**Title:** Adaptation of *Staphylococcus saprophyticus* at the host-pathogen interface

**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. The mutation under selection affects binding to host fibronectin, and we hypothesize that it 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.

**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 suggests 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 that these virulence factors play 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 affects bacterial binding to a protein at the host-pathogen interface and 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 (Supplementary Table 1) 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 an intermediate stage in the emergence of a human-adapted lineage from a zoonotic background.

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*). We did not identify any genes that were significantly associated with the human pathogenic niche after correction for multiple hypothesis testing.

In addition to the variability observed in gene content, analyses of the core genome also indicated relatively frequent recombination among *S. saprophyticus*. We identified recombinant regions with Gubbins (*20*), which indicated that 70% of sites in the *S. saprophyticus* alignment have been affected by recombination. There are 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, Supplementary Figure 1). The region with decreased π /TD corresponds to 1,760,000-1,820,000 bp in *S. saprophyticus* ATCC 15305. Few recombinant tracts are found here (Figure 3), suggesting that the locus may be functionally important. 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 (Supplementary Figure 2).

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 (Supplementary Table 2). 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 a presumably 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*) (Supplementary Figure 3). 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*)). 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 4, (*31*)), and we hypothesized that it is 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, as a control, thrombospondin-1, which binds to this region of the homologous AtlE amidase from *Staphylococcus epidermidis* (*33*). Recombinant proteins comprising Aas R1a1b of the ancestral or derived alleles were adsorbed to plates at varying concentrations to which we added soluble human fibronectin, bovine fibronectin, or human thrombospondin-1. Binding was assessed by ELISA with antibodies to the soluble fibronectin and thrombospondin-1. We confirmed that R1a1b binds to both fibronectin and thrombospondin. There was no difference between alleles of R1a1b in binding to thrombospondin-1, but the derived, human-associated variant bound more fibronectin than the ancestral version across the range of concentrations of bacterial peptide we tested (Figure 5). This difference was seen in binding to both human and bovine fibronectin (not shown).

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 (*34–37*). 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 (*38*) to identify the demographic model that best fit the observed synonymous SFS of lineage U (Figure 6).

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. We used fastGEAR (*39*) to identify recombinant tracts that originated outside of lineage U and contained ancestral alleles. 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) 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 (*40*). We found significant evidence for an expansion in all models (Table 1). 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 ($\nu = N_e/N_{anc}$, $\tau$ = generations/$N_{anc}$, $\nu_A$: 2.9 x $10^{-2}$, $\nu_B$: 4.5 x $10^{-1}$, $\tau_A$: 1.2 x $10^{-1}$, $\tau_B$: 3.1 x $10^{-3}$).

Recombination and positive selection are known to confound the inference of bacterial demography (*41*) and so we used simulations to investigate their effects on our demographic inference for uropathogenic *S. saprophyticus.* We used SFS_CODE (*42*) 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 (*41*, *43*). 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 affects the adhesin's binding to host fibronectin, 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 (*44*), 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 (*45*, *46*). 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 fibronectin binding 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. Given its position on the phylogeny, the Troy lineage may represent an intermediate step in a host shift from livestock to humans. 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.

The Troy bacterium pairs with a modern isolate from cheese, and food-derived isolates are interspersed throughout the phylogeny. Previous studies of the molecular epidemiology of *S.*

*saprophyticus* using pulsed field gel electrophoresis (PFGE) found that clinical isolates and isolates from fermented food had a diversity of PFGE patterns without clustering by sample type or geography (*44*, *47*). Clinical isolates did not share the same patterns as food-associated isolates in these studies. However, our phylogeny, based on whole genome sequencing, suggests that human- and animal-pathogenic strains of *S. saprophyticus* are found in food.

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 fibronectin-binding 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*, *48*). 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 (*49*). For example, multiple clonal replacements have occurred in *Shigella sonnei* populations in Vietnam due to acquisition of resistance to antimicrobials and environmental stress (*50*). Recurrent clonal replacements have also been observed within single hosts during chronic infection of cystic fibrosis patients by *Pseudomonas aeruginosa* (*51*). 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 (*52*). 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 (*53*).

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* (*54*), which is consistent with our finding that clinical phenotypes can represent distinct niche spaces preferentially occupied by sub-populations of bacteria. There is also precedent for a single nucleotide polymorphism to affect host tropism of bacteria, although the mechanism was not identified in that case (*55*).

In sexually reproducing organisms, haplotype based statistics are frequently used to identify selective sweeps because positively selected alleles will also increase the frequency of nearby linked loci faster than recombination can disrupt linkage, producing longer haplotypes for selected alleles (*28*, *56*). We found that *aas*_2206A>C had a longer haplotype than the ancestral variant, but this difference was not extreme relative to other regions of the genome (assessed with the $nS_L$ statistic). Haplotype based statistics have been found to perform poorly in purebred dogs, where linkage across the genome is high (*57*). Relatively low levels of recombination may also contribute to a lack of sensitivity when haplotype-based detection methods are applied to bacteria; linkage of sites is also likely to be disrupted in a less predictable way by bacterial gene conversion than by crossing over (*29*). Based on our findings, we conclude that screening for regional decreases in diversity and distortions of the SFS (i.e. sliding window analyses) and identification of genetic variants with extreme differences in frequency between niches can be useful in identifying candidate sites of positive selection in bacteria.

*S. saprophyticus* encodes a number of adhesins including UafA, UafB, SdrI, and Aas. UafA and Aas are found in all isolates, suggesting that they play important roles in the diverse niches occupied by *S. saprophyticus*. Aas has autolytic, fibronectin binding, and haemagluttinating functions (*9*, *31*, *32*, *58*). We identified a single, non-synonymous polymorphism as a target of selection in the fibronectin binding repeats of *Aas*. This variant is predicted to affect the repeat's structure, and we have shown that the amino acid change associated with *aas*_2206A>C affects binding to both human and bovine fibronectin. Adhesins are plausible candidates for adaptation to the uropathogenic niche, as they are known to be important virulence factors in pathogens causing urinary tract infections (*59*). Fibronectin binding proteins including Aas have been identified as virulence factors in in *S. saprophyticus* and *Enterococcus faecalis* (*32*, *60*, *61*). Adhesion to the uroepithelium is essential for uropathogens to establish themselves in the bladder, where they are subject to strong shear stress (*62*): 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 (*63–65*). FimH binds mannose, providing protection from shear stress through a catch bond mechanism (*66*). 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 (*67*). 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 (*68*).

Further experiments are needed to investigate the effects of variation in Aas on *S. saprophyticus* biology. 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* (*69*), 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 (*41*, *70*). 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 recombination (*41*) and this was confirmed in our analyses of simulated data. We used only synonymous sites in our inference, which should be less affected by selection. 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.

Population expansion after introduction to a new niche has been described in other bacterial pathogens, including an expansion in *S. aureus* after transatlantic spread (*35*), an expansion in *Clostridium difficile* after introduction to the hospital environment (*34*), and an expansion in *S. pyogenes* after horizontal transfer of toxin genes that allow the species to cause scarlet fever (*37*). These previously described expansions were, however, identified with Bayesian skyline plots, which can be affected by recombination and selection (*41*).

We observed an excess of high frequency derived variants in the SFS of lineage U, many of which were the result of lateral gene transfer from outside the lineage as identified by fastGEAR (*37*). After removing sites inferred to have been laterally transferred, some high frequency derived variants remained. While these could represent unidentified recombinant regions or misidentification of the ancestral state, they could also reflect a multiple merger coalescent process driven by positive selection (*43*).

Here we have described adaptation of *S. saprophyticus* that 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. We have further shown that the mutation under selection affects binding to fibronectin, providing a likely mechanism by which it enables invasion of the human urinary tract. 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 and centrigured at 4°C for 10 minutes. We additionally used a SpeedVac for 10 minutes to ensure pellets were dry before re-suspending the pellet in 50 µL water.

*Library preparation and sequencing.* For SSC01, SSC02, and SSC03, library prep was performed using a modified Nextera protocol  as decribed by Baym et al. (*71*) with a reconditioning PCR with fresh primers and polymerase for an additional 5 PCR cycles to minimize chimeras and a two-step bead based size selection with target fragment size of 650 bp and sequenced on an Illumina HiSeq 2500 (paired-end, 150 bp). For 43, SSC04, SCC05, and SSMast, DNA was submitted to the University of Wisconsin-Madison Biotechnology Center for library preparation and were prepared according the TruSeq Nano DNA LT Library Prep Kit (Illumina Inc., San Diego, California, USA) with minor modifications. A maximum of 200 ng of each sample was sheared using a Covaris M220 Ultrasonicator (Covaris Inc, Woburn, MA, USA). Sheared samples were size selected for an average insert size of 550 bp using Spri bead based size exclusion. Quality and quantity of the finished libraries were assessed using an Agilent DNA High Sensitivity chip (Agilent Technologies, Santa Clara, CA) and Qubit dsDNA HS Assay Kit, respectively. Libraries were standardized to 2 µM. Paired-end, 150 bp sequencing was performed using v2 SBS chemistry on an Illumina MiSeq sequencer. Images were analyzed using the Illumina Pipeline, version 1.8.2.

*Reference guided mapping.* When sequencing reads were available, we mapped reads to ATCC 15305 using via a pipeline (available at https://github.com/tracysmith/RGAPepPipe). Briefly, read quality was assessed and trimmed with TrimGalore! v 0.4.0 (www.bioinformatics.babraham.ac.uk/projects/trim_galore), which runs both FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc) and cutadapt. Reads were mapped to using BWA-MEM v 0.7.12 (*72*) and sorted using Samtools v 1.2 (*73*). We used Picard v 1.138 (picard.sourceforge.net) to add read group information and removed duplicates. Reads were locally realigned using GATK v 2.8.1 (*74*). We identified variants using Pilon v 1.16 (*75*) (minimum read depth: 10, minimum mapping quality: 40, minimum base quality: 20).

*Assembly.* We used the iMetAMOS pipeline for *de novo* assembly (*76*).  We chose to compare assemblies from SPAdes (*77*), MaSurCA (*78*), and Velvet (*79*). KmerGenie (*80*) was used to select kmer sizes for assembly. iMetAMOS uses FastQC, QUAST (*81*), REAPR (*82*), LAP (*83*), ALE (*84*), FreeBayes (*85*), and CGAL (*86*) to evaluate the quality of reads and assemblies. We also used Kraken (*87*) to detect potential contamination in sequence data. For all newly assembled isolates (43, SSC01-05, SSMast), the SPAdes assembly was the highest quality.

*Annotation and gene content analyses.* We annotated the de novo assemblies using Prokka v 1.11 (*88*) and used Roary (*89*) 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 (*90*). 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 (*91*). We used the GTRGAMMA substitution model and performed bootstrapping using the autoMR convergence criteria. Tree visualizations were created in ggtree (*92*).

*Population genetics statistics.* To calculate π and Tajima's D, we used EggLib v 2.1.10 (*93*), 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 (*94*) to calculate $nS_L$.

*Recombination analyses.* To identify recombinant regions in the *S. saprophyticus* alignment, we used Gubbins v 2.1.0 (*20*). fastGEAR (*39*) was used with the recommended input specifications to identify recombination events between major lineages of *S. saprophyticus*. We used Circos (*95*) for visualization of recombinant tracts.

*Site frequency spectrum.* We used SNP-sites v 2.0.3 (*96*) to convert the alignment of *S. saprophyticus* isolates to a multi-sample VCF. SnpEff v 4.1j (*97*) 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 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.

*Demography.* We performed demographic inference with the synonymous SFS using ∂a∂i (*38*). Models tested were the standard neutral model, expansion, and exponential growth. Parameters, $\nu$ ($N_e/$) and $\tau$ (time scaled by 2 $N_a$), 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 SFS_CODE (release date 9/10/2015) (*42*) 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 ($\gamma$ = 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 4ºC. 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 (*98*) or platelet-derived thrombospondin-1 (*99*) 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.

**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. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256259. 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 Institute of Health National Research Service Award (T32 GM07215). CSP is supported by National Institutes of Health (R01AI113287).

**References.**

1. W. E. Stamm, S. R. Norrby, Urinary Tract Infections: Disease Panorama and Challenges. *J. Infect. Dis.* **183**, S1–S4 (2001).

2. A. R. Ronald *et al.*, Urinary tract infection in adults: research priorities and strategies. *Int. J. Antimicrob. Agents*. **17**, 343–348 (2001).

3. A. E. Barber, J. P. Norton, A. M. Spivak, M. A. Mulvey, Urinary Tract Infections: Current and Emerging Management Strategies. *Clin. Infect. Dis.* **57**, 719–724 (2013).

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

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

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

7. A. M. Devault *et al.*, A molecular portrait of maternal sepsis from Byzantine Troy. *eLife*. **6**, e20983 (2017).

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

9. H. G. Meyer, U. Wengler-Becker, S. G. Gatermann, The hemagglutinin of Staphylococcus saprophyticus is a major adhesin for uroepithelial cells. *Infect. Immun.* **64**, 3893–3896 (1996).

10. M. Kuroda *et al.*, Whole genome sequence of Staphylococcus saprophyticus reveals the pathogenesis of uncomplicated urinary tract infection. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13272–13277 (2005).

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

12. E. J. Nelson, J. B. Harris, J. Glenn Morris, S. B. Calderwood, A. Camilli, Cholera transmission: the host, pathogen and bacteriophage dynamic. *Nat. Rev. Microbiol.* **7**, 693–702 (2009).

13. Y. Boucher, F. D. Orata, M. Alam, 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** (2015), doi:10.3389/fmicb.2015.01120.

14. F. J. Veyrier, A. Dufort, M. A. Behr, The rise and fall of the Mycobacterium tuberculosis genome. *Trends Microbiol.* **19**, 156–161 (2011).

15. A. McNally, N. R. Thomson, S. Reuter, B. W. Wren, “Add, stir and reduce”: Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.* **14**, 177–190 (2016).

16. P. Larsson *et al.*, Molecular Evolutionary Consequences of Niche Restriction in Francisella tularensis, a Facultative Intracellular Pathogen. *PLoS Pathog*. **5**, e1000472 (2009).

17. F. Lebreton *et al.*, Emergence of Epidemic Multidrug-Resistant Enterococcus faecium from Animal and Commensal Strains. *mBio*. **4**, e00534-13 (2013).

18. J. R. Fitzgerald *et al.*, Recombination-mediated remodelling of host-pathogen interactions during Staphylococcus aureus niche adaptation. *Microb. Genomics* (2015), doi:10.1099/mgen.0.000036.

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

20. N. J. Croucher *et al.*, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15–e15 (2015).

21. J. M. Smith, J. Haigh, The hitch-hiking effect of a favourable gene. *Genet. Res.* (1974), (available at /core/journals/genetics-research/article/the-hitch-hiking-effect-of-a-favourable-gene/918291A3B62BD50E1AE5C1F22165EF1B).

22. J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, W. Stephan, The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*. **140**, 783–796 (1995).

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

24. S. Nair *et al.*, A Selective Sweep Driven by Pyrimethamine Treatment in Southeast Asian Malaria Parasites. *Mol. Biol. Evol.* **20**, 1526–1536 (2003).

25. S. I. Wright, B. S. Gaut, Molecular Population Genetics and the Search for Adaptive Evolution in Plants. *Mol. Biol. Evol.* **22**, 506–519 (2005).

26. J. J. de Simoni Gouveia, M. V. G. B. da Silva, S. R. Paiva, S. M. P. de Oliveira, Identification of selection signatures in livestock species. *Genet. Mol. Biol.* **37**, 330–342 (2014).

27. B. S. Weir, C. C. Cockerham, Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. **38**, 1358–1370 (1984).

28. P. C. Sabeti *et al.*, Detecting recent positive selection in the human genome from haplotype structure. *Nature*. **419**, 832–837 (2002).

29. B. J. Shapiro, in *Ecological Genomics*, C. R. Landry, N. Aubin-Horth, Eds. (Springer Netherlands, 2014; http://link.springer.com/chapter/10.1007/978-94-007-7347-9_17), *Advances in Experimental Medicine and Biology*, pp. 339–359.

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

31. W. Hell, H.-G. W. Meyer, S. G. Gatermann, Cloning of aas, a gene encoding a Staphylococcus saprophyticus surface protein with adhesive and autolytic properties. *Mol. Microbiol.* **29**, 871–881 (1998).

32. S. Gatermann, H. G. Meyer, Staphylococcus saprophyticus hemagglutinin binds fibronectin. *Infect. Immun.* **62**, 4556–4563 (1994).

33. T. P. Kohler *et al.*, Repeating Structures of the Major Staphylococcal Autolysin Are Essential for the Interaction with Human Thrombospondin 1 and Vitronectin. *J. Biol. Chem.* **289**, 4070–4082 (2014).

34. M. He *et al.*, Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc. Natl. Acad. Sci.* **107**, 7527–7532 (2010).

35. U. Nübel *et al.*, A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant Staphylococcus aureus. *PLOS Pathog.* **6**, e1000855 (2010).

36. C. S. Pepperell *et al.*, The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. *PLoS Pathog*. **9**, e1003543 (2013).

37. M. R. Davies *et al.*, Emergence of scarlet fever Streptococcus pyogenes emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat. Genet.* **47**, 84–87 (2015).

38. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*. **5**, e1000695 (2009).

39. R. Mostowy *et al.*, Analysis of recent and ancestral recombination reveals high-resolution population structure in Streptococcus pneumoniae. *bioRxiv*, 059642 (2016).

40. A. J. Coffman, P. H. Hsieh, S. Gravel, R. N. Gutenkunst, Computationally Efficient Composite Likelihood Statistics for Demographic Inference. *Mol. Biol. Evol.* **33**, 591–593 (2016).

41. M. Lapierre, C. Blin, A. Lambert, G. Achaz, E. P. C. Rocha, The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol. Biol. Evol.*, msw048 (2016).

42. R. D. Hernandez, A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*. **24**, 2786–2787 (2008).

43. R. A. Neher, O. Hallatschek, Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci.* **110**, 437–442 (2013).

44. M. Widerstrom, J. Wistrom, S. Ferry, C. Karlsson, T. Monsen, Molecular Epidemiology of Staphylococcus saprophyticus Isolated from Women with Uncomplicated Community-Acquired Urinary Tract Infection. *J. Clin. Microbiol.* **45**, 1561–1564 (2007).

45. P. Hedman, O. Ringertz, M. Lindström, K. Olsson, The origin of Staphylococcus saprophyticus from cattle and pigs. *Scand. J. Infect. Dis.* **25**, 57–60 (1993).

46. E. K. Kastman *et al.*, Biotic Interactions Shape the Ecological Distributions of Staphylococcus Species. *mBio*. **7**, e01157-16 (2016).

47. E. Coton *et al.*, Biodiversity of Coagulase-Negative Staphylococci in French cheeses, dry fermented sausages, processing environments and clinical samples. *Int. J. Food Microbiol.* **137**, 221–229 (2010).

48. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**, 585–595 (1989).

49. B. J. Shapiro, L. A. David, J. Friedman, E. J. Alm, Looking for Darwin's footprints in the microbial world. *Trends Microbiol.* **17**, 196–204 (2009).

50. K. E. Holt *et al.*, Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci.* **110**, 17522–17527 (2013).

51. J. D. Caballero *et al.*, Selective Sweeps and Parallel Pathoadaptation Drive Pseudomonas aeruginosa Evolution in the Cystic Fibrosis Lung. *mBio*. **6**, e00981-15 (2015).

52. M. L. Bendall *et al.*, Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).

53. B. J. Shapiro *et al.*, Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*. **336**, 48–51 (2012).

54. Y.-J. Bao, B. J. Shapiro, S. W. Lee, V. A. Ploplis, F. J. Castellino, Phenotypic differentiation of Streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. *Sci. Rep.* **6**, 36644 (2016).

55. D. Viana *et al.*, A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat. Genet.* **advance online publication** (2015), doi:10.1038/ng.3219.

56. J. J. Vitti, S. R. Grossman, P. C. Sabeti, Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.* **47**, 97–120 (2013).

57. F. Schlamp *et al.*, Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Mol. Ecol.* **25**, 342–356 (2016).

58. H.-G. W. Meyer, J. Müthing, S. G. Gatermann, The hemagglutinin of Staphylococcus saprophyticus binds to a protein receptor on sheep erythrocytes. *Med. Microbiol. Immunol. (Berl.)*. **186**, 37–43 (1997).

59. A. L. Flores-Mireles, J. N. Walker, M. Caparon, S. J. Hultgren, Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat. Rev. Microbiol.* **advance online publication** (2015), doi:10.1038/nrmicro3432.

60. N. P. King *et al.*, UafB is a serine-rich repeat adhesin of Staphylococcus saprophyticus that mediates binding to fibronectin, fibrinogen and human uroepithelial cells. *Microbiology*. **157**, 1161–1175 (2011).

61. R. Torelli *et al.*, The PavA-like Fibronectin-Binding Protein of Enterococcus faecalis, EfbA, Is Important for Virulence in a Mouse Model of Ascending Urinary Tract Infection. *J. Infect. Dis.* **206**, 952–960 (2012).

62. K. J. Wright, S. J. Hultgren, Sticky fibers and uropathogenesis: bacterial adhesins in the urinary tract. *Future Microbiol.* **1**, 75–87 (2006).

63. S. L. Chen *et al.*, Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: A comparative genomics approach. *Proc. Natl. Acad. Sci.* **103**, 5977–5982 (2006).

64. S. L. Chen *et al.*, Positive selection identifies an in vivo role for FimH during urinary tract infection in addition to mannose binding. *Proc. Natl. Acad. Sci.* **106**, 22439–22444 (2009).

65. D. J. Schwartz *et al.*, Positively selected FimH residues enhance virulence during urinary tract infection by altering FimH conformation. *Proc. Natl. Acad. Sci.* **110**, 15530–15537 (2013).

66. O. Yakovenko *et al.*, FimH Forms Catch Bonds That Are Enhanced by Mechanical Force Due to Allosteric Regulation. *J. Biol. Chem.* **283**, 11596–11605 (2008).

67. A. F. Niddam *et al.*, Plasma fibronectin stabilizes Borrelia burgdorferi–endothelial interactions under vascular shear stress by a catch-bond mechanism. *Proc. Natl. Acad. Sci.* **114**, E3490–E3498 (2017).

68. J. A. Messina, J. T. Thaden, B. K. Sharma-Kuinkel, V. G. F. Jr, Impact of Bacterial and Human Genetic Variation on Staphylococcus aureus Infections. *PLOS Pathog.* **12**, e1005330 (2016).

69. V. Da Cunha *et al.*, Streptococcus agalactiae clones infecting humans were selected and fixed through the extensive use of tetracycline. *Nat. Commun.* **5** (2014), doi:10.1038/ncomms5544.

70. J. Hedge, D. J. Wilson, Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. *mBio*. **5**, e02158-14 (2014).

71. M. Baym *et al.*, Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*. **10**, e0128036 (2015).

72. H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM" (arXiv e-print 1303.3997, 2013), (available at http://arxiv.org/abs/1303.3997).

73. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

74. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

75. B. J. Walker *et al.*, Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*. **9**, e112963 (2014).

76. S. Koren, T. J. Treangen, C. M. Hill, M. Pop, A. M. Phillippy, Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*. **15**, 126 (2014).

77. A. Bankevich *et al.*, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

78. A. Zimin *et al.*, The MaSuRCA genome Assembler. *Bioinformatics*, btt476 (2013).

79. D. R. Zerbino, E. Birney, Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

80. R. Chikhi, P. Medvedev, Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. **30**, 31–37 (2014).

81. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies. *Bioinforma. Oxf. Engl.* **29**, 1072–1075 (2013).

82. M. Hunt *et al.*, REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14**, R47 (2013).

83. M. Ghodsi *et al.*, De novo likelihood-based measures for comparing genome assemblies. *BMC Res. Notes*. **6**, 334 (2013).

84. S. C. Clark, R. Egan, P. I. Frazier, Z. Wang, ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*. **29**, 435–443 (2013).

85. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* (2012) (available at http://arxiv.org/abs/1207.3907).

86. A. Rahman, L. Pachter, CGAL: computing genome assembly likelihoods. *Genome Biol.* **14**, R8 (2013).

87. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

88. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, btu153 (2014).

89. A. J. Page *et al.*, Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. **31**, 3691–3693 (2015).

90. S. V. Angiuoli, S. L. Salzberg, Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*. **27**, 334–342 (2011).

91. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).

92. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, *Methods Ecol. Evol.*, in press, doi:10.1111/2041-210X.12628.

93. S. D. Mita, M. Siol, EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* **13**, 27 (2012).

94. Z. A. Szpiech, R. D. Hernandez, selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).

95. M. Krzywinski *et al.*, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

96. A. J. Page *et al.*, “SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments” (biorxiv;038190v1, 2016), (available at http://biorxiv.org/lookup/doi/10.1101/038190).

97. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. **6**, 80–92 (2012).

98. L. M. Maurer, B. R. Tomasini-Johansson, D. F. Mosher, Emerging roles of fibronectin in thrombosis. *Thromb. Res.* **125**, 287–291 (2010).

99. D. S. Annis, J. E. Murphy-Ullrich, D. F. Mosher, Function-blocking antithrombospondin-1 monoclonal antibodies. *J. Thromb. Haemost.* **4**, 459–468 (2006).

**Table 1. Results of demographic inference.** $\nu = N_e/N_{ancestral}$, $\tau$ = generations/ $N_{ancestral}$

| **Model** | **Optimized parameters (standard deviation)** | **Log Likelihood** | **p-value (comparison model)** |
|---|---|---|---|
| constant size | | -562 | |
| instantaneous change | $\nu$: $9.9 \times 10^4$ ($6.0 \times 10^4$)<br>$\tau$: $4.4 \times 10^{-2}$ ($4.7 \times 10^{-3}$) | -455 | 0.0<br>(constant size) |
| exponential change | $\nu$: $9.7 \times 10^4$ ($9.8 \times 10^4$)<br>$\tau$: $4.1 \times 10^{-2}$ ($4.9 \times 10^{-3}$) | -455 | 0.0<br>(constant size) |
| instantaneous change followed by exponential change | $\nu_A$: $2.1 \times 10^{-2}$ ($2.9 \times 10^{-2}$)<br>$\nu_B$: 1.1 ($3.9 \times 10^{-2}$)<br>$\tau$: $1.3 \times 10^{-1}$ ($1.6 \times 10^{-1}$) | -404 | $1.6 \times 10^{-4}$<br>(exponential change) |
| two instantaneous size changes | $\nu_A$: $2.9 \times 10^{-2}$ ($1.2 \times 10^{-2}$)<br>$\nu_B$: $4.5 \times 10^{-1}$ ($2.1 \times 10^{-1}$)<br>$\tau_A$: $1.2 \times 10^{-1}$ ($3.4 \times 10^{-2}$)<br>$\tau_B$: $3.1 \times 10^{-3}$ ($5.3 \times 10^{-3}$) | -393 | $4.6 \times 10^{-7}$<br>(instantaneous change) |

**Figures.**

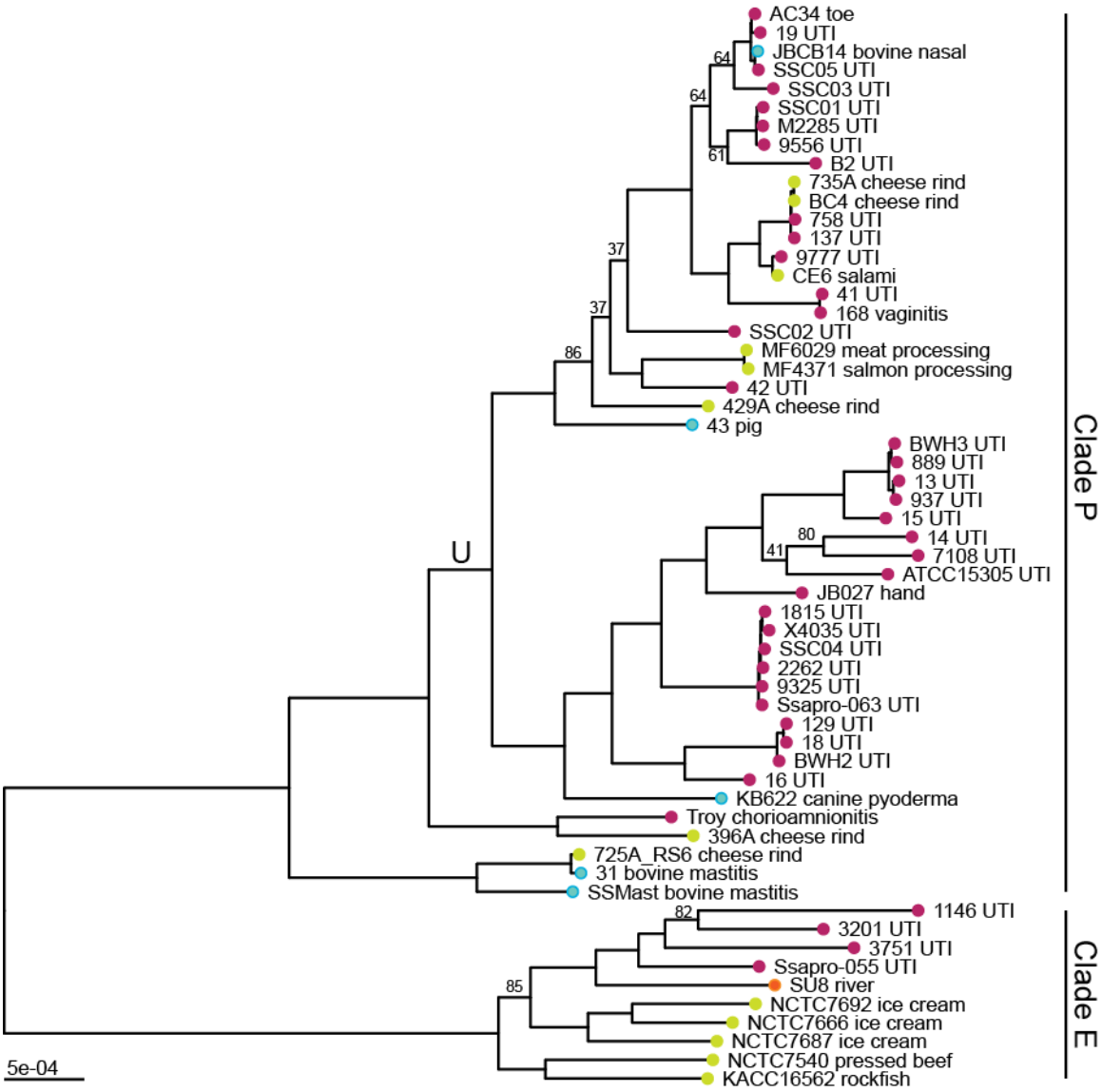



**Figure 1. Maximum likelihood phylogeny of *S. saprophyticus*.** Maximum likelihood phylogenetic analysis was performed in RAxML (*91*) using a whole genome alignment with repetitive regions masked. The phylogeny is midpoint rooted, and nodes with bootstrap values less than 90 are labelled. Branch lengths 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').

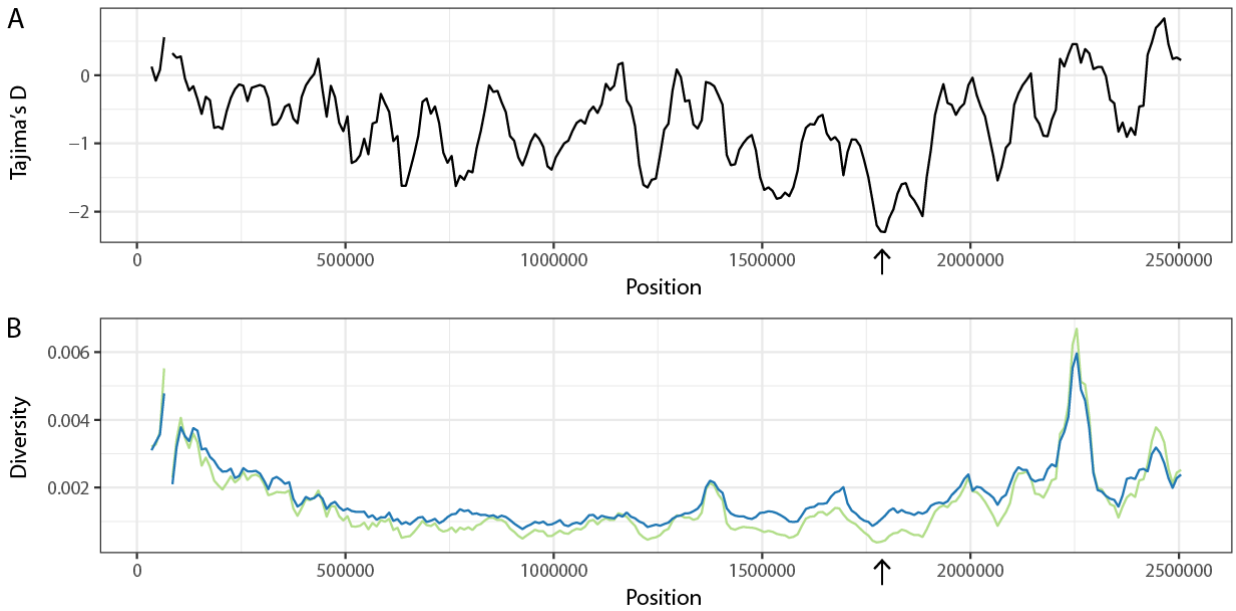



**Figure 2. Sliding window analysis of diversity and neutrality statistics.** Population genetic statistics were calculated for lineage U using EggLib (*93*). Windows were 50 kb in width with a step size of 10 kb. A) Tajima's D. B) π (green) and θ (blue). The lowest values for Tajima's D and π are found in the same window (1,760,000-1,820,000 bp, arrow).

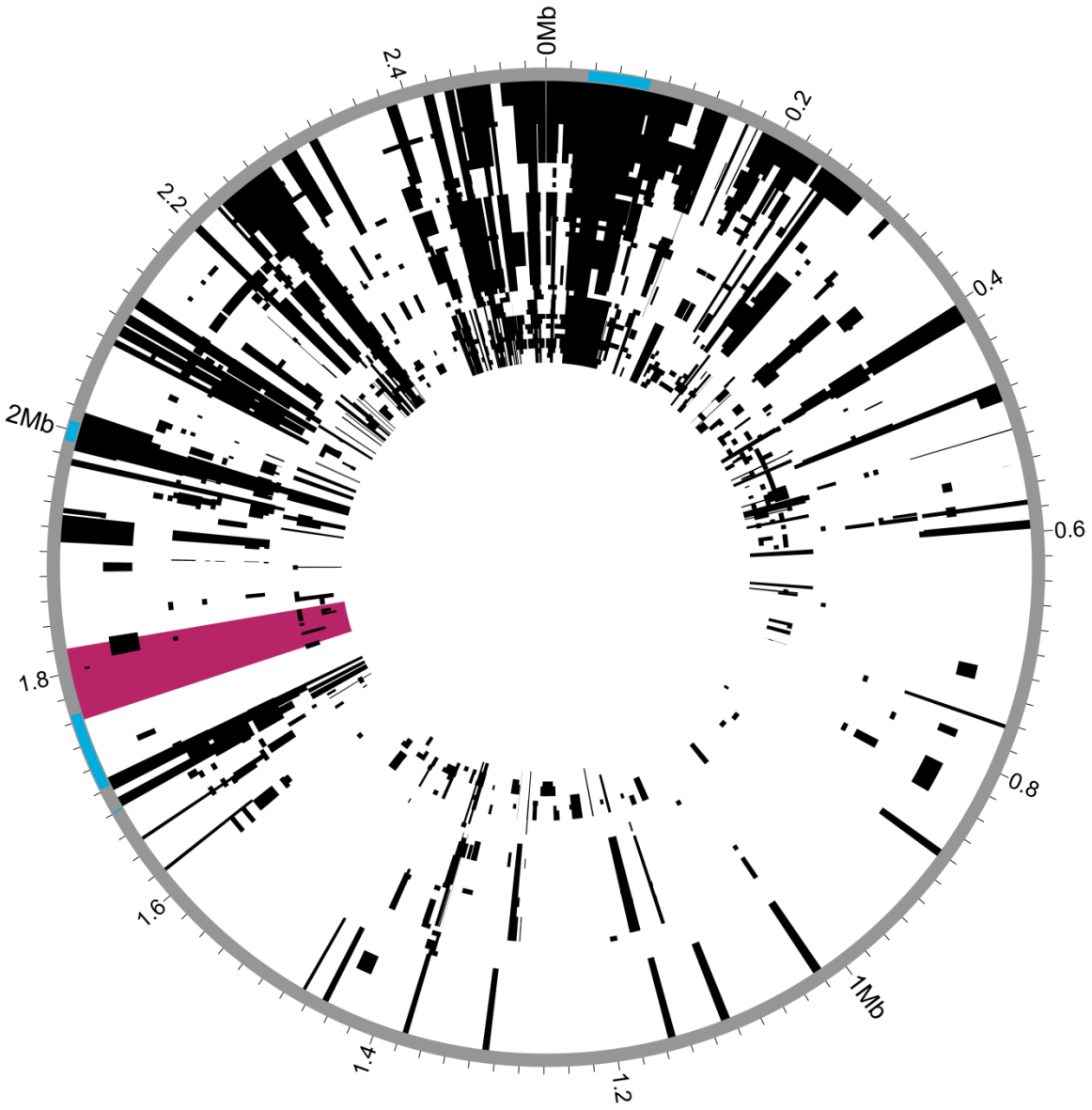



**Figure 3. Recombination in *S. saprophyticus*.** Recombinant regions in the whole genome alignment of *S. saprophyticus* were identified using Gubbins (*20*). Mobile genetic elements are highlighted in blue on the outer rim. The window with low Tajima's D and π is highlighted in pink. Few recombination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

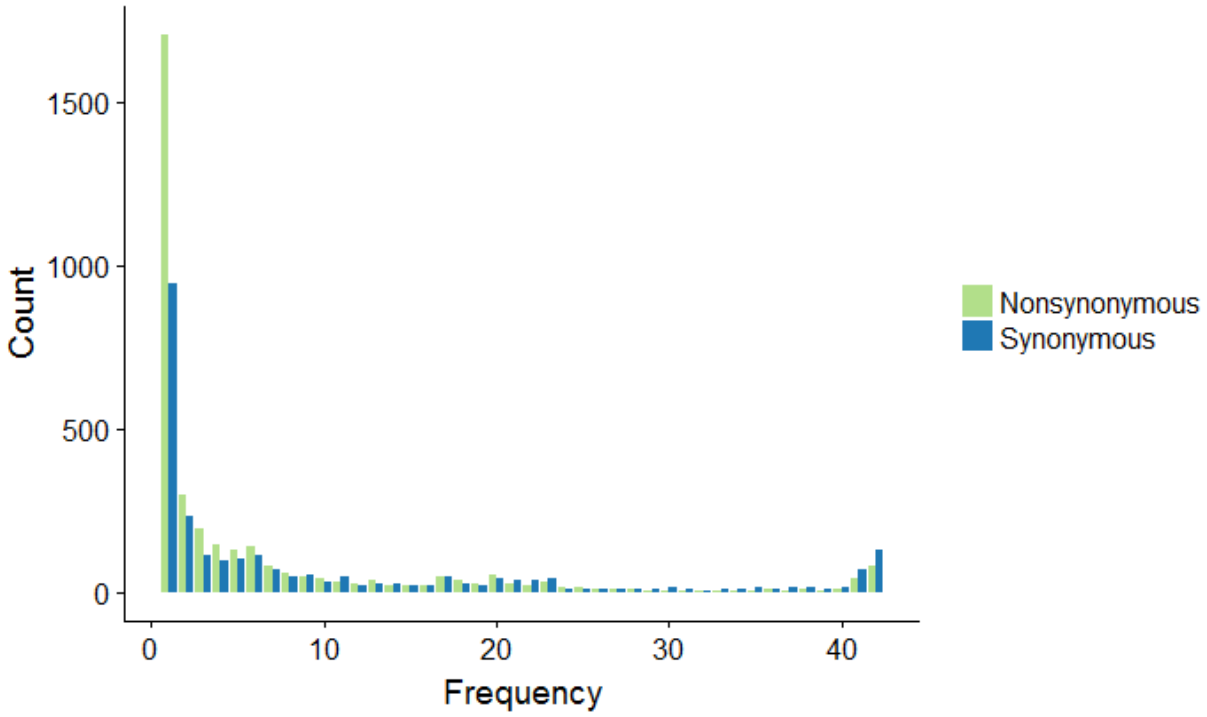



**Figure 4. Site frequency spectrum of lineage U.** The ancient genome (Troy) was used as the outgroup to determine the ancestral state. Synonymous, nonsynonymous, and intergenic sites were identified with SnpEff (*97*). All SFS contain an excess of singletons and high frequency derived variants. The synonymous SFS was used for demographic inference.

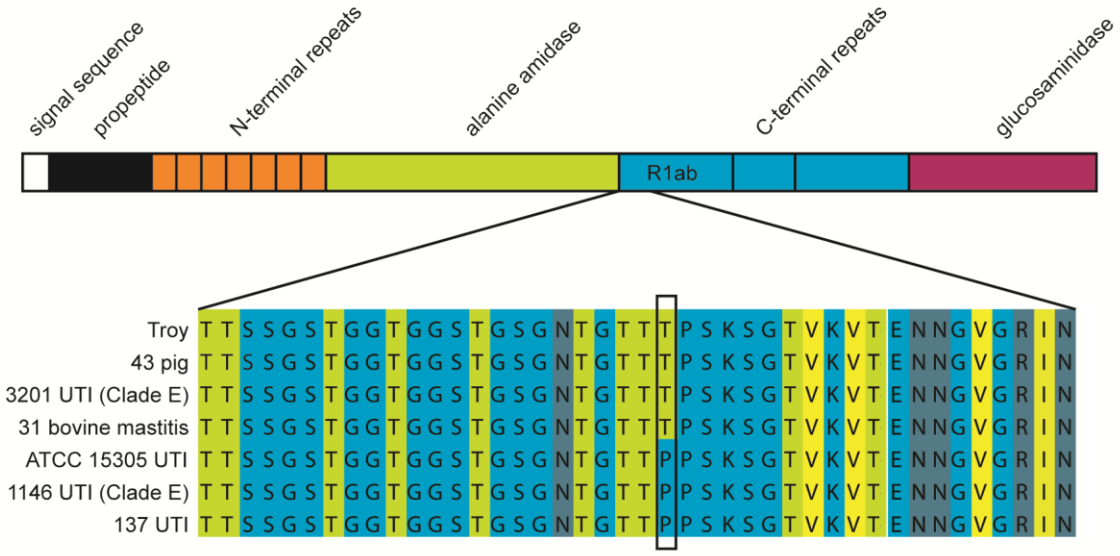



**Figure 5. Non-synonymous variant in Aas fibronectin binding repeat.** Top- Domains of Aas protein adapted from Hell et al. 1998. R1ab is the peptide used in the fibronectin and thrombospondin binding experiments. Bottom- Alignment of a portion of R1 showing amino acid sequence in Aas from selected *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.

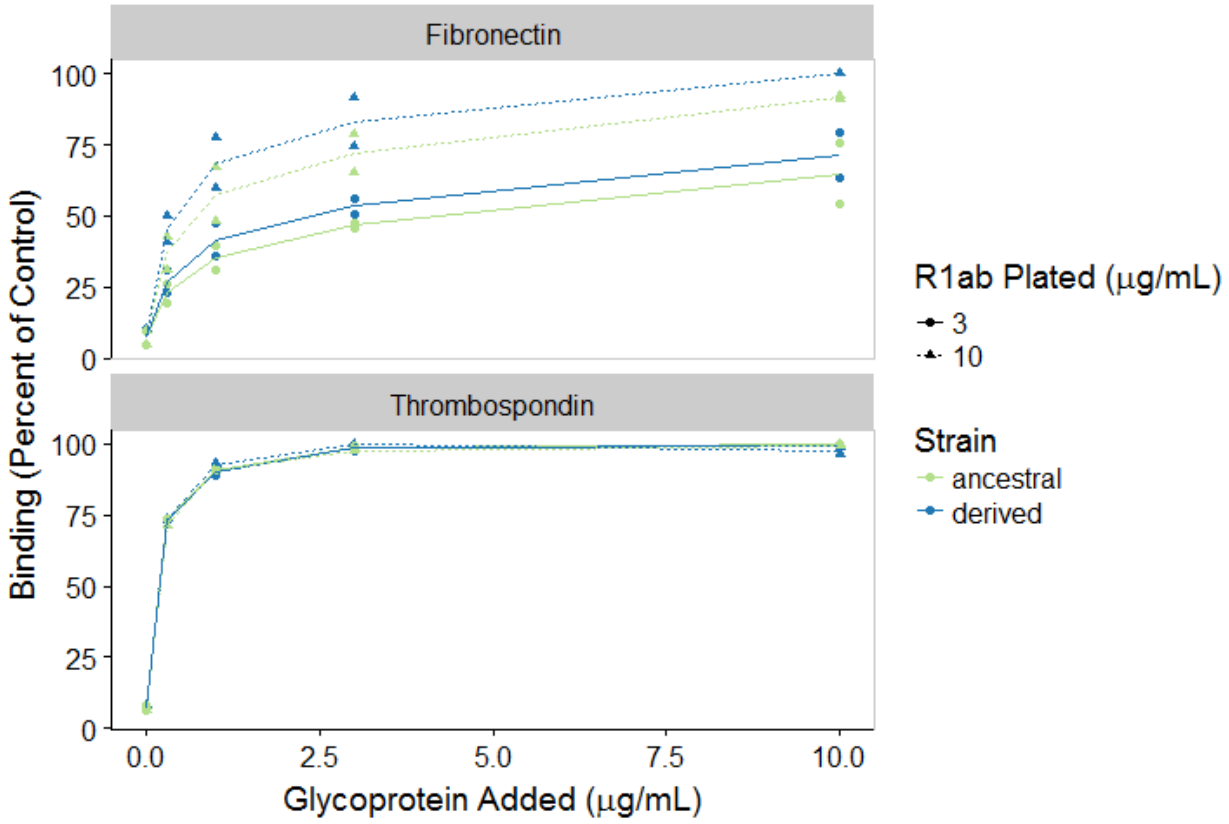
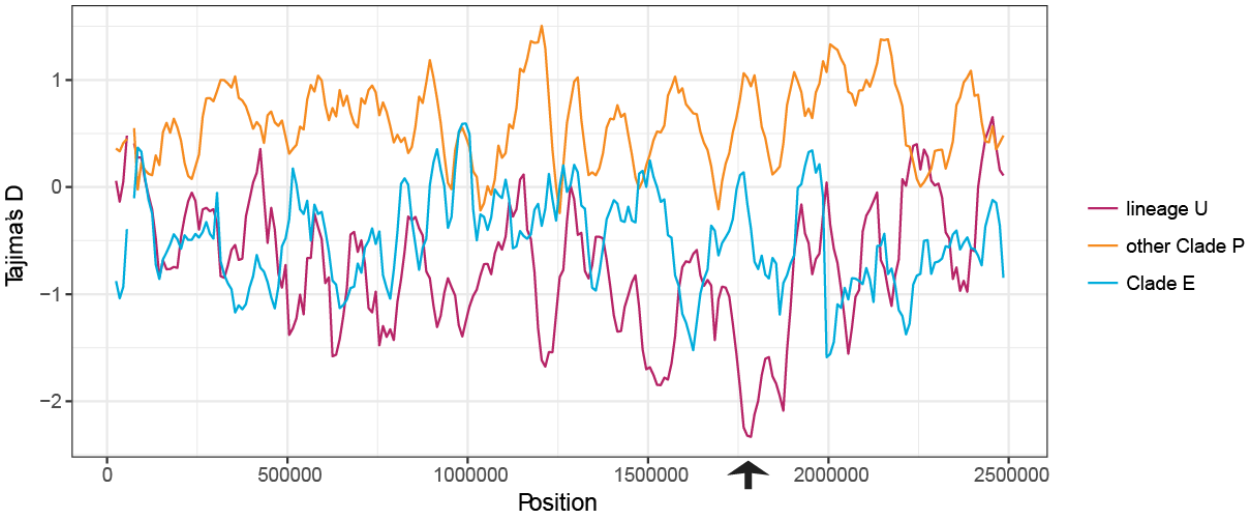


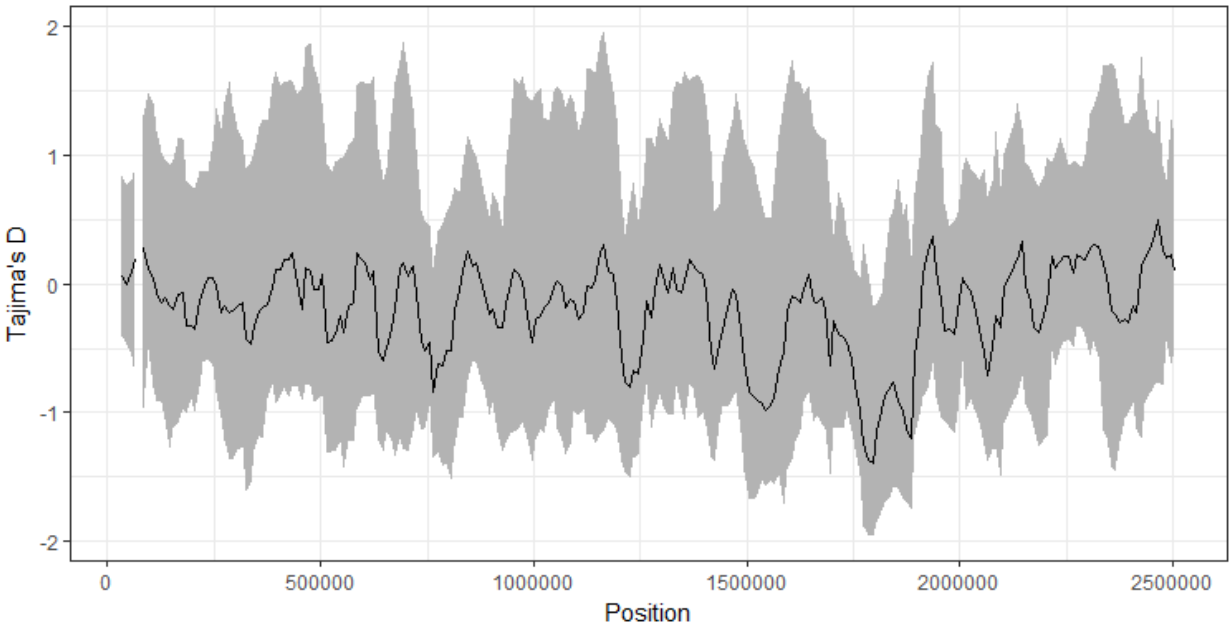

**Figure 6.**

**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. Fibronectin bound the human-associated strain better than the ancestral strain Aas R1ab across a range of concentrations. Thrombospondin bound to the two strains equally well. Symbols represent means of triplicate wells in separate experiments.

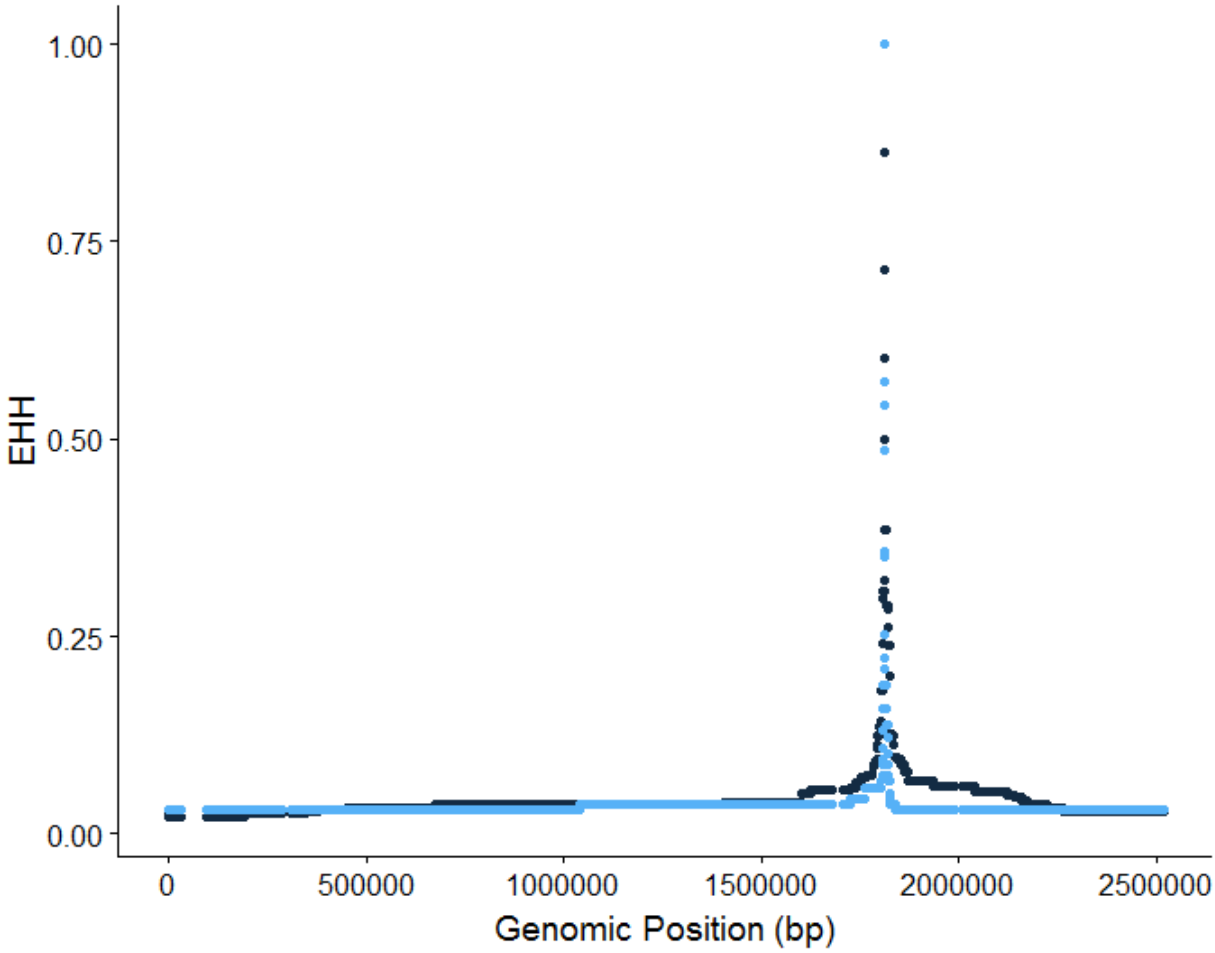
**Supplementary Figures:**




**Supplementary Figure 1.** Sliding window Tajima's D in *S. saprophyticus* populations. Windows were 50 kb in width with a step size of 10 kb. The local decrease in Tajima's D at 1,760,000-1,820,000 bp, arrow is specific to lineage U.




**Supplementary Figure 2.** Sliding window analysis of Tajima's D in subsamples (n=10) of lineage U. Windows were 50 kb in width with a step size of 10 kb. The local decrease in Tajima's D at 1,760,000-1,820,000 bp is not dependent of sampling of any particular *S. saprophyticus* isolates from lineage U.




**Supplementary Figure 3.** 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.

**Supplementary Tables:**

**Supplementary Table 1.** Accession numbers for *S. saprophyticus* isolates.

| Strain | Accession# | Source |
|---|---|---|
| ATCC 15305 | NC_007350.1 | Human urine |
| SU8 | NZ_JXBG00000000.1 | River |
| KACC 16562 | NZ_AHKB00000000.1 | Rockfish |
| NCTC 7692 | ftp://ftp.sanger.ac.uk/pub/project/pathogens/NCTC3000/NCTC7692.gff | Ice cream |
| NCTC 7666 | ftp://ftp.sanger.ac.uk/pub/project/pathogens/NCTC3000/NCTC7666.gff | Ice cream |
| NCTC 7687 | ftp://ftp.sanger.ac.uk/pub/project/pathogens/NCTC3000/NCTC7687.gff | Ice cream |
| NCTC 7540 | ftp://ftp.sanger.ac.uk/pub/project/pathogens/NCTC3000/NCTC7540.gff | Pressed Meat |
| 137 | NZ_CP014057.1 | Human urine |
| 168 | NZ_CP014113.1 | Human urine |
| 758 | NZ_JUUE00000000.1 | Human urine |
| 129 | SRR3057117 | Human urine |
| 13 | SRR4938476 | Human urine |
| 14 | SRR4938475 | Human urine |
| 15 | SRR4938469 | Human urine |
| 16 | SRR4938477 | Human urine |
| 18 | SRR4938478 | Human urine |
| 19 | SRR4938472 | Human urine |
| 41 | SRR4938467 | Human urine |
| 42 | SRR4938473 | Human urine |
| 55 | GCA_001937035 | Human urine |
| 63 | GCA_001937025 | Human urine |
| M | SRR4938468 | Human urine |
| X | SRR4938471 | Human urine |
| K | SRR4938470 | Canine dermatitis |
| 31 | SRR4938474 | Bovine mastitis |

| | | |
|---|---|---|
| JB027 | GCA_001747245 | Human hand |
| 725A_RS6 | GCA_001747665 | Cheese rind |
| BC4 | GCA_001747655 | Cheese rind |
| CE6 | GCA_001747625 | Salami |
| 429A | GCA_001747585 | Cheese rind |
| 396A | GCA_001747975 | Cheese rind |
| 3751 | GCA_001747545 | Human urine |
| 3201 | GCA_001747465 | Human urine |
| 1146 | GCA_001747575 | Human urine |
| 9777 | GCA_001747485 | Human urine |
| AC34 | GCA_001747475 | Human toe |
| JBCB14 | GCA_001747555 | Bovine nasal |
| 9556 | GCA_001747165 | Human urine |
| B2 | GCA_001747235 | Human urine |
| 2262 | GCA_001747325 | Human urine |
| 1815 | GCA_001747225 | Human urine |
| 9325 | GCA_001747315 | Human urine |
| BWH3 | GCA_001747155 | Human urine |
| 889 | GCA_001747145 | Human urine |
| 937 | GCA_001747305 | Human urine |
| 7108 | GCA_001747265 | Human urine |
| BWH2 | GCA_001747175 | Human urine |
| 735A | GCA_001747645 | Cheese rind |
| Troy | PRJNA352376 | Human chorioamnionitis (ancient sample) |

**Supplementary Table 2.** Single nucleotide polymorphisms with $F_{ST}$ values in the top 0.05% 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 | 0.50 | Non-synonymous |
| 1797190 | 0.90 | 0.37 | 0.48 | Synonymous |
| 1808274 | 0.80 | 0.21 | 0.52 | Synonymous |
| 1811585 | 0.72 | 0.16 | 0.50 | Synonymous |
| 1811777 | 0.90 | 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 |