiotics (17). Gene gains via recombination have also allowed *Staphylococcus aureus* ST71 to emerge into a bovine-associated niche (18).

Using contemporary and ancient genomic data from strains of *S. saprophyticus*, we found previously that UTI-associated lineages of *S. saprophyticus* were not attendant with specific gene gains or losses; the evolutionary genetic processes underlying *S. saprophyticus'* adoption of the human-pathogenic niche are likely more subtle than what has been described for canonical pathogens (7). Here we have identified one of the mechanisms underlying *S. saprophyticus*' adaptation to the uropathogenic niche: a selective sweep in the Aas adhesin, which is associated with an apparently large-scale expansion into the human-pathogenic niche. This is, to our knowledge, the first identification of a single nucleotide sweep in a bacterium.

Results.

We reconstructed the phylogeny of *S. saprophyticus* isolates (Table S1) from a whole genome alignment using maximum likelihood inference implemented in RAxML (Figure 1). The bacterial isolates are separated into two clades, which we have previously named Clades P and E (7). In both clades, human associated lineages are nested

among isolates from diverse sources, including food (cheese rind, ice cream, meat), indoor and outdoor environments, and animals. Interestingly, cheese rinds harbor diverse strains of *S. saprophyticus*, which cluster with both human- and animal-pathogenic strains.

Thirty-three of 37 modern, human-pathogenic isolates are found within a single lineage (that we term lineage U, for UTI-associated) to which bovine-pathogenic (mastitis), food-associated isolates, and an ancient genome are basal. Given the association between this lineage and illness in humans, we were curious about its potential adaptation to the human pathogenic niche. The placement of the 800-year old strain between bovine and human-associated lineages suggests it could represent a generalist intermediate between human-adapted and bovid-adapted strains.

Core genome analysis of the 58 isolates of *S. saprophyticus* in our sample showed substantial variability in gene content; the core genome is composed of 1798 genes, and there are an additional 7110 genes in the pan genome. We found previously that uropathogenic isolates of *S. saprophyticus* were not associated with any unique gene content (7). Given the variability in accessory gene content among this larger sample of isolates, we decided to test for relative differences in accessory gene content between human clinical isolates and other isolates using Scoary (19), which performs a genome wide association study (GWAS) using gene presence and absence. We did not identify any genes that were significantly associated with the human pathogenic niche after correction for multiple hypothesis testing using the Bonferroni method.

In addition to the variability observed in gene content, analyses of the core genome also indicated relatively frequent recombination among *S. saprophyticus* (Figure 3). We identified recombinant regions with Gubbins (20), which identifies regions with high densities of substitutions. These results indicated that 70% of sites in the *S. saprophyticus* alignment have been affected by recombination. Recombination can affect bacterial evolution both by introducing novel polymorphisms from outside the population and by reshuffling alleles without increasing overall diversity. Considering sites that are reshuffled within the *S. saprophyticus* sample as recombinant, we estimate a ratio of recombinant to non-recombinant SNPs of 3.4. When considering only

the SNPs that introduce novel diversity as recombinant, our estimate of the ratio of recombinant SNPs to non-recombinant SNPs is 0.51. The mean r/m of branches in the phylogeny is 0.82 as estimated by Gubbins (range: 0-7.6). Removal of recombinant SNPs did not affect the topology of the maximum likelihood phylogeny. We observed regional patterns in the amount of recombination inferred, and, as expected, recombination appears frequent at mobile elements such as the staphylococcal cassette chromosomes ($SCC_{15305RM}$ and $SCC_{15305cap}$) and *v*Ss15305 (10).

Adaptation to a new environment may be facilitated by advantageous mutations that quickly rise in frequency, leaving a characteristic genomic imprint: reduced diversity at the target locus and nearby linked loci (i.e. selective sweep, (21, 22)). In order for positive selection to be evident as a local reduction in diversity, there must be sufficient recombination that the target locus is unlinked from the rest of the genome; for this reason, scans for sweeps have been used primarily for sexually reproducing organisms (23–26). As described above and in prior work (7), we found evidence of frequent recombination among *S. saprophyticus*. We hypothesized that *S. saprophyticus*' transition to the uropathogenic niche may have been driven by selection for one or more mutations that were advantageous in the new environment, and that levels of recombination have been sufficient to preserve the signature of a selective sweep at loci under positive selection. We therefore used a sliding window analysis of diversity along the *S. saprophyticus* alignment as an initial screen for positive selection. We identified a marked regional decrease in nucleotide diversity (π) and Tajima's D (TD) that is specific to lineage U (Figure 2); TD for this window was -0.38 and 0.94 for non-lineage U Clade P isolates and Clade E isolates, respectively. The region with decreased π /TD corresponds to 1,760,000-1,820,000 bp in *S. saprophyticus* ATCC 15305 and has the lowest values of π and TD in the entire alignment. We investigated the sensitivity of our sliding window analyses to sampling by randomly subsampling lineage U isolates to the same size as Clade E (n = 10); we found the results to be robust to changes in sampling scheme and size.

To complement the sliding window analysis and pinpoint candidate variants under positive selection, we used an approach based on allele frequency differences between

bacterial isolates from different niches. We calculated Weir and Cockerham's F_{ST} (27) for single nucleotide polymorphisms (SNPs) in the *S. saprophyticus* genome using human association and non-human association to define populations. The region of low π /TD included three non-synonymous variants in the top 0.05% of F_{ST} values (Table 1). One of these variants was fixed among human-associated isolates in lineage U (position 1811777 in ATCC 15305, F_{ST} = 0.48) and distinct from the ancestral allele found in basal lineages of Clade P, including the ancient strain of *S. saprophyticus* Troy. This suggests that the variant may have been important in adaptation to the human urinary tract. To assess the significance of the F_{ST} value for this variant, we performed permutations by randomly assigning isolates as human associated, and we did not achieve F_{ST} values higher than 0.28 in 100 permutations.

Selective sweeps may be evident as a longer than expected haplotype block, since neutral variants linked to the adaptive mutation will also sweep to high frequencies (28). Given the evidence suggesting there was a selective sweep at this locus, we used haplotype based statistics to test for such a signature in the *S. saprophyticus* alignment. Haplotype-based methods are hypothesized to not be applicable to bacteria due to differences between crossing over and bacterial patterns of recombination (29), but the methods had not been tested in a scenario akin to a classical sweep, in which local changes in diversity and the SFS have been observed. We found that the variant at position 181177 did show a signature of a sweep using the extended haplotype homozygosity (EHH) statistic (28) (Figure 4). However, the variant did not have an extreme value of nS_L, which compares haplotype homozygosity for ancestral and derived alleles (30), after normalization by the allele frequency.

The variant of interest (*aas*_2206A>C) causes a threonine to proline change in the amino acid sequence of Aas, a bifunctional autolysin with a fibronectin binding domain (Figure 5, (31)). There are 8 additional nonsynonymous polymorphisms in the fibronectin binding domain; however, none are as highly associated with human pathogenic isolates (Table 1). Adhesins such as Aas are important in the pathogenesis of *S. saprophyticus* urinary tract infections, and this gene has been previously implicated as a virulence factor (9, 31, 32).

The Aas variant is in a region known to bind fibronectin (Figure 5, (31)) and may be under selection because it affects adhesion to this host protein. We used ELISAs to investigate potential effects of *aas*_2206A>C on binding to fibronectin and thrombospondin-1, which binds to this region of the homologous AtlE amidase from *Staphylococcus epidermidis* (33). Staphylococcal autolysins contains 3 C-terminal repeats (R1-R3), which can each be divided into two subunits (a and b) based on structural information (34). We confirmed that Aas R1a1b binds fibronectin and discovered that it also binds thrombospondin (Figure 6); there was no detectable difference between the ancestral and derived R1a1b alleles in binding to fibronectin or thrombospondin (human or bovine).

Interestingly, we observed several instances of recombination of the *aas* variant. In each case, the recombination event reinforced the association of the derived allele with human infection. Two of the non-human-associated bacterial isolates in lineage U – an isolate from a pig and a second from cheese rind – had evidence of a recombination event at the *aas* locus resulting in acquisition of the ancestral allele. Conversely, one of the human UTI isolates in Clade E (for which the ancestral allele is otherwise fixed) acquired the derived *aas* variant.

Several human pathogens appear to have undergone recent population expansion (35–38). We wondered whether the uropathogenic lineage of *S. saprophyticus* might also have undergone a recent change in its effective size. The genome wide estimate of TD for lineage U was negative (-0.58), which is consistent with population expansion. We used the methods implemented in ∂a∂i (39) to identify the demographic model that best fit the observed synonymous SFS of lineage U (Figure 7).

The synonymous SFS showed an unexpected excess of high frequency derived alleles, which we hypothesized were the result of gene flow from populations with ancestral variants. Within population recombination has been shown to have no effect on SFS-based methods of demographic inference in bacteria (40). However, external sources of recombination were not modeled in previous studies. We used SimBac (41) to simulate bacterial populations with a range of internal and external recombination rates. Similar to previous studies, we did not find that within population recombination had an effect

the value of Tajima's D. However, we found that recombinant tracts from external sources resulted in positive values of Tajima's D (Figure 8). Positive values of Tajima's D are also associated with population bottlenecks and balancing selection.

We used fastGEAR (42) to identify recombinant tracts that originated outside of lineage U, and these sites were removed from the analysis prior to demographic inference. We compared five demographic models (constant size, instantaneous population size change, exponential population size change, instantaneous population size change followed by exponential, and two instantaneous population size changes , Figure 9) and used bootstrapping to estimate the uncertainty of the parameters and to adjust the composite likelihoods using the Godambe Information Matrix implemented in ∂a∂i (43). We found significant evidence for an expansion in all models (Table 2). The best fitting model was an instantaneous contraction followed by an instantaneous expansion, in which the population underwent a tight bottleneck followed by a 15-fold expansion without recovering to its ancestral size ($\nu = N_e/N_{anc}$, τ = generations/N_{anc}, ν_A: 2.9 x 10^{-2}, ν_B: 4.5 x 10^{-1}, τ_A: 1.2 x 10^{-1}, τ_B: 3.1 x 10^{-3}).

Recombination and positive selection are known to confound the inference of bacterial demography (40), so we used simulations to investigate their effects on our demographic inference for uropathogenic *S. saprophyticus.* We used SFS_CODE (44) to simulate positive selection (with a range of recombination rates) and evaluate its effects on the accuracy of demographic inference with ∂a∂i. The method implemented in ∂a∂i relies on inference from the synonymous SFS, but it's possible for synonymous variation to be affected by selection, particularly at low rates of recombination (40, 45). Neutral simulations with gene conversion did not affect demographic inference. We did find that positive selection can affect the synonymous SFS, resulting in inference of population size changes. In simulations of positive selection in a population of constant size, we found the spurious inference to be a bottleneck rather than an expansion. This suggests that the observed synonymous SFS of lineage U has been affected both by positive selection and by demographic expansion.

Discussion.

A central question in the population biology of infectious diseases is how and why pathogenic traits emerge in microbes. Addressing this question is important for understanding novel disease emergence and for identifying the genetic basis of virulence. Here we present evidence suggesting that a mutation in *S. saprophyticus*' *aas*, which binds host matrix proteins, is under positive selection and has enabled emergence and spread of a human pathogenic, UTI-associated lineage of this bacterium.

S. saprophyticus is familiar to medical microbiologists and clinicians as a common cause of UTIs (46), which are associated with significant morbidity, economic costs, and severe complications (4). Despite its strong association with UTIs in humans, *S. saprophyticus* can also be isolated from diverse environments including livestock, food and food processing plants, and the environment (47, 48). Our previous research suggested that pathogenicity to humans is a derived trait in the species (7).

This pattern is replicated here, where phylogenetic analyses link human UTI with two lineages of *S. saprophyticus* that are nested among isolates from diverse, non-human niches (i.e. free living, food- and animal-associated) . The *aas* mutation arose in lineage U, which contains most of the UTI isolates. Two lineages are basal to lineage U: one is bovine-associated, and the other contains an ancient bacterial sequence from a pregnancy-related infection in Late Byzantine Troy. The Troy bacterium has the ancestral, bovine-associated *aas* allele, and we have previously hypothesized (7) that this lineage could be associated with human infections in regions where humans have close contacts – e.g. shared living quarters– with livestock, as they did at Troy during this time.

A second cluster of UTI isolates appears in Clade E. One isolate has acquired the derived *aas* allele, which parallels our finding that two non-human isolates in lineage U acquired the ancestral variant; all recombination events that we observed at this locus reinforced the association between *aas*_2206A>C and human infection.

Several UTI isolates in Clade E do not have the derived *aas* allele and the clustering of UTI isolates suggests there may be a distinct adaptive path to virulence in this clade. Larger and more comprehensive samples will be needed to investigate this hypothesis and to identify the factors shaping the separation of clades P & E.

The *aas* mutation has characteristics associated with a classical selective sweep driven by positive selection, namely a regional reduction in diversity (21) and Tajima's D (22, 49). With the exception of the interesting allelic replacements noted above, there was also relatively little recombination at this locus, consistent with it being functionally important. To our knowledge, this is the first description of a single nucleotide sweep in a bacterium.

Depending on the strength of selection and recombination rate, positive selection in bacteria has been observed to affect the entire genome, resulting in clonal replacements, or to only affect specific regions of the genome (50). For example, multiple clonal replacements have occurred in *Shigella sonnei* populations in Vietnam due to acquisition of resistance to antimicrobials and environmental stress (51). Recurrent clonal replacements have also been observed within single hosts during chronic infection of cystic fibrosis patients by *Pseudomonas aeruginosa* (52). Environmental bacterial populations can also be subject to clonal replacements: a metagenomic time course study of Trout Bog found evidence of clonal replacement occurring in natural bacterial populations but not gene or region specific sweeps (53). However, large regions of low diversity were also observed, suggesting gene-specific selective sweeps had occurred prior to the start of the study. Shapiro et al. identified genomic loci that differentiated *Vibrio cyclitrophicus* associated with distinct niches but that had limited diversity within niches; they concluded that differentiation of these populations had been enabled by recombination events that reinforced the association of alleles with the niche in which they were advantageous (54).

The *aas*_2206A>C mutation is within a group of genetic variants that differentiates bacteria associated with human-pathogenic versus other niches (i.e. F_{ST} outlier). SNPs associated with specific clinical phenotypes were described recently in the pathogen *Streptococcus pyogenes* (55), which is consistent with our finding that clinical

317 phenotypes can represent distinct niche spaces preferentially occupied by sub-
318 populations of bacteria. There is also precedent for a single nucleotide polymorphism to
319 affect host tropism of bacteria (56).

320 In sexually reproducing organisms, haplotype based statistics are frequently used to
321 identify selective sweeps because positively selected alleles will also increase the
322 frequency of nearby linked loci faster than recombination can disrupt linkage, producing
323 longer haplotypes for selected alleles (28, 57). We found that *aas*_2206A>C had a
324 longer haplotype than the ancestral variant, but this difference was not extreme relative
325 to other regions of the genome (assessed with the nS_L statistic). Haplotype based
326 statistics have been found to perform poorly in purebred dogs, where linkage across the
327 genome is high (58). Relatively low levels of recombination may also contribute to a lack
328 of sensitivity when haplotype-based detection methods are applied to bacteria; linkage
329 of sites is also likely to be disrupted in a less predictable way by bacterial gene
330 conversion than by crossing over (29). Based on our findings, we conclude that
331 screening for regional decreases in diversity and distortions of the SFS (i.e. sliding
332 window analyses) and identification of genetic variants with extreme differences in
333 frequency between niches can be useful in identifying candidate sites of positive
334 selection in bacteria.

335 *S. saprophyticus* encodes a number of adhesins including UafA, UafB, SdrI, and Aas.
336 UafA and Aas are found in all isolates, suggesting that they play important roles in the
337 diverse niches occupied by *S. saprophyticus*. Aas has autolytic, fibronectin binding, and
338 haemagluttinating functions (9, 31, 32, 59). We identified a single, non-synonymous
339 polymorphism as a target of selection in the fibronectin binding repeats of *Aas*. This
340 variant is predicted to affect the repeat's structure, as proline has a more rigid structure
341 than other amino acids. Adhesins are plausible candidates for adaptation to the
342 uropathogenic niche, as they are known to be important virulence factors in pathogens
343 causing urinary tract infections (60). Fibronectin binding proteins including Aas have
344 been identified as virulence factors in *S. saprophyticus* and *Enterococcus faecalis* (32,
345 61, 62). Adhesion to the uroepithelium is essential for uropathogens to establish
346 themselves in the bladder, where they are subject to strong shear stress (63): we

hypothesize that *S. saprophyticus* with the derived *aas* variant are better able to colonize the human bladder.

Invasion of the human urinary tract may provide a fitness advantage by allowing relative enrichment of *S. saprophyticus* in a site with little competition from other bacterial species and by providing a mechanism of dispersal in the environment. In analyses of selection in *Escherichia coli*, another bacterium occupying diverse niches, residues in the adhesin FimH were found to be subject to positive selection in uropathogenic strains (64–66). FimH binds mannose, providing protection from shear stress through a catch bond mechanism (67). Interestingly, *Borrelia burgdorferi*'s vascular adherence and resistance to shear stress were recently found to be enabled by interactions between a bacterial adhesin and host fibronectin that also use a catch bond mechanism (68). There are also precedents in *Staphylococcus aureus* for polymorphisms in bacterial fibronectin-binding adhesins to affect the strength of binding, and for these polymorphisms to associate with specific clinical phenotypes (69).

Further experiments are needed to investigate the effects of variation in Aas on *S. saprophyticus* biology. In our preliminary investigations of binding using ELISA assays of recombinant bacterial peptides, we did not detect differences between ancestral and derived alleles in binding of the R1a1b repeat to fibronectin. The variant could still affect fibronectin binding by altering conformation of the protein in a manner analogous to FimH in *E. coli* (66). It's also possible that variants in the peptide affect binding under specific conditions that we did not test. Another possibility is that the variant affects autolysis or other as yet undescribed functions of Aas. The roles of adhesins and other virulence factors in *S. saprophyticus*'s colonization of niches in livestock and the environment are also interesting topics for further study.

Our demographic analysis of the uropathogenic lineage of *S. saprophyticus* showed evidence of a population bottleneck and subsequent expansion. Bottlenecks and expansion of drug resistant clones have previously been shown to affect the population structure of *Streptococcus agalactiae* (70), demonstrating the effects of positive selection on the demographic trajectories of bacterial sub-populations. However, previous work has also shown that selection - and recombination - can produce

spurious results from demographic inference in bacteria (40, 71). We used an SFS-based method to reconstruct the demographic history of *S. saprophyticus*: the accuracy of demographic inference using these methods has been shown to be unaffected by within-population recombination (40) and this was confirmed in our analyses of simulated data. We found that recombination from external sources may result in an excess of intermediate frequency variants, which is also a signature of population bottlenecks, so we masked externally imported sites. However, the frequency of synonymous variants could still be affected by selection on linked non-synonymous sites, including the selective sweep in *aas* that we have described. We performed simulations to address these potential confounders and aid in the interpretation of our demographic inferences. Simulation of a single site under positive selection resulted in the inference of a bottleneck (N_e/N_a = 0.01-0.42), indicating that, at the recombination rates we simulated, diversity was lost from neutrally evolving sites due to their linkage to the site under selection. In inferences from our observed data, a bottleneck was followed by a 15-fold expansion, suggesting that lineage U has undergone both a selective sweep and demographic expansion.

Here we have described adaptation of *S. saprophyticus* that may have enabled its expansion into a human pathogenic niche. Mutation of a single nucleotide within the *aas* adhesin appears to have driven a selective sweep, and allele frequency differences at the locus are consistent with niche-specific adaptation. Lateral gene transfer events in *aas* reinforced the association of the positively selected allele with human infection. These results provide new insights into the emergence of virulence in bacteria and outline an approach for discovering the molecular basis of adaptation to the human pathogenic niche.

Methods.

DNA extraction. After overnight growth in TSB at 37°C in a shaking incubator, cultures were pelleted and resuspended in 140 µL TE buffer. Cells were incubated overnight with 50 units of mutanolysin. We used the MasterPure Gram Positive DNA Purification Kit (EpiCentre) for DNA extraction.For DNA precipitation we used 1 mL 70% ethanol

406 and centrigured at 4°C for 10 minutes. We additionally used a SpeedVac for 10 minutes
407 to ensure pellets were dry before re-suspending the pellet in 50 µL water.

408 *Library preparation and sequencing.* For SSC01, SSC02, and SSC03, library prep was
409 performed using a modified Nextera protocol as decribed by Baym et al. (72) with a
410 reconditioning PCR with fresh primers and polymerase for an additional 5 PCR cycles to
411 minimize chimeras and a two-step bead based size selection with target fragment size
412 of 650 bp and sequenced on an Illumina HiSeq 2500 (paired-end, 150 bp). For 43,
413 SSC04, SCC05, and SSMast, DNA was submitted to the University of Wisconsin-
414 Madison Biotechnology Center for library preparation and were prepared according the
415 TruSeq Nano DNA LT Library Prep Kit (Illumina Inc., San Diego, California, USA) with
416 minor modifications. A maximum of 200 ng of each sample was sheared using a
417 Covaris M220 Ultrasonicator (Covaris Inc, Woburn, MA, USA). Sheared samples were
418 size selected for an average insert size of 550 bp using Spri bead based size exclusion.
419 Quality and quantity of the finished libraries were assessed using an Agilent DNA High
420 Sensitivity chip (Agilent Technologies, Santa Clara, CA) and Qubit dsDNA HS Assay
421 Kit, respectively. Libraries were standardized to 2 µM. Paired-end, 150 bp sequencing
422 was performed using v2 SBS chemistry on an Illumina MiSeq sequencer. Images were
423 analyzed using the Illumina Pipeline, version 1.8.2.

424 *Reference guided mapping.* We mapped reads to ATCC 15305 using via a pipeline
425 (available at https://github.com/pepperell-lab/RGAPepPipe). Briefly, read quality was
426 assessed and trimmed with TrimGalore! v 0.4.0
427 (www.bioinformatics.babraham.ac.uk/projects/trim_galore), which runs both FastQC
428 (www.bioinformatics.babraham.ac.uk/projects/fastqc) and cutadapt. Reads were
429 mapped to using BWA-MEM v 0.7.12 (73) and sorted using Samtools v 1.2 (74). We
430 used Picard v 1.138 (picard.sourceforge.net) to add read group information and
431 removed duplicates. Reads were locally realigned using GATK v 2.8.1 (75). We
432 identified variants using Pilon v 1.16 (76) (minimum read depth: 10, minimum mapping
433 quality: 40, minimum base quality: 20).

434 *Assembly.* We used the iMetAMOS pipeline for *de novo* assembly (77). We chose to
435 compare assemblies from SPAdes (78), MaSurCA (79), and Velvet (80). KmerGenie

(81) was used to select kmer sizes for assembly. iMetAMOS uses FastQC, QUAST (82), REAPR (83), LAP (84), ALE (85), FreeBayes (86), and CGAL (87) to evaluate the quality of reads and assemblies. We also used Kraken (88) to detect potential contamination in sequence data. For all newly assembled isolates (43, SSC01-05, SSMast), the SPAdes assembly was the highest quality. Assembly statistics are reported in Table S2.

Annotation and gene content analyses. We annotated the de novo assemblies using Prokka v 1.11 (89) and used Roary (90) to identify orthologous genes in the core and accessory genomes. To look for associations between accessory gene content and human association, we used Scoary (19). For the analysis, we used human association as our trait, and we adjusted the p-value for multiple comparisons using the Bonferroni method.

Alignment. When short read data for reference guided mapping were unavailable, whole genome alignment of genomes to ATCC 15305 was performed using Mugsy v 2.3 (91). Repetitive regions in the reference genome greater than 100 bp were identified using nucmer, and these regions were masked in the alignment used in downstream analyses.

Maximum likelihood phylogenetic analysis. Maximum likelihood phylogenetic trees were inferred using RAxML 8.0.6 (92). We used the GTRGAMMA substitution model and performed bootstrapping using the autoMR convergence criteria. Tree visualizations were created in ggtree (93).

Population genetics statistics. To calculate π and Tajima's D, we used EggLib v 2.1.10 (94), a Python package for population genetic analyses. A script to perform the sliding window analysis is available at https://github.com/tatumdmortimer/popgen-stats/slidingWindowStats.py. We used vcflib (https://github.com/vcflib/vcflib) to calculate F_{ST} and EHH and selscan v 1.1.0b (95) to calculate nS_L.

Recombination analyses. To identify recombinant regions in the *S. saprophyticus* alignment, we used Gubbins v 2.1.0 (20). fastGEAR (42) was used with the recommended input specifications to identify recombination events between major

lineages of *S. saprophyticus*. We used Circos (96) for visualization of recombinant tracts.

Site frequency spectrum. We used SNP-sites v 2.0.3 (97) to convert the alignment of *S. saprophyticus* isolates to a multi-sample VCF. SnpEff v 4.1j (98) was used to annotate variants in this VCF as synonymous, non-synonymous, or intergenic. Using the Troy genome as an outgroup, we calculated an unfolded site frequency spectrum (SFS) for lineage U for each category of sites. To reduce the impact of lateral gene transfer on the SFS, we removed sites where the origin was outside lineage U based on results of fastGEAR.

*Ancestral reconstruction of aas*_2206A>C. We used TreeTime (99) to reconstruct the evolutionary history of the variant in *aas* using the maximum likelihood phylogeny inferred using RAxML.

Demography. We performed demographic inference with the synonymous SFS using ∂a∂i (39). Models tested were the standard neutral model, expansion, and exponential growth. Parameters, ν (N_e/ N_a) and τ (time scaled by 2), were optimized for both the expansion and exponential growth models. Significance of the expansion and exponential growth model compared to the standard neutral model was evaluated using a likelihood ratio test. The scripts used to perform this analysis are available at https://github.com/tatumdmortimer/popgen-stats.

Simulations. Simulations were performed in SimBac (41) to evaluate the effect of external recombination on the SFS. Populations were simulated with sample size and θ equivalent to our sample of lineage U (n = 44, θ = 0.003). The length of internal recombinant tracts was 6500 bp (median of Gubbins output), and the length of external recombination events was 3000 bp (median of fastGEAR output). Internal recombination was simulated at rates ranging from 0 to 0.03 (r/m = 10). External recombination was simulated at rates ranging from 0 to 0.003. The lower bound of difference for external recombination was 0, and the upper bound was simulated at ranges from 0.25-1.0. Simulations were performed in SFS_CODE (release date 9/10/2015) (44) to evaluate the power of ∂a∂i to accurately estimate demographic

parameters in the presence of gene conversion and selection. We simulated a locus of length 100 kb with theta 0.003, gene conversion tract length of 1345 bp, and a range of recombination/mutation ratios (0.0002-2.0). In addition to neutral simulations with gene conversion, we also performed simulations with a single site under selection (γ = 10-1000) with the same parameters as neutral simulations.

Expression of R1ab. Human-associated and ancestral strain R1ab were cloned into the expression vector pET-ELMER (LM Maurer 2010, JBC), transformed into BL21(DE3)cells (EMD, Gibbstown, NJ) for expression and induced with 1mM IPTG. Bacteria were lysed in 100 mM NaH_2PO_4, 10 mM Tris, 8M urea, 1 mM β-mercapto-ethanol, 5 mM imidazole, pH 8.0 (lysis buffer). The cleared lysate was incubated overnight with nickel-nitrilotriacetic acid agarose (Qiagen), washed and eluted in lysis buffer pH 7.0 plus 300 mM imidazole.

ELISA. Antigen was diluted to 10 µg/ml in TBS (10 mM Tris, 150 mM, pH 7.4) and used to coat 96-well microtiter plates (Costar 3590 high binding, Corning Inc., Corning, NY) with 50 µl per well, for 16 h at 4□C. The plates were blocked with 1% BSA in TBS plus 0.05% Tween-20 (TBST) for 1 h. After washing three times with TBST, purified plasma fibronectin (100) or platelet-derived thrombospondin-1 (101) diluted to 10, 3, 1, or 0.3µg/ml in TBST plus 0.1% BSA, were added to the plates and incubated for 2 h. Plates were washed four times with TBST. Rabbit anti-fibronectin and rabbit anti-thrombospondin antibodies diluted in TBST plus 0.1% BSA were added to the appropriate wells and incubated for 1 h. Plates were washed four times with TBST. Peroxidase-conjugated secondary antibody was incubated with the plates for 1 h. Plates were washed four times with TBST and 50 µl per well of SureBlue TMB peroxidase substrate (KLP) was added to each well. Color development was monitored for 10-30 min, 50 µl of TMB stop solution (KLP) was added, followed by measurement of absorbance at 450nm.

Funding Information.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256259 to TDM and

MBO. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. TDM is also supported by National Institutes of Health National Research Service Award (T32 GM07215). CSP is supported by National Institutes of Health (R01AI113287).

Acknowledgments.

We thank Andrew Kitchen (University of Iowa), Jeniel Nett (UW-Madison), JD Sauer (UW-Madison), and Rod Welch (UW-Madison) for their helpful input on this study. We also thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing library preparation and sequencing facilities and services.

References.

1. Stamm WE, Norrby SR. 2001. Urinary Tract Infections: Disease Panorama and Challenges. J Infect Dis 183:S1–S4.

2. Ronald AR, Nicolle LE, Stamm E, Krieger J, Warren J, Schaeffer A, Naber KG, Hooton TM, Johnson J, Chambers S, Andriole V. 2001. Urinary tract infection in adults: research priorities and strategies. Int J Antimicrob Agents 17:343–348.

3. Barber AE, Norton JP, Spivak AM, Mulvey MA. 2013. Urinary Tract Infections: Current and Emerging Management Strategies. Clin Infect Dis 57:719–724.

4. Foxman B. 2014. Urinary tract infection syndromes: occurrence, recurrence, bacteriology, risk factors, and disease burden. Infect Dis Clin North Am 28:1–13.

5. Wallmark G, Arremark I, Telander B. 1978. Staphylococcus saprophyticus: A Frequent Cause of Acute Urinary Tract Infection among Female Outpatients. J Infect Dis 138:791–797.

6. Kahlmeter G. 2003. An international survey of the antimicrobial susceptibility of pathogens from uncomplicated urinary tract infections: the ECO·SENS Project. J Antimicrob Chemother 51:69–76.

7. Devault AM, Mortimer TD, Kitchen A, Kiesewetter H, Enk JM, Golding GB, Southon J, Kuch M, Duggan AT, Aylward W, Gardner SN, Allen JE, King AM, Wright G, Kuroda M, Kato K, Briggs DE, Fornaciari G, Holmes EC, Poinar HN, Pepperell CS. 2017. A molecular portrait of maternal sepsis from Byzantine Troy. eLife 6:e20983.

8. Gatermann S, John J, Marre R. 1989. Staphylococcus saprophyticus urease: characterization and contribution to uropathogenicity in unobstructed urinary tract infection of rats. Infect Immun 57:110–116.

9. Meyer HG, Wengler-Becker U, Gatermann SG. 1996. The hemagglutinin of Staphylococcus saprophyticus is a major adhesin for uroepithelial cells. Infect Immun 64:3893–3896.

10. Kuroda M, Yamashita A, Hirakawa H, Kumano M, Morikawa K, Higashide M, Maruyama A, Inose Y, Matoba K, Toh H, Kuhara S, Hattori M, Ohta T. 2005. Whole genome sequence of Staphylococcus saprophyticus reveals the pathogenesis of uncomplicated urinary tract infection. Proc Natl Acad Sci U S A 102:13272–13277.

11. Kleine B, Gatermann S, Sakinc T. 2010. Genotypic and phenotypic variation among Staphylococcus saprophyticus from human and animal isolates. BMC Res Notes 3:163.

12. Nelson EJ, Harris JB, Glenn Morris J, Calderwood SB, Camilli A. 2009. Cholera transmission: the host, pathogen and bacteriophage dynamic. Nat Rev Microbiol 7:693–702.

13. Boucher Y, Orata FD, Alam M. 2015. The out-of-the-delta hypothesis: dense human populations in low-lying river deltas served as agents for the evolution of a deadly pathogen. Front Microbiol 6.

14. Veyrier FJ, Dufort A, Behr MA. 2011. The rise and fall of the Mycobacterium tuberculosis genome. Trends Microbiol 19:156–161.

15. McNally A, Thomson NR, Reuter S, Wren BW. 2016. “Add, stir and reduce”: Yersinia spp. as model bacteria for pathogen evolution. Nat Rev Microbiol 14:177–190.

16. Larsson P, Elfsmark D, Svensson K, Wikström P, Forsman M, Brettin T, Keim P, Johansson A. 2009. Molecular Evolutionary Consequences of Niche Restriction in Francisella tularensis, a Facultative Intracellular Pathogen. PLoS Pathog 5:e1000472.

17. Lebreton F, Schaik W van, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L, Saif S, Young S, Zeng Q, Wortman J, Birren B, Willems RJL, Earl AM, Gilmore MS. 2013. Emergence of Epidemic Multidrug-Resistant Enterococcus faecium from Animal and Commensal Strains. mBio 4:e00534-13.

586 18. Fitzgerald JR, Nutbeam-Tuffs S, Richardson E, Lee CY, Gupta RK, Richards AC,
587 Black NS, Corander J, McAdam PR, Spoor LE, O'Gara JP, Mendonca C, Wilson
588 GJ. 2015. Recombination-mediated remodelling of host-pathogen interactions
589 during Staphylococcus aureus niche adaptation. Microb Genomics.

590 19. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in
591 microbial pan-genome-wide association studies with Scoary. Genome Biol 17:238.

592 20. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J,
593 Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant
594 bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15–e15.

595 21. Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res.

596 22. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The
597 hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics
598 140:783–796.

599 23. Harr B, Kauer M, Schlötterer C. 2002. Hitchhiking mapping: A population-based
600 fine-mapping strategy for adaptive mutations in Drosophila melanogaster. Proc Natl
601 Acad Sci 99:12949–12954.

602 24. Nair S, Williams JT, Brockman A, Paiphun L, Mayxay M, Newton PN, Guthmann J-
603 P, Smithuis FM, Hien TT, White NJ, Nosten F, Anderson TJC. 2003. A Selective
604 Sweep Driven by Pyrimethamine Treatment in Southeast Asian Malaria Parasites.
605 Mol Biol Evol 20:1526–1536.

606 25. Wright SI, Gaut BS. 2005. Molecular Population Genetics and the Search for
607 Adaptive Evolution in Plants. Mol Biol Evol 22:506–519.

608 26. de Simoni Gouveia JJ, da Silva MVGB, Paiva SR, de Oliveira SMP. 2014.
609 Identification of selection signatures in livestock species. Genet Mol Biol 37:330–
610 342.

611 27. Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of
612 Population Structure. Evolution 38:1358–1370.

613 28. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel
614 SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler
615 D, Cooper R, Kwiatkowski D, Ward R, Lander ES. 2002. Detecting recent positive
616 selection in the human genome from haplotype structure. Nature 419:832–837.

617 29. Shapiro BJ. 2014. Signatures of Natural Selection and Ecological Differentiation in
618 Microbial Genomes, p. 339–359. *In* Landry, CR, Aubin-Horth, N (eds.), Ecological
619 Genomics. Springer Netherlands.

620 30. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On Detecting
621 Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Mol Biol
622 Evol 31:1275–1291.

623 31. Hell W, Meyer H-GW, Gatermann SG. 1998. Cloning of aas, a gene encoding a
624 Staphylococcus saprophyticus surface protein with adhesive and autolytic
625 properties. Mol Microbiol 29:871–881.

626 32. Gatermann S, Meyer HG. 1994. Staphylococcus saprophyticus hemagglutinin
627 binds fibronectin. Infect Immun 62:4556–4563.

628 33. Kohler TP, Gisch N, Binsker U, Schlag M, Darm K, Völker U, Zähringer U,
629 Hammerschmidt S. 2014. Repeating Structures of the Major Staphylococcal
630 Autolysin Are Essential for the Interaction with Human Thrombospondin 1 and
631 Vitronectin. J Biol Chem 289:4070–4082.

632 34. Zoll S, Schlag M, Shkumatov AV, Rautenberg M, Svergun DI, Götz F, Stehle T.
633 2012. Ligand-Binding Properties and Conformational Dynamics of Autolysin
634 Repeat Domains in Staphylococcal Cell Wall Recognition. J Bacteriol 194:3789–
635 3802.

636 35. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-
637 Smith HMB, Quail MA, Rance R, Brooks K, Churcher C, Harris D, Bentley SD,

638 Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, Tonder A van,
639 Walker D, Wren BW, Dougan G, Parkhill J. 2010. Evolutionary dynamics of
640 Clostridium difficile over short and long time scales. Proc Natl Acad Sci 107:7527–
641 7532.

642 36. Nübel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, Žemličková H,
643 Leblois R, Wirth T, Jombart T, Balloux F, Witte W. 2010. A Timescale for Evolution,
644 Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-
645 Resistant Staphylococcus aureus. PLOS Pathog 6:e1000855.

646 37. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren
647 B, Galagan J, Feldman MW. 2013. The Role of Selection in Shaping Diversity of
648 Natural M. tuberculosis Populations. PLoS Pathog 9:e1003543.

649 38. Davies MR, Holden MT, Coupland P, Chen JHK, Venturini C, Barnett TC, Zakour
650 NLB, Tse H, Dougan G, Yuen K-Y, Walker MJ. 2015. Emergence of scarlet fever
651 Streptococcus pyogenes emm12 clones in Hong Kong is associated with toxin
652 acquisition and multidrug resistance. Nat Genet 47:84–87.

653 39. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring
654 the Joint Demographic History of Multiple Populations from Multidimensional SNP
655 Frequency Data. PLoS Genet 5:e1000695.

656 40. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The impact of
657 selection, gene conversion, and biased sampling on the assessment of microbial
658 demography. Mol Biol Evol msw048.

659 41. Brown T, Didelot X, Wilson DJ, De Maio N. 2016. SimBac: simulation of whole
660 bacterial genomes with homologous recombination. Microb Genomics 2.

661 42. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. 2016.
662 Analysis of recent and ancestral recombination reveals high-resolution population
663 structure in Streptococcus pneumoniae. bioRxiv 059642.

664 43. Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN. 2016. Computationally Efficient
665 Composite Likelihood Statistics for Demographic Inference. Mol Biol Evol 33:591–
666 593.

667 44. Hernandez RD. 2008. A flexible forward simulator for populations subject to
668 selection and demography. Bioinformatics 24:2786–2787.

669 45. Neher RA, Hallatschek O. 2013. Genealogies of rapidly adapting populations. Proc
670 Natl Acad Sci 110:437–442.

671 46. Widerstrom M, Wistrom J, Ferry S, Karlsson C, Monsen T. 2007. Molecular
672 Epidemiology of Staphylococcus saprophyticus Isolated from Women with
673 Uncomplicated Community-Acquired Urinary Tract Infection. J Clin Microbiol
674 45:1561–1564.

675 47. Hedman P, Ringertz O, Lindström M, Olsson K. 1993. The origin of
676 Staphylococcus saprophyticus from cattle and pigs. Scand J Infect Dis 25:57–60.

677 48. Kastman EK, Kamelamela N, Norville JW, Cosetta CM, Dutton RJ, Wolfe BE.
678 2016. Biotic Interactions Shape the Ecological Distributions of Staphylococcus
679 Species. mBio 7:e01157-16.

680 49. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by
681 DNA polymorphism. Genetics 123:585–595.

682 50. Shapiro BJ, David LA, Friedman J, Alm EJ. 2009. Looking for Darwin's footprints in
683 the microbial world. Trends Microbiol 17:196–204.

684 51. Holt KE, Nga TVT, Thanh DP, Vinh H, Kim DW, Tra MPV, Campbell JI, Hoang
685 NVM, Vinh NT, Minh PV, Thuy CT, Nga TTT, Thompson C, Dung TTN, Nhu NTK,
686 Vinh PV, Tuyet PTN, Phuc HL, Lien NTN, Phu BD, Ai NTT, Tien NM, Dong N,
687 Parry CM, Hien TT, Farrar JJ, Parkhill J, Dougan G, Thomson NR, Baker S. 2013.
688 Tracking the establishment of local endemic populations of an emergent enteric
689 pathogen. Proc Natl Acad Sci 110:17522–17527.

690 52. Caballero JD, Clark ST, Coburn B, Zhang Y, Wang PW, Donaldson SL, Tullis DE,
691 Yau YCW, Waters VJ, Hwang DM, Guttman DS. 2015. Selective Sweeps and
692 Parallel Pathoadaptation Drive Pseudomonas aeruginosa Evolution in the Cystic
693 Fibrosis Lung. mBio 6:e00981-15.

694 53. Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J,
695 Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG,
696 Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR.
697 2016. Genome-wide selective sweeps and gene-specific sweeps in natural
698 bacterial populations. ISME J 10:1589–1601.

699 54. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz
700 MF, Alm EJ. 2012. Population Genomics of Early Events in the Ecological
701 Differentiation of Bacteria. Science 336:48–51.

702 55. Bao Y-J, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ. 2016. Phenotypic
703 differentiation of Streptococcus pyogenes populations is induced by recombination-
704 driven gene-specific sweeps. Sci Rep 6:36644.

705 56. Viana D, Comos M, McAdam PR, Ward MJ, Selva L, Guinane CM, González-
706 Muñoz BM, Tristan A, Foster SJ, Fitzgerald JR, Penadés JR. 2015. A single natural
707 nucleotide mutation alters bacterial pathogen host tropism. Nat Genet advance
708 online publication.

709 57. Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting Natural Selection in Genomic
710 Data. Annu Rev Genet 47:97–120.

711 58. Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW.
712 2016. Evaluating the performance of selection scans to detect selective sweeps in
713 domestic dogs. Mol Ecol 25:342–356.

714 59. Meyer H-GW, Müthing J, Gatermann SG. 1997. The hemagglutinin of
715 Staphylococcus saprophyticus binds to a protein receptor on sheep erythrocytes.
716 Med Microbiol Immunol (Berl) 186:37–43.

717 60. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. 2015. Urinary tract
718 infections: epidemiology, mechanisms of infection and treatment options. Nat Rev
719 Microbiol advance online publication.

720 61. King NP, Beatson SA, Totsika M, Ulett GC, Alm RA, Manning PA, Schembri MA.
721 2011. UafB is a serine-rich repeat adhesin of Staphylococcus saprophyticus that
722 mediates binding to fibronectin, fibrinogen and human uroepithelial cells.
723 Microbiology 157:1161–1175.

724 62. Torelli R, Serror P, Bugli F, Sterbini FP, Florio AR, Stringaro A, Colone M, Carolis
725 ED, Martini C, Giard J-C, Sanguinetti M, Posteraro B. 2012. The PavA-like
726 Fibronectin-Binding Protein of Enterococcus faecalis, EfbA, Is Important for
727 Virulence in a Mouse Model of Ascending Urinary Tract Infection. J Infect Dis
728 206:952–960.

729 63. Wright KJ, Hultgren SJ. 2006. Sticky fibers and uropathogenesis: bacterial
730 adhesins in the urinary tract. Future Microbiol 1:75–87.

731 64. Chen SL, Hung C-S, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T,
732 Meyer RR, Ozersky P, Armstrong JR, Fulton RS, Latreille JP, Spieth J, Hooton TM,
733 Mardis ER, Hultgren SJ, Gordon JI. 2006. Identification of genes subject to positive
734 selection in uropathogenic strains of Escherichia coli: A comparative genomics
735 approach. Proc Natl Acad Sci 103:5977–5982.

736 65. Chen SL, Hung CS, Pinkner JS, Walker JN, Cusumano CK, Li Z, Bouckaert J,
737 Gordon JI, Hultgren SJ. 2009. Positive selection identifies an in vivo role for FimH
738 during urinary tract infection in addition to mannose binding. Proc Natl Acad Sci
739 106:22439–22444.

740 66. Schwartz DJ, Kalas V, Pinkner JS, Chen SL, Spaulding CN, Dodson KW, Hultgren
741 SJ. 2013. Positively selected FimH residues enhance virulence during urinary tract
742 infection by altering FimH conformation. Proc Natl Acad Sci 110:15530–15537.

743 67. Yakovenko O, Sharma S, Forero M, Tchesnokova V, Aprikian P, Kidd B, Mach A,
744 Vogel V, Sokurenko E, Thomas WE. 2008. FimH Forms Catch Bonds That Are
745 Enhanced by Mechanical Force Due to Allosteric Regulation. J Biol Chem
746 283:11596–11605.

747 68. Niddam AF, Ebady R, Bansal A, Koehler A, Hinz B, Moriarty TJ. 2017. Plasma
748 fibronectin stabilizes Borrelia burgdorferi–endothelial interactions under vascular
749 shear stress by a catch-bond mechanism. Proc Natl Acad Sci 114:E3490–E3498.

750 69. Messina JA, Thaden JT, Sharma-Kuinkel BK, Jr VGF. 2016. Impact of Bacterial
751 and Human Genetic Variation on Staphylococcus aureus Infections. PLOS Pathog
752 12:e1005330.

753 70. Da Cunha V, Davies MR, Douarre P-E, Rosinski-Chupin I, Margarit I, Spinali S,
754 Perkins T, Lechat P, Dmytruk N, Sauvage E, Ma L, Romi B, Tichit M, Lopez-
755 Sanchez M-J, Descorps-Declere S, Souche E, Buchrieser C, Trieu-Cuot P, Moszer
756 I, Clermont D, Maione D, Bouchier C, McMillan DJ, Parkhill J, Telford JL, Dougan
757 G, Walker MJ, Devani Consortium, Holden MTG, Poyart C, Glaser P. 2014.
758 Streptococcus agalactiae clones infecting humans were selected and fixed through
759 the extensive use of tetracycline. Nat Commun 5.

760 71. Hedge J, Wilson DJ. 2014. Bacterial Phylogenetic Reconstruction from Whole
761 Genomes Is Robust to Recombination but Demographic Inference Is Not. mBio
762 5:e02158-14.

763 72. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. 2015.
764 Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. PLOS
765 ONE 10:e0128036.

766 73. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with
767 BWA-MEM. 1303.3997. arXiv e-print.

768 74. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
769 Durbin R. 2009. The Sequence Alignment/Map format and SAMtools.
770 Bioinformatics 25:2078–2079.

771 75. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis
772 AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM,
773 Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework
774 for variation discovery and genotyping using next-generation DNA sequencing
775 data. Nat Genet 43:491–498.

776 76. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA,
777 Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for
778 Comprehensive Microbial Variant Detection and Genome Assembly Improvement.
779 PLOS ONE 9:e112963.

780 77. Koren S, Treangen TJ, Hill CM, Pop M, Phillippy AM. 2014. Automated ensemble
781 assembly and validation of microbial genomes. BMC Bioinformatics 15:126.

782 78. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
783 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,
784 Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm
785 and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

786 79. Zimin A, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The
787 MaSuRCA genome Assembler. Bioinformatics btt476.

788 80. Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly
789 using de Bruijn graphs. Genome Res 18:821–829.

790 81. Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for
791 genome assembly. Bioinformatics 30:31–37.

792 82. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment
793 tool for genome assemblies. Bioinforma Oxf Engl 29:1072–1075.

794 83. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. 2013. REAPR: a
795 universal tool for genome assembly evaluation. Genome Biol 14:R47.

796 84. Ghodsi M, Hill CM, Astrovskaya I, Lin H, Sommer DD, Koren S, Pop M. 2013. De
797 novo likelihood-based measures for comparing genome assemblies. BMC Res
798 Notes 6:334.

799 85. Clark SC, Egan R, Frazier PI, Wang Z. 2013. ALE: a generic assembly likelihood
800 evaluation framework for assessing the accuracy of genome and metagenome
801 assemblies. Bioinformatics 29:435–443.

802 86. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read
803 sequencing. ArXiv12073907 Q-Bio.

804 87. Rahman A, Pachter L. 2013. CGAL: computing genome assembly likelihoods.
805 Genome Biol 14:R8.

806 88. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence
807 classification using exact alignments. Genome Biol 15:R46.

808 89. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics
809 btu153.

810 90. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M,
811 Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan
812 genome analysis. Bioinformatics 31:3691–3693.

813 91. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related
814 whole genomes. Bioinformatics 27:334–342.

815 92. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-
816 analysis of large phylogenies. Bioinformatics 30:1312–1313.

817 93. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2016. ggtree: an R package for
818 visualization and annotation of phylogenetic trees with their covariates and other
819 associated data. Methods Ecol Evol n/a-n/a.

820 94. Mita SD, Siol M. 2012. EggLib: processing, analysis and simulation tools for
821 population genetics and genomics. BMC Genet 13:27.

822 95. Szpiech ZA, Hernandez RD. 2014. selscan: An Efficient Multithreaded Program to
823 Perform EHH-Based Scans for Positive Selection. Mol Biol Evol 31:2824–2827.

824 96. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ,
825 Marra MA. 2009. Circos: An information aesthetic for comparative genomics.
826 Genome Res 19:1639–1645.

827 97. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016.
828 SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments.
829 biorxiv;038190v1.

830 98. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X,
831 Ruden DM. 2012. A program for annotating and predicting the effects of single
832 nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80–92.

833 99. Sagulenko P, Puller V, Neher R. 2017. TreeTime: maximum likelihood
834 phylodynamic analysis. bioRxiv 153494.

835 100. Maurer LM, Tomasini-Johansson BR, Mosher DF. 2010. Emerging roles of
836 fibronectin in thrombosis. Thromb Res 125:287–291.

837 101. Annis DS, Murphy-Ullrich JE, Mosher DF. 2006. Function-blocking
838 antithrombospondin-1 monoclonal antibodies. J Thromb Haemost 4:459–468.

839

840

841 **Table 1. Single nucleotide polymorphisms with F_{ST} values in the top 0.05%**

842 **between 1,760,000 and 1,820,000 bp in ATCC 15305.**

Position	Frequency in human associated isolates	Frequency in non-human associated isolates	F_{ST} Value	Type
1772616	0.72	0.16	0.5	Non-synonymous
1797190	0.9	0.37	0.48	Synonymous
1808274	0.8	0.21	0.52	Synonymous
1811585	0.72	0.16	0.5	Synonymous
1811777	0.9	0.37	0.48	Non-synonymous
1813204	0.74	0.16	0.53	Synonymous
1816895	0.77	0.05	0.71	Non-synonymous
1818150	0.77	0.16	0.56	Intergenic
1818151	0.77	0.16	0.56	Intergenic
1818156	0.77	0.16	0.56	Intergenic

843

844 **Table 2. Results of demographic inference.** $\nu = N_e/N_{ancestral}$, τ = generations/ $N_{ancestral}$

Model	Optimized parameters (standard deviation)	Log Likelihood	p-value (comparison model)
constant size		-562	
instantaneous change	ν: 9.9×10^{4} (6.0×10^{4}) τ: 4.4×10^{-2} (4.7×10^{-3})	-455	0.0 (constant size)
exponential change	ν: 9.7×10^{4} (9.8×10^{4}) τ: 4.1×10^{-2} (4.9×10^{-3})	-455	0.0 (constant size)
instantaneous change followed by exponential change	ν_A: 2.1×10^{-2} (2.9×10^{-2}) ν_B: 1.1 (3.9×10^{-2}) τ: 1.3×10^{-1} (1.6×10^{-1})	-404	1.6×10^{-4} (exponential change)
two instantaneous size changes	ν_A: 2.9×10^{-2} (1.2×10^{-2}) ν_B: 4.5×10^{-1} (2.1×10^{-1}) τ_A: 1.2×10^{-1} (3.4×10^{-2}) τ_B: 3.1×10^{-3} (5.3×10^{-3})	-393	4.6×10^{-7} (instantaneous change)

845

846

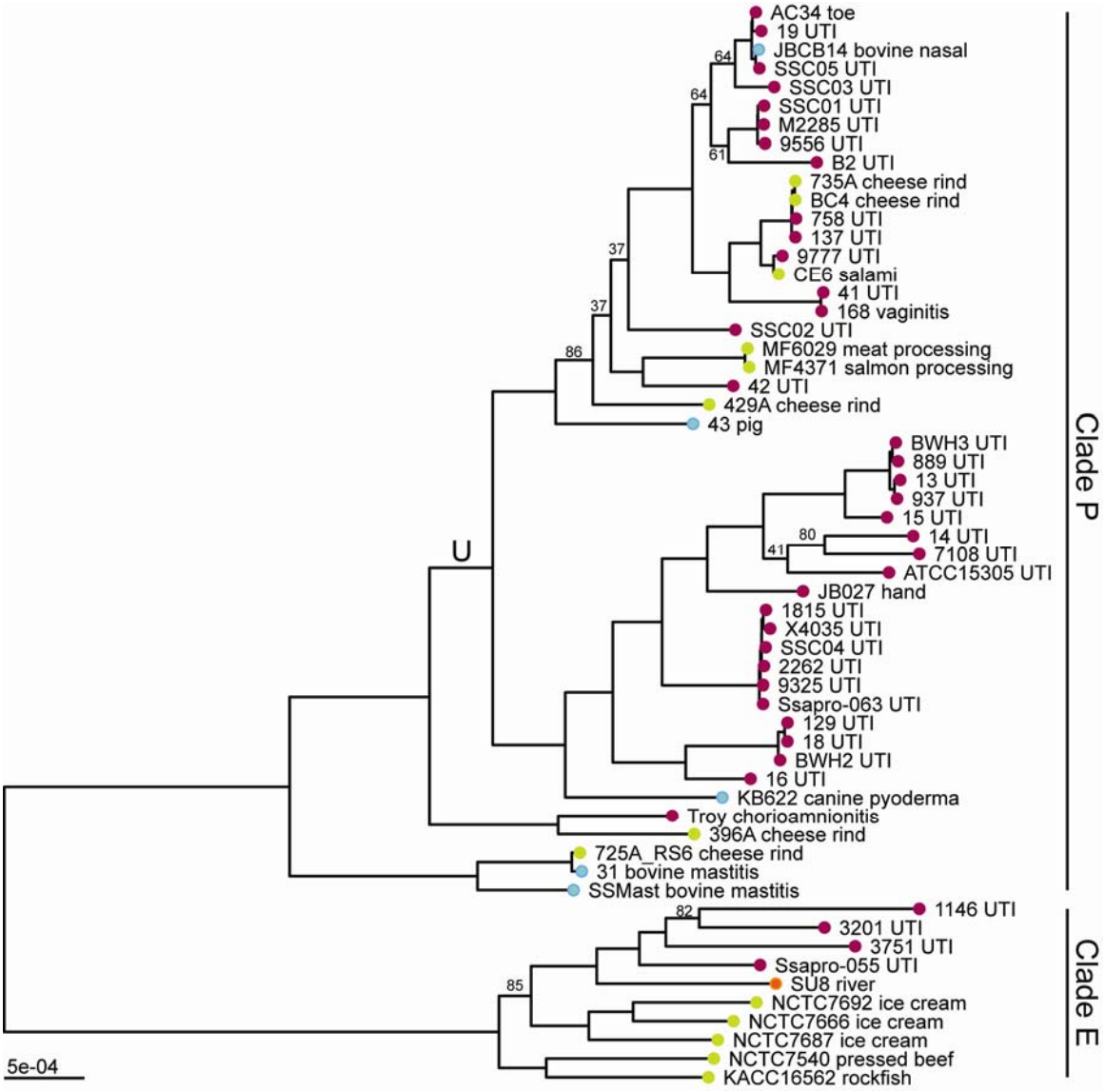


Figure 1. Maximum likelihood phylogeny of *S. saprophyticus*. Maximum likelihood phylogenetic analysis was performed in RAxML (92) using a whole genome alignment with repetitive regions masked. The phylogeny is midpoint rooted, and nodes with bootstrap values less than 90 are labelled. Branch lengths are scaled by substitutions per site. Tips are colored based on the isolation source (pink- human, blue- animal, green- food, orange- environment). Tips are labeled with isolate name and detailed source information. *S. saprophyticus* contains two major clades (Clade P and Clade E). Within Clade P, there is a lineage enriched in human pathogenic isolates (lineage U, branch labeled 'U').

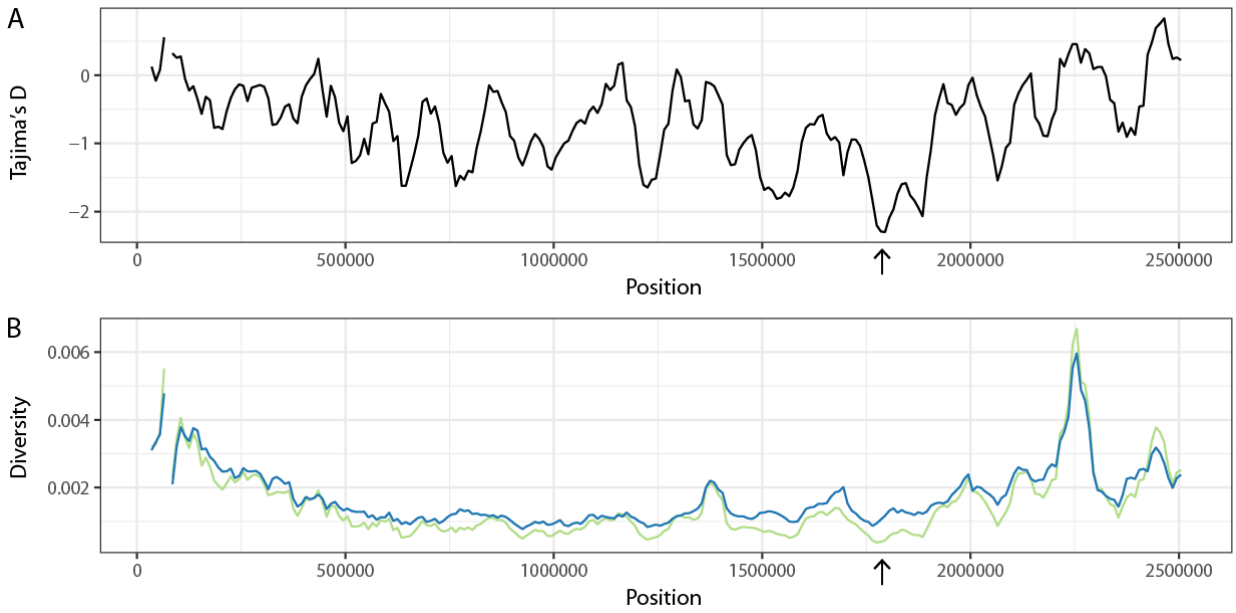
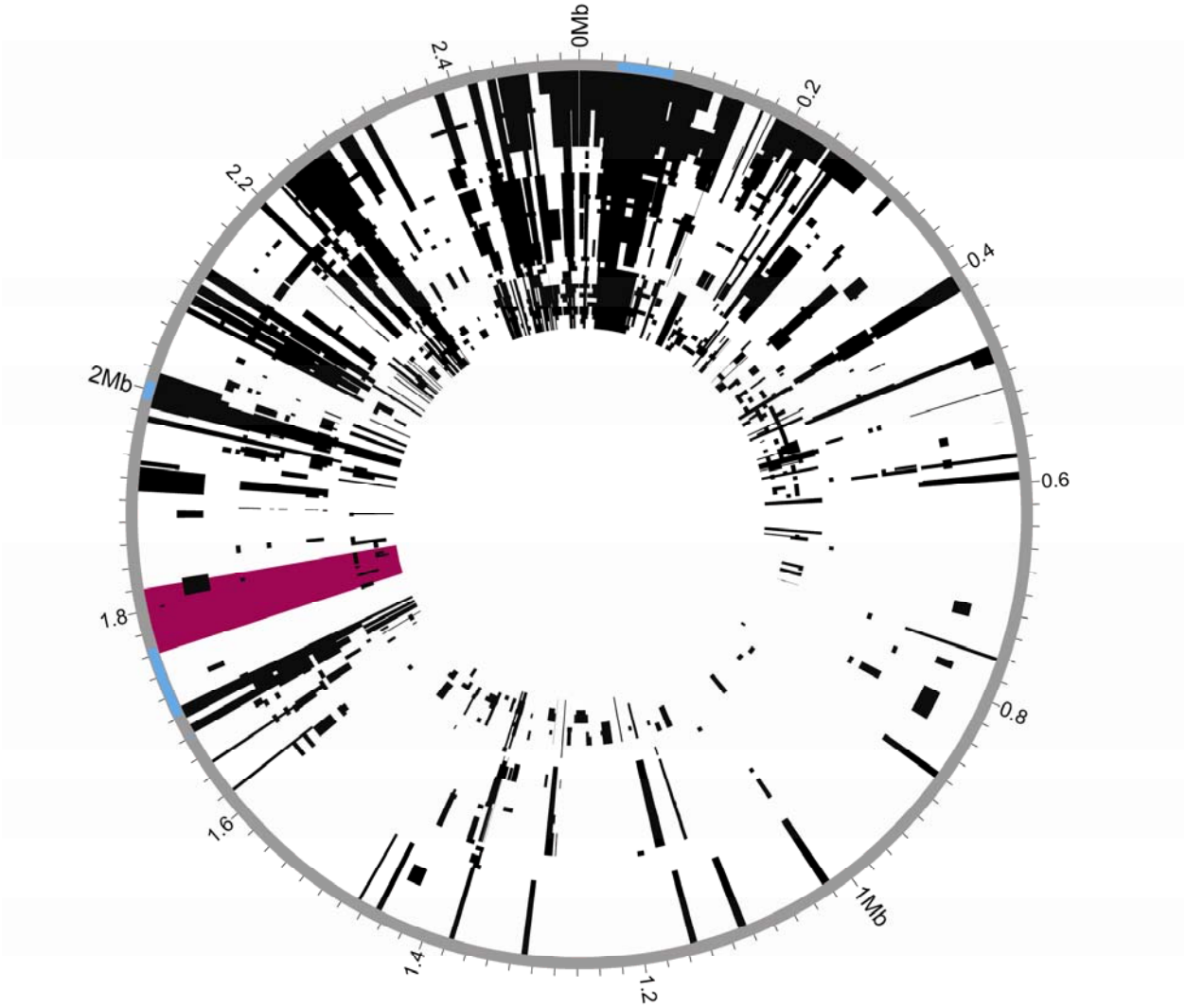


Figure 2. Sliding window analysis of diversity and neutrality statistics. Population genetic statistics were calculated for lineage U using EggLib (94). Windows were 50 kb in width with a step size of 10 kb. A) Tajima's D. B) π (green) and θ (blue). The lowest values for Tajima's D and π are found in the same window (1,760,000-1,820,000 bp, arrow).



863
864 **Figure 3. Recombination in *S. saprophyticus*.** Recombinant regions in the whole
865 genome alignment of *S. saprophyticus* were identified using Gubbins (20). Mobile
866 genetic elements are highlighted in blue on the outer rim. The window with low Tajima's
867 D and π is highlighted in pink. Few recombination events are inferred within this region.

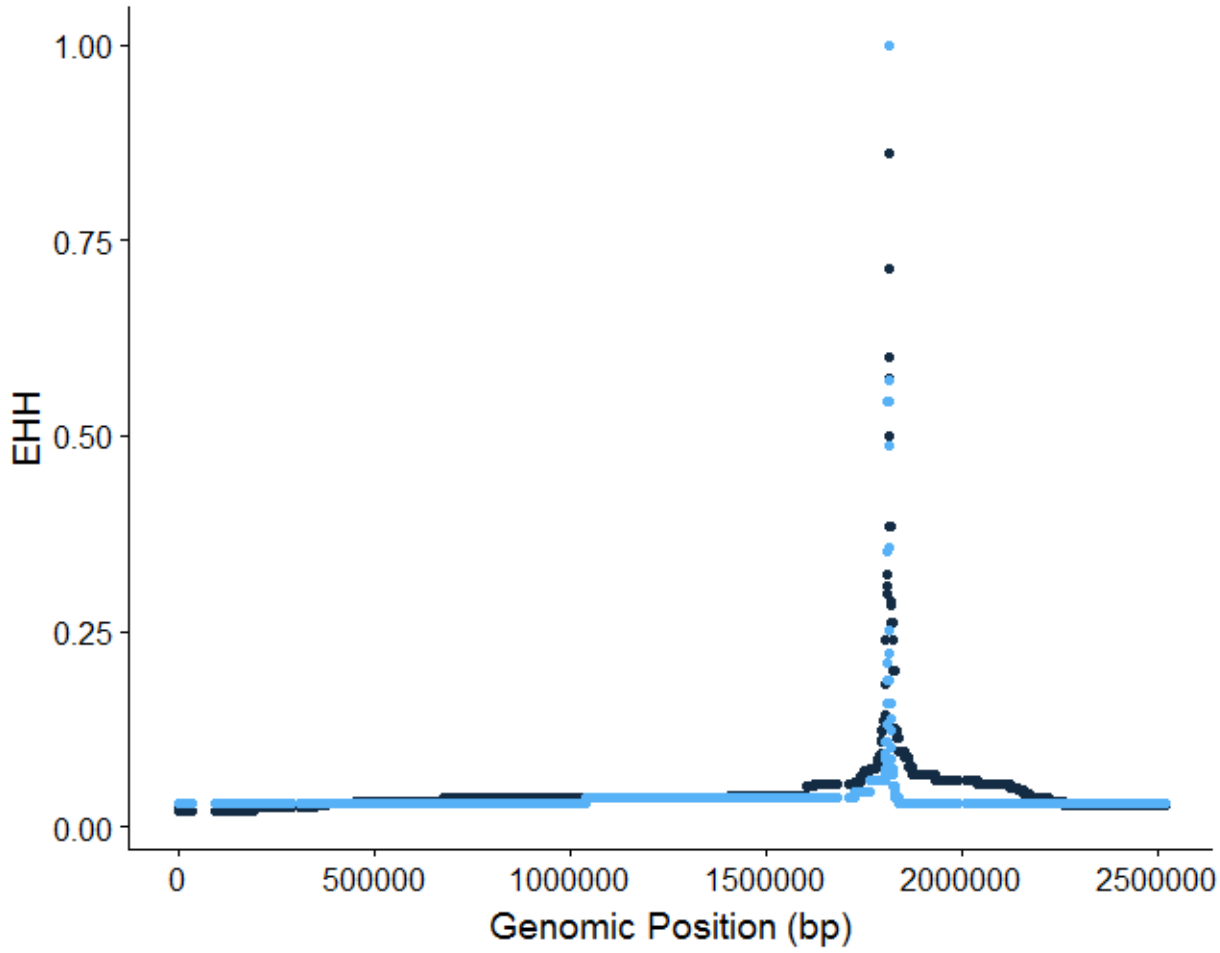


Figure 4. Extended Haplotype Homozygosity (EHH) of single nucleotide polymorphism at position 1811777. EHH values for the ancestral allele are in light blue. EHH values for the derived allele are in dark blue.

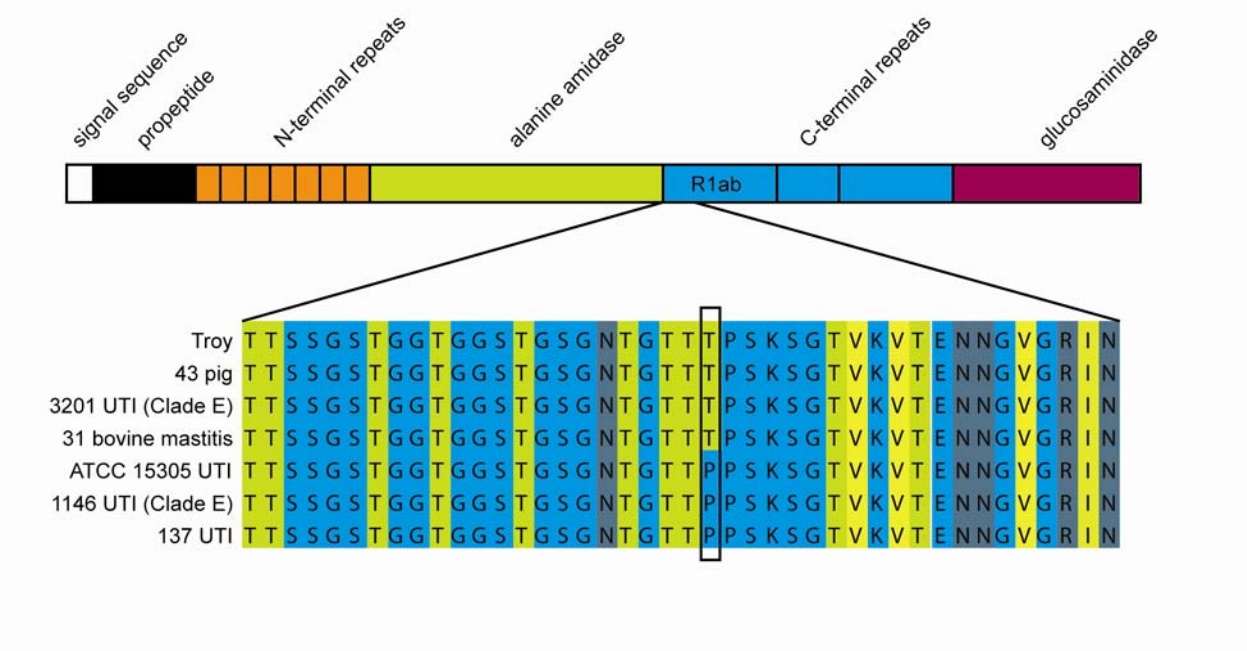


Figure 5. Non-synonymous variant in Aas fibronectin binding repeat. Top- Domains of Aas protein adapted from Hell et al. 1998. R1ab is the peptide used in the fibronectin and thrombospondin binding experiments. Bottom- Alignment of a portion of R1 showing amino acid sequence in Aas from selected *S. saprophyticus* strains. Amino acids are colored based on their propensity to form beta strands (light green=high propensity, light blue=low propensity). The alignment visualization was created in JalView.

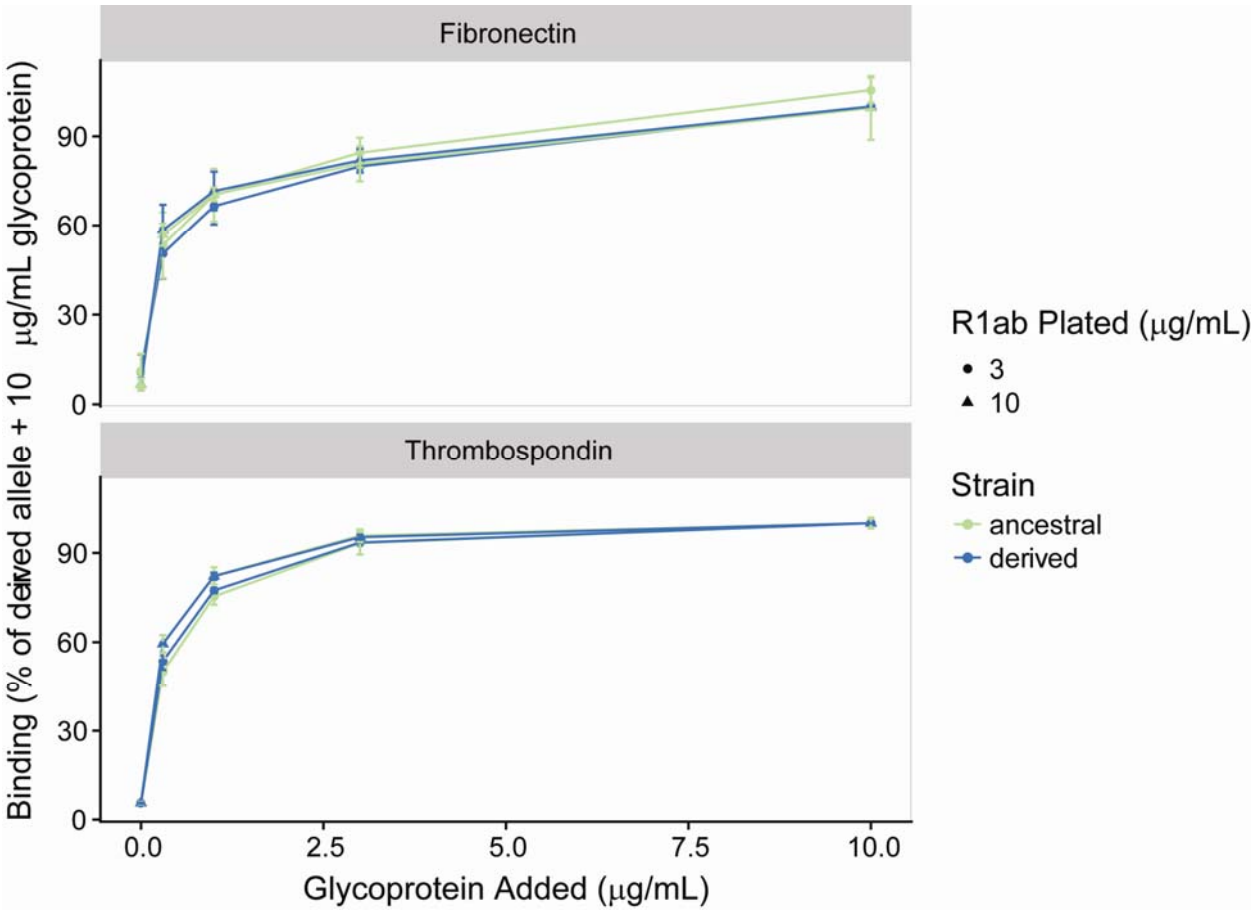
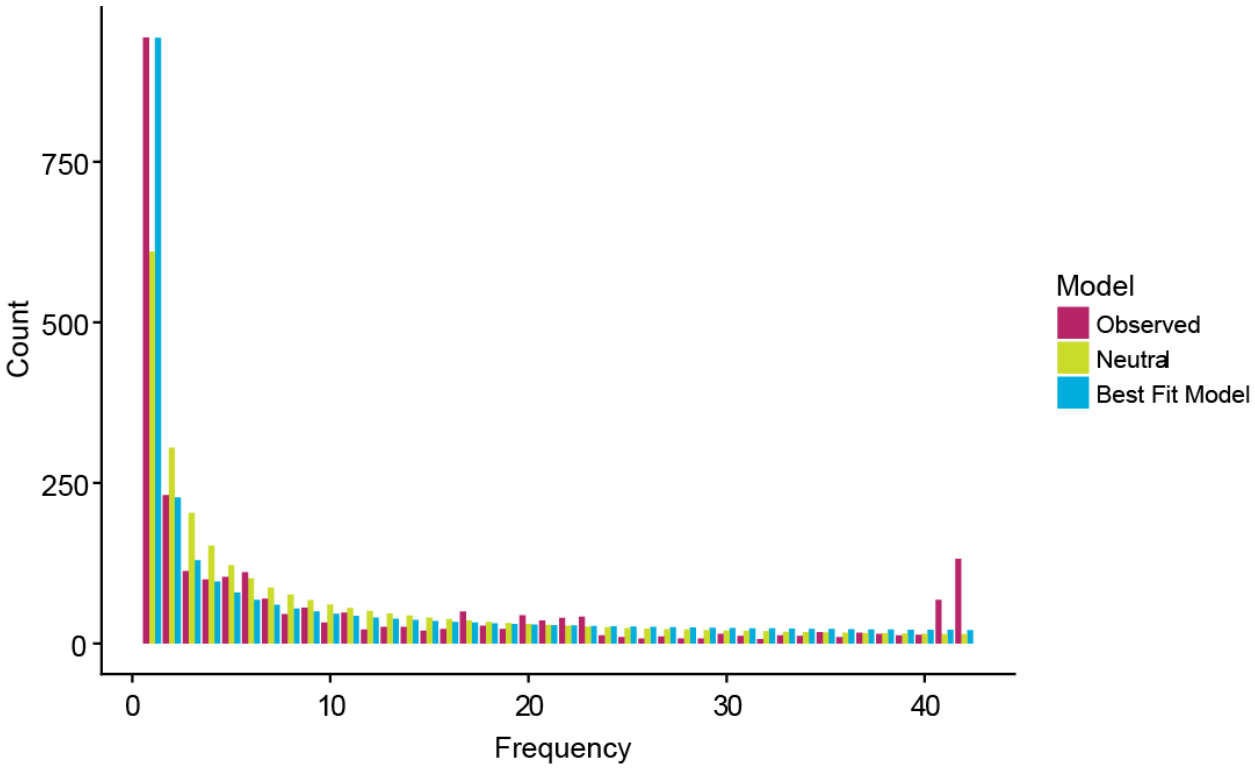


Figure 6. Fibronectin and thrombospondin binding to human-associated and ancestral strain Aas R1ab. ELISAs detecting the binding of soluble human fibronectin and thrombospondin to plates coated with Aas R1ab at 3 and 1 ug/ml. Results normalized to percent of binding of 10ug/ml glycoprotein to human-associated strain R1ab. Human and bovine fibronectin and human thrombospondin bound to the two constructs equally well.



887

888 **Figure 7. Site frequency spectrum of lineage U.** The ancient genome (Troy) was
889 used as the outgroup to determine the ancestral state. Synonymous, nonsynonymous,
890 and intergenic sites were identified with SnpEff (98). The observed synonymous SFS
891 contain an excess of singletons and high frequency derived variants. Both the observed
892 SFS and the SFS predicted by the best fitting model have an excess of singletons
893 compared to the SFS expected under the standard neutral model with no population
894 size change.

895

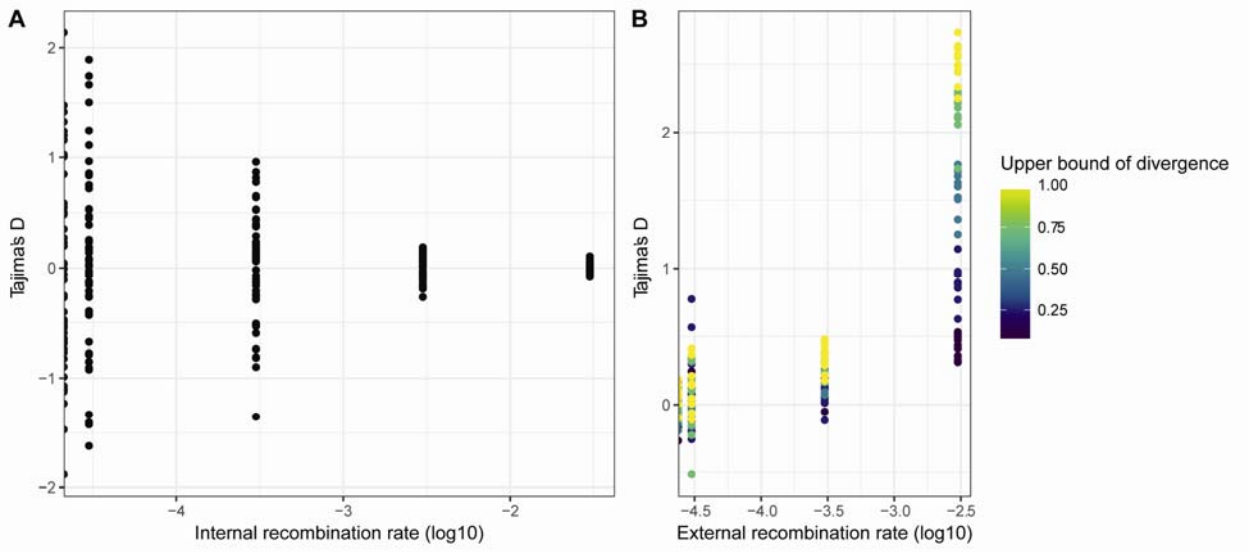


Figure 8. Effects of internal and external recombination on Tajima's D. Bacterial populations with a range of recombination rates were simulated with SimBac. A) Taijma's D values from simulations of internal recombination rates ranging from 0-0.03 and no external recombination. B) Tajima's D values from simulations with an internal recombination rate of 0.003 (r/m = 1) and external recombination rates ranging from 0-0.003. Points are filled according to the upper limit of diversity in external recombinant fragments.

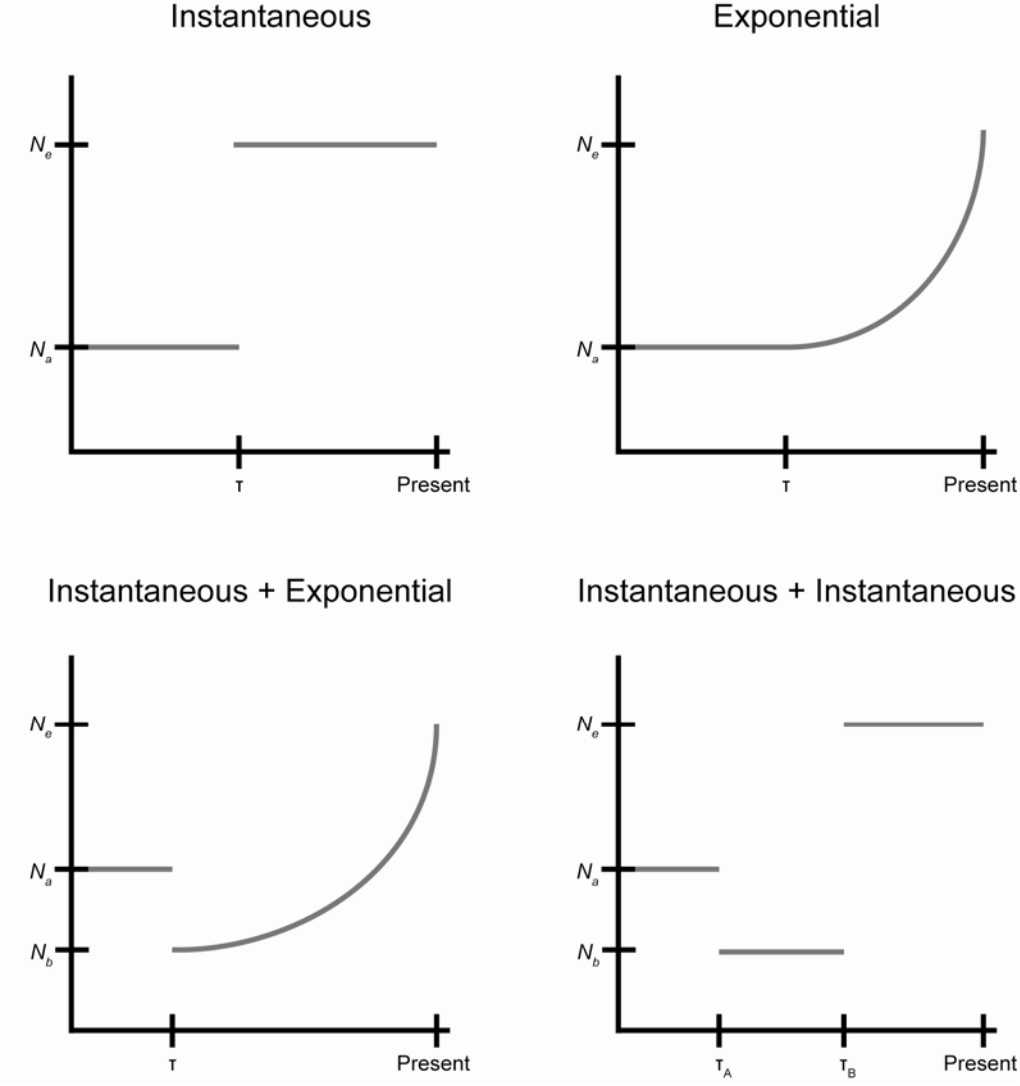


Figure 9. Cartoon of fitted demographic models. The observed synonymous SFS was fit to 5 demographic models including constant size, instantaneous population size change, exponential population size change, instantaneous population size change followed by exponential, and two instantaneous population size changes**.** Parameters for the instantaneous and exponential models are the magnitude of the population size change ($\nu = N_e/N_{ancestral}$) and the timing of the change (τ = generations/ $N_{ancestral}$**).** For models with two population size changes, magnitudes are reported as $\nu_A = N_b/N_{ancestral}$ and $\nu_b = N_e/N_{ancestral}$.

916

917 **Supplementary Tables:**

918 **Table S1.** Accession numbers for *S. saprophyticus* isolates.

919 **Table S2.** Assembly statistics for *S. saprophyticus* genomes.

920