



Genome specialization and decay of the stranglers pathogen, *Streptococcus equi*, is driven by persistent infection

Simon R Harris, Carl Robinson, Karen F Steward, et al.

Genome Res. published online July 9, 2015

Access the most recent version at doi:[10.1101/gr.189803.115](https://doi.org/10.1101/gr.189803.115)

P<P	Published online July 9, 2015 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Comprehensive immune receptor profiling.
Discover the **DriverMap™ AIR Assay** difference.



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

TITLE:

Genome specialization and decay of the strangles pathogen, *Streptococcus equi*, is driven by persistent infection

AUTHORS:

Simon R. Harris¹, Carl Robinson², Karen F. Steward², Katy S. Webb², Romain Paillot², Julian Parkhill¹, Matthew T.G. Holden^{1,3}, and Andrew S. Waller²

AFFILIATIONS:

¹ The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom.

² The Animal Health Trust, Lanwades Park, Kentford, Newmarket, CB8 7UU, United Kingdom.

³ School of Medicine, University of St Andrews, North Haugh, St Andrews, KY16 9TF, United Kingdom.

Correspondence should be addressed to A.S.W (andrew.waller@ahd.org.uk; +44 (0)1638 751000).

RUNNING TITLE:

Genome specialization of *Streptococcus equi*

KEYWORDS:

Genome decay, host-restriction, population replacement, convergent evolution, *Streptococcus equi*

ABSTRACT:

Strangles, the most frequently diagnosed infectious disease of horses worldwide, is caused by *Streptococcus equi*. Despite its prevalence, the global diversity and mechanisms underlying the evolution of *S. equi* as a host-restricted pathogen remain poorly understood. Here we define the global population structure of this important pathogen and reveal a population replacement in the late 19th or early 20th century. Our data reveal a dynamic genome that continues to mutate and decay, but also to amplify and acquire genes despite the organism having lost its natural competence and become host-restricted.

The lifestyle of *S. equi* within the horse is defined by short-term acute disease, strangles, followed by long-term infection. Population analysis reveals evidence of convergent evolution in isolates from post-acute disease samples, as a result of niche adaptation to persistent infection within a host. Mutations that lead to metabolic streamlining and the loss of virulence determinants are more frequently found in persistent isolates, suggesting that the pathogenic potential of *S. equi* reduces as a consequence of long term residency within the horse post-acute disease. An example of this is the deletion of the equibactin siderophore locus that is associated with iron acquisition, which occurs exclusively in persistent isolates, and renders *S. equi* significantly less able to cause acute disease in the natural host. We identify several loci that may similarly be required for the full virulence of *S. equi*, directing future research towards the development of new vaccines against this host-restricted pathogen.

INTRODUCTION:

Bacterial pathogens live a double life, balancing the requirements of acute disease with persistence and colonisation in order to optimise their ability to transmit to naïve animals. These high and low virulence states evolve over time to fit with each host species. However, the impact of evolution towards a persistent state on streamlining the genome of a virulent organism has not been described. Strangles is one of the oldest recognized infectious diseases of horses and continues to cause significant welfare and economic costs throughout the world. The causative agent, *Streptococcus equi* subspecies *equi* (*S. equi*), was first identified in 1888 (Schutz 1888). However, the clinical signs of strangles, typified by pyrexia, followed by abscessation of lymph nodes in the head and neck, were first reported by Jordanus Ruffus in 1251 (Rufus 1251) and detailed descriptions of the clinical course of disease are provided in old veterinary texts (Solleysel 1664; White 1825; Anon 1831; Blaine 1841; Spooner 1842; Miles 1868; Williams 1884). The historical clinical signs of strangles exactly mimic those observed in contemporary populations of horses and it is likely that the disease has been caused by the same species of bacteria throughout this 700-year period. Rupture of lymph node abscesses releases highly infectious pus that can spread the infection from one horse to another, but this direct transmission from acute disease cases in itself cannot explain the global success of *S. equi*. Rather, it is believed that incomplete drainage of abscess material from the retropharyngeal lymph nodes permits the organism to persistently infect the adjacent guttural pouches of convalescent horses, usually by residing within dried balls of pus known as chondroids (Newton et al. 1997). *S. equi* can persist in this low nutrient state in the absence of clinical signs for several years and potentially the remaining lifetime of the horse, providing the organism with prolonged opportunity to be shed into the environment and transmit to naïve animals (Newton et al. 2000). The disease cycle of acute disease followed by persistent infection has underpinned the success of *S. equi*, balancing the requirements of both acute and persistent states.

The worldwide population of *S. equi* is almost clonal by multilocus sequence typing (MLST) (Webb et al. 2008), consisting of only two sequence types (ST), ST179 and ST151, and has been proposed to have evolved recently from an ancestral strain of *Streptococcus equi* subspecies

zooepidemicus (*S. zooepidemicus*) (Webb et al. 2008). Both *S. equi* and *S. zooepidemicus* belong to the same group of pyogenic streptococci as the important human pathogen *Streptococcus pyogenes* (Group A *Streptococcus*) and share many common virulence mechanisms, with evidence of cross-species horizontal DNA exchange (Holden et al. 2009; Lefebure et al. 2012). Comparison of the complete genomes of *S. equi* strain 4047 (Se4047) and *S. zooepidemicus* strain H70 (SzH70) showed that the pathogenic specialization of *S. equi* has been shaped by a combination of gene loss due to nonsense mutations and deletions, and gene gain through the acquisition of mobile genetic elements (MGEs) (Holden et al. 2009). However, the effects of the disease cycle transitions between acute disease and persistent infection on the genome of *S. equi* remain unknown.

In this study we utilize whole genome sequencing of a large collection of *S. equi* strains from around the globe. The collection contains isolates from horses that were part of outbreaks on large farms in the UK, including isolates from acute and persistent states, and also multiple isolates from the same animal. Using these data we investigate the evolutionary history of the species, reveal the complexity of disease outbreaks and uncover evidence of common pathways of within-host adaptation associated with persistent infection, which together shed new light on the evolution of this important animal pathogen towards host-restriction.

RESULTS:

Population structure of a global collection of *S. equi*

The genomes of a collection of 224 *S. equi* isolates were sequenced (Supplementary Table 1). Forty geographically and temporally diverse isolates from outbreaks in Australia, Belgium, Canada, Ireland, New Zealand, Saudi Arabia, Sweden, and the USA were included to provide a global snapshot of the diversity of *S. equi*. Allied with this, 180 acute infection and persistent isolates from 41 counties across the UK were included to determine the diversity of *S. equi* within individual outbreaks. Finally, isolates from commercially available live attenuated vaccines, Equilis StrepE (Jacobs et al. 2000) and Pinnacle IN (high and low capsule) (Walker and Timoney 2002), and the

Se1866 isolate, which is the basis of the new multi-component Strangvac vaccine (Guss et al. 2009), were included to determine the relationships of these vaccines to the currently circulating population of *S. equi*, and to identify mutations responsible for attenuation of the Pinnacle IN strain.

The indexed sequencing reads from each isolate were mapped against the Se4047 reference genome (Holden et al. 2009) to identify single-nucleotide polymorphisms (SNPs), insertions and deletions (indels). 3,109 polymorphic sites were identified across the genome. 58.8% of these sites ($n = 1,844$) were in accessory regions of the Se4047 reference genome, which corresponds to only 16.4% of the length of the genome. Much of this variation in accessory regions is likely to have arisen by recombination and horizontal gene transfer of alternative MGEs. To remove confounding signals in these regions, all phylogenetic analyses were carried out using the 1,265 polymorphic sites (Supplementary Table 2) in the 1.9 Mb of the core genome.

A Maximum likelihood (ML) phylogenetic reconstruction of the core variable sites of the sequenced isolates is shown in Fig. 1a. Isolates are colored according to clusters defined using a tree-independent Bayesian method for subdividing populations based on sequence similarity (Corander et al. 2008). Clusters one to three correspond to the three major clades in the ML tree, with the fourth cluster representing the remaining isolates. The increasingly prevalent ST151 formed a clade within cluster 1, distant from the Equilis StrepE and Pinnacle IN vaccine samples, which fell into clusters 3 and 4, respectively (Fig. 1a). Two features of the tree were particularly striking. First, the observed sequence diversity was surprisingly low for a global collection of isolates of a pathogen with a historical record that suggests it dates back to at least the 13th century. Second, the tree includes a number of anomalously long branches towards its tips, which appeared inconsistent with the isolation dates of the samples.

Variation in the rate of substitution in the *S. equi* core genome

Plotting root-to-tip distance against isolation date for each isolate for which dating information was available revealed a poor correlation between accumulation of substitutions and time (Fig. 1b) ($R^2 = 0.00868$), although a tip-date permutation test (Firth et al. 2010) confirmed significant temporal

signature in the data ($p = 0.001$ from 1,000 permutations). Similar observations in other streptococci have been the result of the effect of homologous recombination, which can import large amounts of variation into the genome *en masse*, leading to elongation of branches and the loss of clock-like mutational signals. Although it has been thought that *S. equi* has lost natural competence (Holden et al. 2009), we investigated the possibility that recombination was the cause of the observed long branches. By reconstructing sequences for ancestral nodes in the tree, evidence for past recombination events can be identified in two ways: recombination events from donors outside of the diversity of the tree will cause imports of large numbers of clustered substitutions on individual branches of the tree, while recombination events between donors and recipients within the diversity of the tree will lead to accumulation of homoplasies. In contrast to other streptococci (Croucher et al. 2011; Marttinen et al. 2011; Cornejo et al. 2013), we observed low levels of both SNP clustering (Supplementary Fig. 1) and homoplasy (only 47 or 3.6% of core variable sites), ruling out recombination as the cause of the long branches. The highest density of SNPs (3.2% of SNPs, $p < 0.0001$) and homoplasies (28% of homoplastic SNPs, $p < 0.0001$) in the core genome coincided with the 5' 1.2 kb region (0.06% of the core genome length) of the SeM gene that encodes a virulence factor that binds to fibrinogen and immunoglobulin. The promoter region of the same gene also contained three homoplasies. Similarly, genes encoding the fibronectin- or fibrinogen-binding proteins, FneE and SzPSe also exhibit high SNP density and homoplasy, highlighting the importance to *S. equi* of generating variation in these proteins.

Given the observed lack of homologous recombination, the poor root-to-tip correlation must have resulted from variation in substitution rate across the tree. Plotting the residuals for the root-to-tip analysis (Fig. 1c) showed that seven isolates, subtended by four long branches when mapped onto the tree (Fig. 1a), fell outside of two standard deviations of the mean. These included two isolates of the Pinnacle IN vaccine strain; two isolates extracted from the right guttural pouch of a horse (*JKS121*; italicized text indicates the ID of the horse, standard text indicates the *S. equi* isolate from this animal, in this case isolates, JKS121a and JKS121b) sampled during a strangles outbreak in Leicestershire, two isolates from the guttural pouches of a horse (*JKS628*) sampled during an outbreak in Essex, and one isolate from the right guttural pouch of a horse (*JKS731*)

sampled during an outbreak in Lincolnshire. Excluding these seven isolates increased the root-to-tip R^2 value to 0.11372 (Fig. 1d). The long branch associated with the Pinnacle IN isolates resulted from 68 shared unique SNPs in the core genome of these isolates. This diversity could be explained by the methods used to make the vaccine, which was a live attenuated vaccine generated by chemical mutagenesis of *S. equi* with *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine (NTG) (Timoney 1985). As expected for NTG-treatment, the substitution spectra for the SNPs unique to the Pinnacle isolates was significantly enriched for C->T and G->A, but deficient in A->G and T->C transitions (Harper and Lee 2012) when compared to other branches in the tree (Supplementary Fig. 2). In wild-type isolates, disruption of the mismatch repair system can also lead to increased substitution rates, but is characterized by an increase in the accumulation of A->G and T->C transitions. We found no significant differences between the substitution spectra of the three long branches leading to clinical isolates, including two isolates from JKS121 which exhibited a nonsynonymous mutation in the mismatch repair gene, *mutS* (Supplementary Fig. 2). Consistent with this, no significant differences ($p = 0.093$) in resistance mutation frequency were found when comparing the long-branch isolates from *JKS121*, *JKS628*, *JKS731*, the Pinnacle IN vaccine isolates and the reference *Se4047 in vitro* (Supplementary Fig. 3).

The two isolates grown from the Pinnacle IN vaccine differed phenotypically in their amount of hyaluronic acid capsule (Walker and Timoney 2002), hemolysis of red blood cells and by three SNPs, including a nonsynonymous substitution (Ile206Thr) in a homologue of the *csrS* two-component system sensor that has been demonstrated to regulate virulence factors, including capsule, in *S. pyogenes* (Graham et al. 2002). Both isolates shared a second SNP in *csrS*, which is predicted to result in a Leu257Phe amino acid substitution in addition to one nonsense mutation and a further 40 nonsynonymous mutations elsewhere in the genome (Supplementary Table 2). The combination of these SNPs likely explain the attenuation of Pinnacle, and could now form the basis of a diagnostic test to resolve cases when the vaccine strain has been suggested to have caused adverse reactions or reverted to virulence (Borst et al. 2011; Patty and Cursons 2014).

Sequencing resolved the complex epidemiology of strangles outbreaks

JKS121, *JKS628* and *JKS731* were all sampled during strangles outbreaks in the UK. The inclusion of multiple isolates from outbreaks allows us to utilize the resolution that whole genome sequencing provides to conduct detailed epidemiological analysis of strangles. In most cases, outbreak isolates were highly clonal, and differed by only a small number of SNPs (Supplementary Fig. 4), consistent with an import from a single source. However, in some cases, including the outbreaks involving *JKS121*, *JKS628* and *JKS731*, both active strangles strains, and persistent strains from chondroids were identified from horses at the same stables during a strangles outbreak (Supplementary Fig. 4). In one case, 10 isolates recovered from a small outbreak in Essex over a 5 month period all fell into cluster 3 on our phylogeny. However, within this cluster the isolates grouped into three distinct sub-clades differentiating an outbreak strain and two persistent strains (Fig. 2). Interestingly, of the four isolates recovered from *JKS628*, including the two outliers in our root-to-tip analysis, two fell into each of the persistent clades possibly reflecting separate, long-standing persistent infections prior to the purchase of this horse some 15 months before the outbreak. A similar situation was observed for six isolates from an outbreak in Leicestershire in 2007. Four isolates, from both persistent infection and disease, formed a clade in cluster 3, representing the outbreak strain. However, the two root-to-tip outliers from *JKS121* were distinct, falling in cluster 1. These two persistent isolates possibly originated from a previous outbreak during the 8-year residency of this horse on the affected premises. Finally, *JKS731* was sampled during a large outbreak in Lincolnshire that persisted from 2006 to 2008 and affected over 200 horses. Twenty-seven isolates from this outbreak formed a single sub-clade within cluster 2 on our tree. Two of these isolates, including the persistent isolate from *JKS731*, which was the final outlier in our root-to-tip analysis, shared a much deeper common ancestor than the rest of the outbreak isolates, which suggests these may be persistent isolates from a historic outbreak of a similar genotype, possibly occurring during the 15 years residency of this horse at the affected premises. The outbreak strain may have been the result of reinfection from an external source, or more likely from a persistent infection at the stables. As with the Essex and Leicestershire outbreaks, isolates taken during the Lincolnshire outbreak also included two persistent lineages that clustered far away from the main outbreak clade, illustrating the prevalence of long-term persistence of *S. equi*.

Substitution rates vary between acute infection and persistence

Remarkably, despite the five clinical isolates identified as outliers by our root-to-tip analysis being sampled during outbreaks, all were isolated from chondroids extracted from the guttural pouches of healthy horses, and therefore were representative of persistent populations that originated from historic acute infections. This observation raised the possibility that the poor temporal signal identified in our root-to-tip analysis was the result of *S. equi* displaying altered substitution rates during persistent infection. To shed further light on this, we analyzed the subset of our dataset for which accurate isolation dates were available with the Bayesian phylogenetics software, BEAST (Drummond and Rambaut 2007), which allows modeling of molecular clock rates to provide estimations of divergence dates on nodes of a phylogenetic tree. Removal of artificially attenuated vaccine isolates and isolates without dating information produced a dataset of 209 isolates with 1184 polymorphic core genome positions. BEAST includes a number of relaxed molecular clock models that permit modeling of variation in substitution rates on different branches of the tree, allowing us to correct for the observed rate variation in our data, and also to identify other branches exhibiting particularly high substitution rates. Indeed, the combination of a skyline population model and relaxed-exponential clock model was found to be the best fit to our data based on Bayes Factors using the harmonic mean estimator. The topology of the maximum clade credibility (MCC) tree generated from combined data post burn-in from three independent runs of this model was concordant with the ML tree of all isolates (Supplementary Fig. 5). The mean substitution rate per core genome site per year was calculated as 5.22×10^{-7} (95 % highest posterior density (HPD): 4.04×10^{-7} to 6.51×10^{-7}), slower than the core genome rates reported for other streptococci, including *S. pyogenes* (1.1×10^{-6}) (Davies et al. 2015) and *Streptococcus pneumoniae* (1.57×10^{-6}) (Croucher et al. 2011), and many other Gram-positive bacteria, including *Staphylococcus aureus* (3.3×10^{-6}) (Harris et al. 2010; Holden et al. 2013). Interestingly, the mutation rate of *S. equi* is closer to those calculated for *Clostridium difficile* (1.47×10^{-7} to 5.33×10^{-7}) (He et al. 2013) and *Mycobacterium tuberculosis* (8.0×10^{-8} to 1.25×10^{-7}) (Bryant et al. 2013), Gram-positive bacteria which can undergo periods of dormancy and long-term persistence. The analysis provided a median estimate for the time of the most recent common ancestor

(tMRCA) of our global sample of *S. equi* to 1909 (95 % HPD: 1819 to 1946). Given the historical record of strangles suggests that *S. equi* dates back to at least the 13th century, our data indicate that a global population replacement occurred during the 19th or early 20th centuries, corresponding to a time when horses were a major mode of transport and played important roles in a number of global conflicts (Fig. 3). To ensure this result was not an artifact of the clock or population models employed in our analysis, we calculated the mean, median and 95 % HPD tMRCA for a range of model combinations and found that our chosen model exhibited the widest HPD, which encompassed the 95 % HPDs of all other model combinations (Fig. 3).

To test our hypothesis that persistent isolates exhibited increased substitution rates, we performed a two-sided Mann-Whitney *U* test of median substitution rates for branches in our BEAST tree leading to persistent isolates against those leading to isolates from acute infection (Supplementary Fig. 6a). We found a significant difference ($p = 0.0004$) in rates between the two populations, with persistent isolates exhibiting a significantly increased rate. A linear regression of substitution rate against time for persistent, acute and other (where the site was unknown) branches showed an increasing rate over time for persistent isolates ($R^2 = 0.20156$, $p < 0.0001$), while the rates for acute ($R^2 = 0.01459$, $p = 0.3693$) and other ($R^2 = 0.00323$, $p = 0.3815$) branches showed no increase with time (Supplementary Fig. 6b). A similar disparity was identified in the types of substitutions occurring in acute and persistent infection, with more nonsynonymous mutations accruing over time in persistent infection when compared with isolates from acute infection and others (Supplementary Fig. 7a). In fact, whereas acute infection and other branches showed a relatively constant dN/dS ratio of around 0.5 to 0.7 over time, persistent isolates showed an increase in dN/dS with time (Supplementary Fig. 7b), supporting the hypothesis that these isolates are undergoing diversifying selection.

Variation in the accessory genome was not associated with persistence or infection

Having observed distinct differences in the mutation rates of core genomes of the persistent and acute populations, we undertook a comparative genomic analysis to investigate if there were also

differences in the dynamics of the accessory genome. Amongst the mobile elements described in the reference Se4047 isolate, the prophages ϕ Seq2-4, and the integrative conjugative elements ICESe1 and ICESe2 were the most stable, being almost ubiquitously conserved in the population (Supplementary Fig. 8). In contrast, the prophage ϕ Seq1 appeared more dynamic, with evidence of alternative, but related, prophages in the accessory genomes of isolates belonging to all four clusters. Although the pattern of presence of these prophage elements generally followed the phylogeny, there was evidence of lateral transfer or convergent gain of some prophages in multiple clusters. The deletion of an ancestral CRISPR locus, prior to the emergence of *S. equi* from *S. zooepidemicus*, potentially contributes to the poly-lysogeny of this species (Holden et al. 2009; Waller and Robinson 2013).

Although ϕ Seq2-4, ICESe1 and ICESe2 were generally conserved (Supplementary Fig. 8), we found evidence of both deletions and duplications affecting regions of these elements. Several isolates showed duplications of the phospholipase A₂ gene, *slaA*, on ϕ Seq2, suggesting that these strains had acquired additional distinct prophages that also contain this putative virulence cargo (Holden et al. 2009). Convergent loss of the *seeL* and *seeM* superantigen-encoding genes in six isolates from three distinct lineages appears to have occurred twice by loss or replacement by diverse alternative prophages of ϕ Seq3, as evidenced by only <25 % of ϕ Seq3 having matches in the genome sequences of the isolates (Supplementary Fig. 8), and once by horizontal replacement of ϕ Seq3 with an alternative, similar prophage, as evidenced by 60-70% of ϕ Seq3 having matches in the genome sequence of the isolates (Supplementary Fig. 8), and an elevated SNP density (mean of 1 SNP every 168 matched bases vs 1 SNP every 6130 matched bases for other isolates with at least one SNP).

Insertion sequence elements mediate genome decay and gene amplification during persistence

Whilst our comparative analysis revealed that, as would be expected, prophages were mediating the loss and gain of genes, it also revealed that genes were being lost and duplicated in the core genome. The mechanism that appears to be driving the observed deletion and copy-number

variation is homologous recombination between insertion sequence (IS) elements. The *S. equi* genome has large expansions of a limited number of IS families, which is hypothesized to result from the evolutionary bottleneck associated with its host-adaptation and pathogenic specialization (Holden et al. 2009). We identified an additional 37 novel IS element insertions belonging to existing families across the population, eight of which inserted at the same genomic location in multiple lineages (Supplementary Fig. 9 and 10). IS elements flanked four of the 15 deleted loci and 10 of the 16 amplified regions in the core genome, showing that they play an important role in variation of gene content.

Consistent with our hypothesis that the *S. equi* genome diversifies within the guttural pouch environment, large deletion and duplication events were far more common in persistent isolates than those from acute infection (Supplementary Fig. 9). Within the core genome, the *has* operon, encoding hyaluronic acid capsule biosynthesis components, was the locus most frequently affected by deletions and duplications, with seven independent deletion variants and eight independent amplifications of the locus, which is bordered by IS3 elements (Fig. 4). The same region also exhibited high SNP densities and four independent nonsense mutations. Importantly, the SNPs, deletions and amplifications at this locus were significantly associated with isolates recovered from persistently infected horses rather than those with acute disease ($p < 0.0001$). To determine the effects of these changes on the production of the hyaluronic acid capsule, we quantified the transcription across the *has* locus and the amount of hyaluronic acid present on the surface of five representative strains. The genome of strain Se4047 has been sequenced to completion and was included as a reference (Holden et al. 2009). Strain JKS551, recovered from a persistently infected horse, contains a deletion of *hasA* and the 3' region of *hasB* (Fig. 4) relative to strain JKS540a, which was recovered from a horse suffering from acute disease during the same outbreak of strangles as JKS551. Reverse transcription qPCR confirmed a lack of *hasA* transcription in strain JKS551 ($p = 0.0017$) and this strain had significantly less hyaluronic acid associated with its cell surface ($p = 0.04$) (Supplementary Fig. 11). Strain 851, which contains an amplification of the *has* locus, had significantly higher levels of transcription of the *has* genes ($p = 0.005$, $p = 0.0001$ and $p = 0.001$ for *hasA*, *hasB* and *hasC*, respectively) and a corresponding

increase in the amount of hyaluronic acid on its surface ($p = 0.0005$) relative to strain 1165, which was recovered from the same persistently infected animal two weeks later and contains a wild-type *has* locus (Supplementary Fig. 11). A specific PCR that spanned the IS3 element that borders the *has* locus was used to confirm that the amplification of the *has* locus in strain 851 occurs in tandem.

Comparison of the multiple isolates recovered from persistently infected horses, showed that even these extremely closely related isolates varied by deletions and/or amplifications (Supplementary Fig. 12), demonstrating that microevolution of *S. equi* in the guttural pouch yields a mixture of variant strains. One example is the horse JKS628, within which two of the four persistent isolates, JKS628c and JKS628d, recovered from the right and left guttural pouches respectively, contained a 39.5 kb deletion (SEQ_1232 to SEQ_1253) in ICESe2 that included the entire equibactin locus. This locus encodes a putative siderophore, which is known to increase the ability of *S. equi* to acquire iron (Heather et al. 2008), and may enhance survival within, and abscessation of, lymph nodes (Holden et al. 2009). This siderophore has been proposed to represent the key speciation event in the evolution of *S. equi* (Holden et al. 2009). Three other deletion events (Fig. 5) that are predicted to disrupt equibactin production were identified in isolates recovered from three more, unrelated persistently infected horses. In contrast, the equibactin region was not deleted in any of the acute isolates ($p = 0.0528$), suggesting that loss of equibactin may preclude transmission to acute infection.

To determine if the loss of the equibactin locus represented a dead end mutation that would prohibit these isolates from being able to cause acute disease and therefore continue to be transmitted via the strangles disease cycle, we challenged two groups of seven Welsh mountain ponies with 1×10^8 cfu of Se4047 or an *eqbE* deletion mutant, which is unable to produce equibactin (Heather et al. 2008). Ponies were monitored carefully for early signs of disease, such as pyrexia and a preference for hay over pelleted food, and euthanized before the onset of later clinical signs of strangles (Guss et al. 2009). Deletion of *eqbE* significantly reduced the amount of pyrexia ($p = 0.021$) and pathology at post mortem examination ($p = 0.041$) (Fig. 6). However, loss

of *eqbE* did not completely prevent abscess formation with two $\Delta eqbE$ -infected ponies developing bilateral retropharyngeal lymph node abscesses compared with all seven wild-type infected animals ($p = 0.021$). One $\Delta eqbE$ -challenged pony developed a submandibular lymph node abscess, two ponies had micro-abscesses without clinical signs of disease and two had no signs of infection. The attenuation of the $\Delta eqbE$ deletion mutant in Welsh mountain ponies demonstrates that the acquisition of ICESe2 was indeed a key step in the evolution of *S. equi* and that strains containing deletions in the equibactin locus, which are shed from persistently infected horses, have a lower virulence potential.

DISCUSSION:

Analysis of a global collection of *S. equi* isolates, has, for the first time, allowed important questions regarding the complex epidemiology of strangles outbreaks to be resolved, and provided compelling evidence of co-infection of individual horses. Variation between isolates in the collection was strikingly low given that the historical record of strangles suggests *S. equi* dates back over 700 years. Our data indicate that a global population replacement occurred during the 19th or early 20th centuries, corresponding to a period of history during which horses were, for the most part, the major mode of transportation and performed critical roles in a series of global conflicts (Clabby 1976). Strangles was, and remains, a particular problem in army horses (Todd 1910; Bazeley 1940). Indeed, the Leicestershire outbreak represented in this study that included horse *JKS121* occurred at the Defense Animal Centre in Leicestershire. Large global conflicts, such as World War I, led to the mobilization and mixing of large populations of horses from around the world in conditions ideally suited to the transmission and mixing of *S. equi* clones. In conjunction with this, dramatic levels of host mortality would have increased pathogen population bottlenecks and the potential for loss of population diversity. Later replacement of lost horses, through initiatives such as the establishment of the National Stud in the UK in 1916 for the breeding of thoroughbreds, produced naïve young horses on an unprecedented scale, providing new hosts for successful *S. equi* clones to colonize. Thus, we speculate that the unique

circumstances created by 19th and early 20th century global conflicts would have provided ideal conditions for the emergence of the contemporary population of *S. equi*.

The crucial factor allowing *S. equi* to propagate the disease cycle of strangles is its ability to persist for long periods in a subclinical, but infectious state. The importance of the establishment of persistent infections to the capability of *S. equi* to transmit to naïve animals was first recognized by Captain Todd of the British Army Veterinary Corps as early as 1910 (Todd 1910) and is undoubtedly a major factor in its continued global success. However, it was not until 1997 that the anatomical site of such persistent infections was shown to be the guttural pouch (Newton et al. 1997). Our data suggest that the environment within the guttural pouch, and the chondroid in which *S. equi* persists, drives both the diversification and decay of its genome, and potentially explains the path to host restriction. Although adaptation to the guttural pouch environment, such as loss of the equibactin siderophore, may reduce virulence and transmissibility, leading to an evolutionary dead-end, it is clear that strangles outbreaks can be founded by long-term persistent strains. Indeed, persistence may also bring beneficial effects beyond simply providing time for contact with new naïve hosts. For example, mutation of the antigenic, sortase-processed surface proteins SeM, SzPSe and SEQ_0402 during persistence could assist the evasion of acquired immunity facilitating persistence in the guttural pouch and potentially the recrudescence of strangles. Increased shedding and/or more efficient transmission of new variant strains in this manner provides one explanation for the raised SNP densities observed in the genes encoding these proteins and highlights further genes that exhibit similar raised SNP densities, such as *fneE*, which will direct the design of improved vaccines against strangles. An interesting observation from the whole genome phylogeny is that the increasingly prevalent ST151 clone of *S. equi* is found in a distinct clade to both the Equilis StrepE and Pinnacle IN vaccine isolates. The first identified isolates of ST151 were recovered from three horses with strangles less than three months after they were vaccinated with Equilis StrepE (Kelly et al. 2006), raising the prospect that this lineage may escape current vaccines and should be closely monitored.

The loss and inactivation of genes required for hyaluronic acid capsule biosynthesis exclusively in isolates of *S. equi* recovered from persistently infected horses is paralleled by the loss of these genes in human disease isolates of *S. pyogenes* serotype M4 and M22 (Flores et al. 2012) and the presence of SNPs in the *has* locus of isolates recovered from cases of pharyngitis rather than invasive disease caused by *S. pyogenes* serotype M3 (Shea et al. 2011). Our data highlight similarities between these two streptococcal species and provide further evidence that *S. equi* infection can serve as a model for human disease. The tandem amplification of the *has* locus led to increased transcription of *hasA*, *hasB* and *hasC* and significantly more hyaluronic acid on the surface of strain 851 relative to strain 1165, which was recovered from the same horse twelve days later. Theoretically the recombination of tandemly arranged amplified loci permits the restoration of a wild-type phenotype. This scenario could explain why later isolates from this animal (2424 and 3446) also contained a wild-type *has* locus. However, we cannot rule out the possibility that the population of *S. equi* from this horse contained more than one *has* variant on each occasion. We propose that the amplification of the *has* locus may assist *S. equi* to establish a persistent infection within the guttural pouch in the immediate period post-recovery when the equine immune response may be at its most effective. However, we speculate that the restoration of a wild-type phenotype or the loss of capsule production may be favored over time during persistent infection, possibly as chondroids form, nutrient levels decline or the immune response becomes gradually less effective. We speculate that amplifications of loci, including *has*, may be present in isolates of *S. pyogenes* recovered from cases of pharyngitis and that this strategy provides a novel bet-hedging solution to survival in fluctuating environments.

The lifecycles of many pathogens involve both acute and persistent phases of infection. Balancing the conflicting genetic requirements of transmission and clinical disease with survival in the face of a convalescent immune response represents a significant biological challenge that may ultimately lead towards host-restriction. By sequencing multiple isolates from the same infected horses our analyses have, for the first time, shed light on the complex population dynamics of *S. equi* during persistence within the host. The observation of an increased substitution rate within persistent infection mirrors the situation in human immunodeficiency virus (HIV) infection where mutations

accrue faster within hosts than at the epidemic level (Herbeck et al. 2006; Lemey et al. 2006). Mathematical modeling of this phenomenon supports the hypothesis that ancestral HIV sequences are preferentially transmitted, purifying out mutations generated within the host and reducing the apparent substitution rate at transmission (Lythgoe and Fraser 2012). Our data suggest a similar process has shaped the evolution of *S. equi*, and may be important in all pathogens the lifecycle of which includes long periods of infection between transmission events. For example, clear parallels can be drawn with the discrepancy between the slow observed substitution rate in *M. tuberculosis* transmission chains (Bryant et al. 2013; Walker et al. 2013) and its ability to quickly develop antibiotic resistance mutations within a host (Sun et al. 2012; Casali et al. 2014). For *S. equi*, the ability to adapt and persist in healthy animals for long periods while remaining infectious is a crucial component of the disease cycle of this pathogen. The fluidity of its genome plays a key role in ensuring *S. equi* remains a persistent threat to equine health with the mobility of modern horse populations facilitating its transmission around the world.

METHODS:

Study collection

Se4047 was isolated from a horse with strangles in the New Forest, Hampshire, UK in 1990. The origins and details of the 224 isolates of *S. equi* that were sequenced in this study are listed in Supplementary Table 1. Isolates from acute cases of strangles were recovered from clinical samples submitted to the Animal Health Trust's Diagnostic Laboratory Service citing clinical signs that were indicative of acute disease or were recovered during veterinary examination of acute clinical cases in the UK, Sweden, Belgium or Saudi Arabia. Persistent isolates were recovered from the guttural pouches of horses exhibiting no clinical signs of infection, from chondroids recovered from a guttural pouch, or from swabs where the attending veterinarian stated that the animal was suspected to be persistently infected. β -hemolytic colonies of *S. equi* strains were recovered from glycerol stocks following overnight growth on COBA strep select plates (bioMérieux). Their identity was confirmed by a lack of fermentation of trehalose, ribose and sorbitol in Purple broth (Becton Dickinson).

DNA preparation

A single colony of each *S. equi* strain was grown overnight in 3 ml of Todd Hewitt (TH) broth containing 30 µg/ml hyaluronidase (Sigma) at 37 °C in a 5 % CO₂ enriched atmosphere, centrifuged, the pellet re-suspended in 200 µl Gram +ve lysis solution (GenElute, Sigma) containing 250 units/ml mutanolysin and 2 x 10⁶ units/ml lysozyme and incubated for 1 hour at 37 °C to allow efficient cell lysis. DNA was then purified using GenElute spin columns according to manufacturer's instructions (all Sigma).

DNA Sequencing and variation detection

Library construction for Illumina sequencing was carried out as described previously (Quail et al. 2012). One isolate, NCTC9682, was sequenced on an Illumina GAII for 54 cycles from each end, producing paired-end reads with an expected insert size of 250 bp. The remaining isolates were mixed in pools of between 20 and 24 to produce multiplexed libraries that were sequenced on the Illumina HiSeq platform for 75 cycles from each end plus an 8-base index-sequence read. Again the expected insert size between the paired-end reads was 250 bp. Short reads were mapped against the reference *Se4047* genome (Accession numbers: FM204883 and NC_012471) using SMALT v0.7.4 (<https://www.sanger.ac.uk/resources/software/smalt>). Locations of deletions and short insertions were predicted using Pindel (Ye et al. 2009), and then validated by comparing the mapping of reads spanning indels to the reference genome and to a version of the same reference with the predicted indel included. If the inclusion of the indel improved mapping, that indel was retained, and reads realigned around it as per the remapping. SNPs were identified using a combination of SAMtools (Li et al. 2009) mpileup and BCFtools as previously described (Harris et al. 2010). Large deletions and duplications were identified for each isolate based on read coverage along the reference genome using a continuous hidden Markov model with three states: 0x coverage, 1x coverage and ≥2x coverage. Initial and transition frequencies were fitted to the data using a Baum-Welch optimisation, and the most likely sequence of hidden states was calculated using the Viterbi algorithm (Viterbi 1967). IS element insertion locations were identified by mapping read data to a set of IS sequences known to occur in *S. equi* and *S. zooepidemicus*. Where one

read of a pair mapped to an IS element, the non-mapping paired read was mapped to the reference Se4047 genome to identify the insertion site. To remove false positive insertion sites that may be caused by chimeric read pairs, a minimum of five reads were required to map to the same insertion location for it to be accepted.

Phylogenomic Analysis

Maximum likelihood phylogenetic reconstruction of variable sites was performed using RAxML v7.0.3 (Stamatakis 2006) under a general time reversible (GTR) evolutionary model using a gamma correction for among-site rate variation. 100 random bootstrap replicates were run to provide a measure of support for relationships in the maximum likelihood tree. A linear regression of root-to-tip distance versus isolation date was used to assess the fit of a strict molecular clock to the data, and gave a weak correlation for isolates with known isolation dates, excluding vaccine isolates (coefficient of determination, $R^2 = 0.25$). To test the null hypothesis that such an R^2 arose by chance alone, we repeated the root-to-tip analysis 1000 times with the tip dates of the isolates randomly permuted each time (Firth et al. 2010). In all cases, the data with random permutations gave lower R^2 values than the real data, so that we could reject our null hypothesis at the 0.001 level and accept the alternative hypothesis that the real data contains significant temporal signal. Bayesian reconstruction in BEAST v1.7 (Drummond and Rambaut 2007) was used to estimate substitution rates and times for divergences of internal nodes on the tree under a GTR model with a gamma correction for among-site rate variation. All combinations of strict, relaxed lognormal, relaxed exponential and random clock models and constant, exponential, expansion, logistic and skyline population models were evaluated. For each, three independent chains were run for 100 million generations, sampling every 10 generations. On completion each model was checked for convergence, both by checking ESS values were greater than 200 for key parameters, and by checking independent runs had converged on similar results. Models that failed to converge, including logistic population models, were discarded. Models were compared for their fit to the data using Bayes Factors based on the harmonic mean estimator as calculated by the program Tracer v1.4 from the BEAST package. The best-fit model combination was found to be a relaxed exponential clock model and a skyline population model, and so this combination was used for all

further analysis. A burn-in of 10 million states was removed from each of the three independent runs of this model before combining the results from those runs with the logcombiner program from the BEAST package. A maximum clade credibility (MCC) tree was created from the resulting combined trees using the treeAnnotator program, also from the BEAST package.

Accessory Genome Assembly

A pan genome for *S. equi* was created by identifying novel regions from *de novo* assemblies of each isolate. Assemblies were created with Velvet v1.2.09 (Zerbino and Birney 2008) using the VelvetOptimiser.pl v2.2.5 (<http://bioinformatics.net.au/software.shtml>) script to optimise the kmer length, expected coverage and coverage cut-off parameters based on the N50 statistic. A core genome was defined by removing the mobile prophages, ϕ Seq1-4, and ICE elements, ICESe1 and ICESe2, from the Se4047 reference genome. This core genome was then mapped to each assembly using NUCmer (Delcher 2002). All regions >200 bp in each assembly that did not match to the core genome were extracted and retained as accessory regions. All accessory regions for each isolate and the accessory regions from the reference Se4047 genome were then mapped against each other using NUCmer. Where two regions were identical in length and matched along their entire length, one was retained. Where one region was completely contained within another the longer region was retained. Any region of novel sequence >200 bp was also retained. Using this process for all pairwise comparisons led to production of a non-redundant set of accessory genome contigs. After filtering, these accessory contigs were appended to the core genome to form a pan genome. Finally, each assembly and the reference Se4047 genome were mapped to the pan genome using NUCmer to identify the regions of the pan genome present in each isolate.

Mutation frequency

Cultures of test bacteria were grown over night in TH broth at 37 °C in a 5 % CO₂ enriched atmosphere. Cultures were diluted to an OD_{600nm} of 0.5. The number of viable bacteria was enumerated by plating 100 µl of a 10⁻⁶ dilution onto each of five TH agar plates that were grown at 37 °C in a 5 % CO₂ enriched atmosphere for 24 hours. 100 µl of undiluted culture was plated onto each of five TH agar plates containing 0.03 µg ml⁻¹ rifampicin and grown at 37 °C in a 5 % CO₂

enriched atmosphere for 24 hours. Colonies were enumerated, averaged and the resistance frequency was calculated by dividing the total number of colonies by the number of rifampicin resistant colonies. The experiment was repeated in triplicate and an average resistance frequency was calculated across the three experiments. Statistical significance was calculated using an ANOVA test on the whole population.

Extraction of RNA

S. equi strains were inoculated onto COBA plates and grown overnight at 37 °C in an atmosphere containing 5 % CO₂. A single colony of each strain was inoculated into 5 ml Todd Hewitt broth containing 0.2 % yeast extract (THY) (Oxoid) and grown overnight at 37 °C in an atmosphere containing 5 % CO₂. Cultures were diluted 1/10 in fresh THY and grown to an OD_{600nm} of 0.6. 2 ml of each culture was then mixed with 4 ml of RNA protect (Qiagen) and pelleted by centrifugation for 10 minutes at 5000 x g and then 8000 x g for a further 10 minutes. Each pellet was re-suspended in 200 µl of tris-EDTA buffer (TE) (Fluka) containing 3 mg of lysozyme (Sigma), 500 units of mutanolysin (Sigma) and 60 µg hyaluronidase (Sigma). Total RNA was extracted using RNeasy mini, and DNase kits (all Qiagen) as per the manufacturer's instructions with the following alterations. After incubation at room temperature for 45 minutes with repeated vortexing, 700 µl of RLT buffer from the RNeasy mini kit was added and the tube vortexed briefly. The lysis mixture was transferred to a tube containing 0.05 g of acid washed glass beads (Sigma) and vortexed constantly for 5 minutes. The lysis mixture was centrifuged for 10 seconds at 16100 x g and the supernatant removed for RNA extraction with an RNeasy mini kit. RNA purity and quantity were determined using a Nanodrop ND1000 spectrophotometer (NanoDrop Technologies). Three biological replicates were performed.

Reverse Transcription

100 ng of RNA per sample were used to synthesise cDNA by reverse transcription (RT) in 20 µl reactions with random hexamers using a Verso cDNA kit (Thermo Scientific) as per the manufacturer's instructions.

qPCR to quantify *has* transcription

Transcription of SEQ_0269 (*hasA*), SEQ_0270 (*hasB*), SEQ_0271 (*hasC*) and the housekeeping gene *gyrA* (SEQ_1170) were quantified by quantitative polymerase chain reaction (qPCR). Reactions contained 1 x Kapa SYBR fast qPCR mix (Anachem), 0.3 μ M forward and reverse primers (Table S3) and 6 μ l of a 1/10 dilution of cDNA made up to 20 μ l with ddH₂O and thermocycled at 95 °C for 3 minutes followed by 40 cycles of 95 °C for 30 seconds, 60 °C 10 seconds then 95 °C for 15 seconds. A ramp step from 60 °C to 95 °C with SYBR reads every 0.3 °C was performed to calculate the dissociation curves of products. No template and no RT controls were included to confirm the absence of contaminating DNA and RNA in samples. Copy numbers were calculated from standard curves and normalised to *gyrA*. The mean values for the three biological replicates were calculated.

PCR to detect tandem amplification of the *has* locus

PCR reactions contained 1 x Sigma Taq buffer, 2 mM deoxyribonucleotide triphosphate (dNTP) mix, 2 mM MgCl₂, 1 μ M forward and reverse primers (Supplementary Table 3), 0.1 μ l Sigma Taq DNA polymerase and 4 μ l DNA in a total volume of 20 μ l. Reactions were thermocycled at 95 °C for 2 minutes followed by 35 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds and 72 °C for 5 minutes and finally 72 °C for 5 minutes. Products were resolved on a 0.7 % agarose gel.

Quantification of surface hyaluronic acid

S. equi strains were grown overnight in THY at 37 °C in an atmosphere containing 5 % CO₂ and used to inoculate 10 ml of THY. Triplicate cultures were grown to an OD₆₀₀ of 0.6 at 37 °C in an atmosphere containing 5 % CO₂. Bacteria were washed twice with water, re-suspended in 0.5 ml of water and hyaluronic acid released by vigorous shaking with 1 ml of chloroform for 15 seconds. The mixture was allowed to stand at room temperature for 1 hour before centrifugation at 16000 x g. The aqueous phase was removed and the amount of capsule determined as described previously (Schrager et al. 1996; Darmstadt et al. 2000). DNA was extracted from 1.5 ml of the OD_{600nm} 0.6 culture using the GenElute spin columns as described above. The DNA was quantified using a Nanodrop ND1000 spectrophotometer and the number of bacteria was determined using

the life technologies DNA copy number calculator (<https://www.lifetechnologies.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/dna-copy-number-calculator.html>). The amount of hyaluronic acid was expressed as femtograms per copy of bacterial DNA.

Allelic replacement

The generation of the *eqbE* deletion mutant, $\Delta eqbE$, has been described previously (Heather et al. 2008).

Experimental infection of ponies

Ponies of 18 months of age were transferred to a containment unit three days before challenge. Each pony was challenged with *S. equi* strain 4047 or the $\Delta eqbE$ deletion mutant via the spraying of a 2 ml culture containing 5×10^7 cfu into each nostril. Bacteria were grown overnight in Todd Hewitt broth and 10 % fetal calf serum (THBS) in a 5 % carbon dioxide enriched atmosphere at 37 °C, diluted 40-fold in fresh pre-warmed THBS, further cultivated and harvested at an $OD_{600nm} = 0.3$. This infection dose has been shown to optimize the infection rate, whilst avoiding overwhelming the host immune response (Hamilton et al. 2006; Guss et al. 2009). Ponies were monitored for the onset of clinical signs of disease over a period of three weeks post challenge by daily physical examination, rectal temperature, lymph node swelling and nasal discharge scoring. Blood samples were taken for quantification of fibrinogen and neutrophil levels by total white blood count performed on Beckman-Coulter ACTdiff analyzer with a manual differential count to calculate % neutrophils. Post mortem examination was performed on all ponies following the onset of early clinical signs of infection such as pyrexia and a reluctance to eat dry pelleted food, preferring haylage or water and prior to the onset of later signs according to strict welfare guidelines at the Animal Health Trust, or on reaching the study endpoint at 3 weeks post challenge. The severity of disease pathology was quantified according to a scoring system described previously (Guss et al. 2009). This work was conducted under the auspices of a Home Office Project License and

following ethical review and approval by the Animal Health Trust's Animal Welfare and Ethical Review Body (RPP 01_12).

DATA ACCESS:

Short reads for all sequenced isolates have been submitted to the European Nucleotide Archive (ENA) under the accession number ERP000812 (<http://www.ebi.ac.uk/ena/data/view/ERP000812&display=html>). Individual accession numbers of sequences for all isolates are listed in Supplementary Table 1.

ACKNOWLEDGEMENTS:

The authors would like to thank the Horserace Betting Levy Board for funding the analysis of the *eqbE* mutant (ref: Vet/prj/758). The Horse Trust funded the collection of isolates from UK outbreaks of strangles (G1606). We are grateful to Prof. John Timoney (University of Kentucky, USA), Prof. Tom Buckley (Irish Equine Centre, Ireland), Nicolas De Brauwere (Redwings Horse Sanctuary, UK), Dr. Jennifer Stewart (Animal Health Trust, UK), Prof. Bengt Guss (Swedish University of Agricultural Sciences, Sweden), Prof. Gunther van Loon (Ghent University, Belgium), Maria Cristina Marotti Campi (Al Khaledia Equine Hospital, Saudi Arabia) and Els Acke (Massey University, New Zealand) for contributing isolates to this study. We thank the core sequencing and informatics teams at the Sanger Institute for their assistance and The Wellcome Trust for its support of the Sanger Institute Pathogen Genomics and Biology groups. SRH, JP and MTGH were supported by Wellcome Trust grant 098051.

AUTHOR CONTRIBUTIONS:

S.R.H. analysed and interpreted the data and wrote the paper. C.R., K.F.S., K.S.W and R.P. performed experiments, carried out data analyses and cultured the isolates. J.P. helped interpret the data and write the paper. M.T.G.H. jointly conceived the project with A.S.W., facilitated sequencing of the isolates, carried out data analysis and helped write the paper. A.S.W. conceived and ran the project, collected samples and helped write the paper.

REFERENCES:

- Anon. 1831. The Strangles. In *The Horse with a Treatise on Draught*, pp. 149-151. Baldwin and Cradock, London.
- Bazeley PL. 1940. Studies with equine streptococci 1. *The Australian Veterinary Journal* **16**: 140-146.
- Blaine D. 1841. Strangles. In *Outlines of the Veterinary Art*, pp. 432-433. Longman, Orme and Co., London.
- Borst LB, Patterson SK, Lanka S, Barger AM, Fredrickson RL, Maddox CW. 2011. Evaluation of a commercially available modified-live *Streptococcus equi* subsp *equi* vaccine in ponies. *Am J Vet Res* **72**(8): 1130-1138.
- Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SA, Siezen RJ, Borgdorff M et al. 2013. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis* **13**: 110.
- Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S et al. 2014. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* **46**(3): 279-286.
- Clabby BJ. 1976. A short history of the Royal Army Veterinary Corps. *Proc R Soc Med* **69**(2): 93-96.
- Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC bioinformatics* **9**: 539.
- Cornejo OE, Lefebure T, Bitar PD, Lang P, Richards VP, Eilertson K, Do T, Beighton D, Zeng L, Ahn SJ et al. 2013. Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol* **30**(4): 881-893.
- Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS et al. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**(6016): 430-434.
- Darmstadt GL, Mentele L, Podbielski A, Rubens CE. 2000. Role of group A streptococcal virulence factors in adherence to keratinocytes. *Infect Immun* **68**(3): 1215-1221.
- Davies MR, Holden MT, Coupland P, Chen JH, Venturini C, Barnett TC, Zakour NL, Tse H, Dougan G, Yuen KY et al. 2015. Emergence of scarlet fever *Streptococcus pyogenes* emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat Genet* **47**(1): 84-87.
- Delcher AL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic acids research* **30**(11): 2478-2483.
- Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* **7**: 214.
- Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol* **27**(9): 2038-2051.
- Flores AR, Jewell BE, Fittipaldi N, Beres SB, Musser JM. 2012. Human disease isolates of serotype M4 and M22 group A streptococcus lack genes required for hyaluronic acid capsule biosynthesis. *MBio* **3**(6): e00413-00412.
- Graham MR, Smoot LM, Migliaccio CA, Virtaneva K, Sturdevant DE, Porcella SF, Federle MJ, Adams GJ, Scott JR, Musser JM. 2002. Virulence control in group A *Streptococcus* by a two-component gene regulatory system: global expression profiling and *in vivo* infection modeling. *Proc Natl Acad Sci U S A* **99**(21): 13855-13860.
- Guss B, Flock M, Frykberg L, Waller AS, Robinson C, Smith KC, Flock JI. 2009. Getting to grips with strangles: an effective multi-component recombinant vaccine for the protection of horses from *Streptococcus equi* infection. *PLoS Pathog* **5**(9): e1000584.
- Hamilton A, Robinson C, Sutcliffe IC, Slater J, Maskell DJ, Davis-Poynter N, Smith K, Waller A, Harrington DJ. 2006. Mutation of the maturase lipoprotein attenuates the virulence of *Streptococcus equi* to a greater extent than does loss of general lipoprotein lipidation. *Infect Immun* **74**(12): 6907-6919.
- Harper M, Lee CJ. 2012. Genome-wide analysis of mutagenesis bias and context sensitivity of *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine (NTG). *Mutat Res* **731**(1-2): 64-67.

- Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**(5964): 469-474.
- He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D, Bamford KB et al. 2013. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet* **45**(1): 109-113.
- Heather Z, Holden MT, Steward KF, Parkhill J, Song L, Challis GL, Robinson C, Davis-Poynter N, Waller AS. 2008. A novel streptococcal integrative conjugative element involved in iron acquisition. *Mol Microbiol* **70**(5): 1274-1292.
- Herbeck JT, Nickle DC, Learn GH, Gottlieb GS, Curlin ME, Heath L, Mullins JI. 2006. Human immunodeficiency virus type 1 *env* evolves toward ancestral states upon transmission to a new host. *J Virol* **80**(4): 1637-1644.
- Holden MT, Heather Z, Paillot R, Steward KF, Webb K, Ainslie F, Jourdan T, Bason NC, Holroyd NE, Mungall K et al. 2009. Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathogens* **5**(3): e1000346.
- Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H et al. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res* **23**(4): 653-664.
- Jacobs AA, Goovaerts D, Nuijten PJ, Theelen RP, Hartford OM, Foster TJ. 2000. Investigations towards an efficacious and safe strangles vaccine: submucosal vaccination with a live attenuated *Streptococcus equi*. *Vet Rec* **147**(20): 563-567.
- Kelly C, Bugg M, Robinson C, Mitchell Z, Davis-Poynter N, Newton JR, Jolley KA, Maiden MC, Waller AS. 2006. Sequence variation of the SeM gene of *Streptococcus equi* allows discrimination of the source of strangles outbreaks. *J Clin Microbiol* **44**(2): 480-486.
- Lefebure T, Richards VP, Lang P, Pavinski-Bitar P, Stanhope MJ. 2012. Gene repertoire evolution of *Streptococcus pyogenes* inferred from phylogenomic analysis with *Streptococcus canis* and *Streptococcus dysgalactiae*. *PLoS One* **7**(5): e37607.
- Lemey P, Rambaut A, Pybus OG. 2006. HIV evolutionary dynamics within and among hosts. *AIDS Rev* **8**(3): 125-140.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
- Lythgoe KA, Fraser C. 2012. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proc Biol Sci* **279**(1741): 3367-3375.
- Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2011. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* **40**(1): e6.
- Miles WJ. 1868. Strangles. In *Modern practical farriery, a complete system of the veterinary art* pp. 290-291. William Mackenzie, London.
- Newton JR, Verheyen K, Talbot NC, Timoney JF, Wood JL, Lakhani KH, Chanter N. 2000. Control of strangles outbreaks by isolation of guttural pouch carriers identified using PCR and culture of *Streptococcus equi*. *Equine Vet J* **32**(6): 515-526.
- Newton JR, Wood JL, Dunn KA, DeBrauwere MN, Chanter N. 1997. Naturally occurring persistent and asymptomatic infection of the guttural pouches of horses with *Streptococcus equi*. *Vet Rec* **140**(4): 84-90.
- Patty O, Cursons R. 2014. The molecular identification of *Streptococcus equi* subsp. *equi* strains isolated within New Zealand. *N Z Vet J* **62**(2): 63-67.
- Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, Oyola SO. 2012. Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9**(1): 10-11.
- Rufus J, ed. 1251. *De Medicina Equorum*.
- Schrager HM, Rheinwald JG, Wessels MR. 1996. Hyaluronic acid capsule and the role of streptococcal entry into keratinocytes in invasive skin infection. *J Clin Invest* **98**(9): 1954-1958.
- Schutz JW. 1888. The Streptococcus of Strangles. *The Journal of Comparative Pathology and Therapeutics* **1**(3): 191 - 208.

- Shea PR, Beres SB, Flores AR, Ewbank AL, Gonzalez-Lugo JH, Martagon-Rosado AJ, Martinez-Gutierrez JC, Rehman HA, Serrano-Gonzalez M, Fittipaldi N et al. 2011. Distinct signatures of diversifying selection revealed by genome analysis of respiratory tract and invasive bacterial populations. *Proc Natl Acad Sci U S A* **108**(12): 5039-5044.
- Solleysel J. 1664. *Le parfait maréchal*. Gervais Clousier, Paris.
- Spooner WC. 1842. Strangles. In *A Compendium of The Veterinary Art*, pp. 129-131. Longman, Brown, Green and Longmans, London.
- Stamatakis A. 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21): 2688-2690.
- Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry CE, 3rd et al. 2012. Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J Infect Dis* **206**(11): 1724-1733.
- Timoney JF. 1985. Protection of equines against *Streptococcus equi*. Cornell Research Foundation Inc., USA.
- Todd AG. 1910. Strangles. *J Comp Pathol Therap* **23**: 212-229.
- Viterbi AJ. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**(2): 260-269.
- Walker JA, Timoney JF. 2002. Construction of a stable non-mucoid deletion mutant of the *Streptococcus equi* Pinnacle vaccine strain. *Vet Microbiol* **89**(4): 311-321.
- Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW et al. 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* **13**(2): 137-146.
- Waller AS, Robinson C. 2013. *Streptococcus zooepidemicus* and *Streptococcus equi* evolution: the role of CRISPRs. *Biochem Soc Trans* **41**(6): 1437-1443.
- Webb K, Jolley KA, Mitchell Z, Robinson C, Newton JR, Maiden MC, Waller A. 2008. Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group. *Microbiology* **154**(Pt 10): 3016-3024.
- White J. 1825. Strangles. In *A Compendium of the Veterinary Art*, pp. 180-181. Longman, Hurst, Rees, Orme, Brown and Green, London.
- Williams W. 1884. Febris Pyogenica, or Strangles. In *The Principles and Practice of Veterinary Surgery*, (ed. JA Churchill), pp. 43-51. John Menzies and Co., Edinburgh and Glasgow.
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**(21): 2865-2871.
- Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18**(5): 821-829.

Figure 1. Phylogenetic reconstruction and assessment of temporal signal A) Maximum-likelihood phylogeny of *S. equi* based on core genome SNPs. Clusters are colored as reconstructed by BAPs. Columns adjacent to the tree represent the BAPS cluster and MLST type associated with the strains. Dashed lines linking branches to these columns are for ease of viewing and are colored with the appropriate BAPs cluster color. Branch colors represent branches corresponding to isolates from horses *JKS628* (red), *JKS731* (orange), *JKS121* (green) and two Pinnacle vaccine strains (blue). Thick dashed lines indicate vaccine strains: Blue = Pinnacle IN, green = Equilis Strep E and red = *Se1866*, the basis of the Strangvac vaccine. The position of the reference isolate, *Se4047*, is indicated by an asterisk. B) Root-to-tip plot of all isolates in the ML tree in panel A. C) Residual plot of root-to-tip plot in B. Unbroken red line indicates the mean, and dashed red lines two standard deviations from the mean. D) Root-to-tip plot with the isolates falling outside of the 95 % confidence intervals in B excluded. Colored points in panels B to D correspond to the isolates on colored branches in A.

Figure 2. Detailed epidemiology of three *S. equi* outbreaks in the UK. Isolates highlighted on the phylogeny were collected from three stables experiencing strangles outbreaks, as indicated by the key. In each case, isolates did not fall into single clades on the phylogeny, instead being grouped into outbreak (O) and persistent (P) clusters. The positions of the outlier branches identified by root-to-tip analysis are labeled in blow-ups of the relevant clades. For the Essex outbreak, the number of SNPs within and between clades are indicated by the numbers in the gray boxes. Colored branches in the tree correspond to those in Fig. 1.

Figure 3. Box plots showing mean (dots), median (center horizontal line) and 95 % HPD (highest posterior density) for the time to the most recent common ancestor (tMRCA) of all isolates in the study under various combinations of clock and population model in BEAST. The dates of the Crimean War, World War I (WWI) and World War II (WWII) are indicated by the gray shaded regions.

Figure 4. Convergent deletion and amplification of the *has* locus in isolates recovered from independent persistently infected horses. A) A representation of the genome annotation around the *has* locus. B) Eight isolates, all from different horses, independently exhibited duplications of the region between IS repeats (indicated by gray columns). This plot represents mapped sequence read coverage of JKS628d, showing an increase in coverage spanning the *has* operon, representing duplication of the region. Read mapping reduction in the gray repeat regions is due to an inability to map reads uniquely to those regions. C) 12 isolates from 8 horses exhibited 7 independent deletions of different regions of the *has* locus. Plots i to vii show read mapping of the *has* locus in example isolates for each of the deletions. Names to the right of the plots indicate isolates exhibiting the same deletion. Isolate names in bold are the examples used for the mapping plot. Read mapping drops to zero in the gray repeat regions due to an inability to map reads uniquely to this region. Deletion breakpoints, identified by finding breakpoints in reads, are indicated by vertical black lines joined by dashed bridges. Brackets indicate where one breakpoint location is uncertain due to falling in a repetitive region. The blue arrow in vii indicates the location of a novel IS insertion.

Figure 5. Convergent deletion of the equibactin locus in 7 isolates recovered from four independent persistently infected horses. A) A representation of the genome annotation around the equibactin cluster. B) Read mapping of the equibactin locus in example isolates exhibiting the four deletion variants. Names to the right of the plots indicate isolates exhibiting the same deletion. Isolate names in bold are the examples used for the mapping plot. Read mapping drops to zero in the gray repeat region due to an inability to map reads uniquely to this region.

Figure 6. Pathology score of two groups of seven Welsh mountain ponies challenged with wild-type Se4047 or the $\Delta eqbE$ deletion mutant. The box illustrates the 50% distribution, the line indicates the median value, the cross indicates the mean and error bars indicate the maximum and minimum distribution.

ADownloaded from genome.cshlp.org on July 15, 2026 . Published by Cold Spring Harbor Laboratory Press

KEY:

BAPS cluster

- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4

ST

- ST179
- ST151

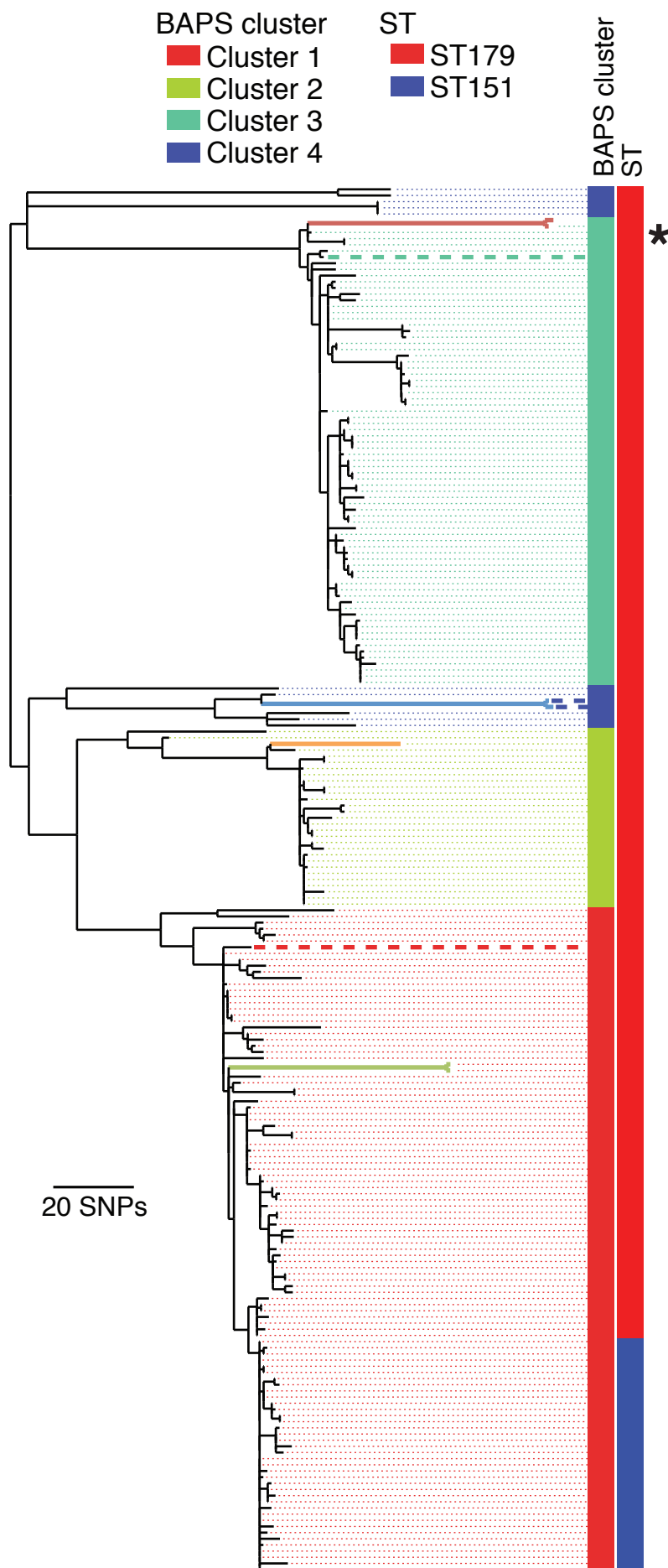
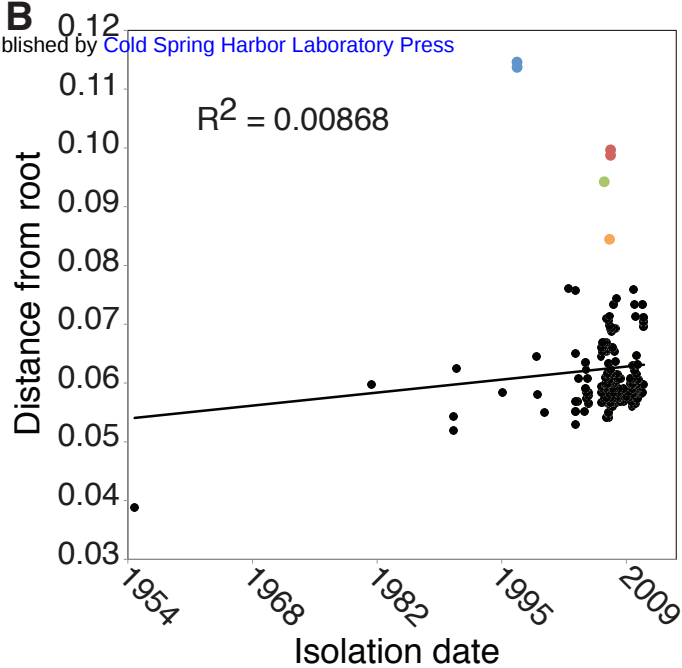
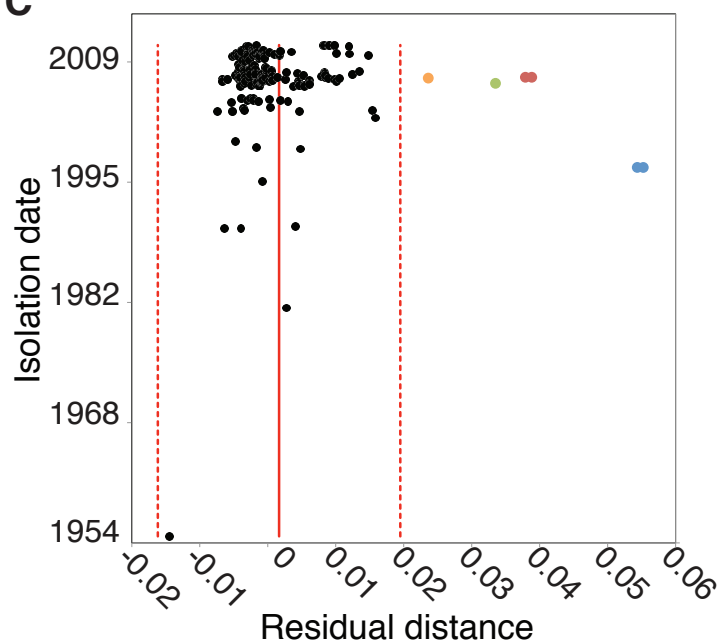
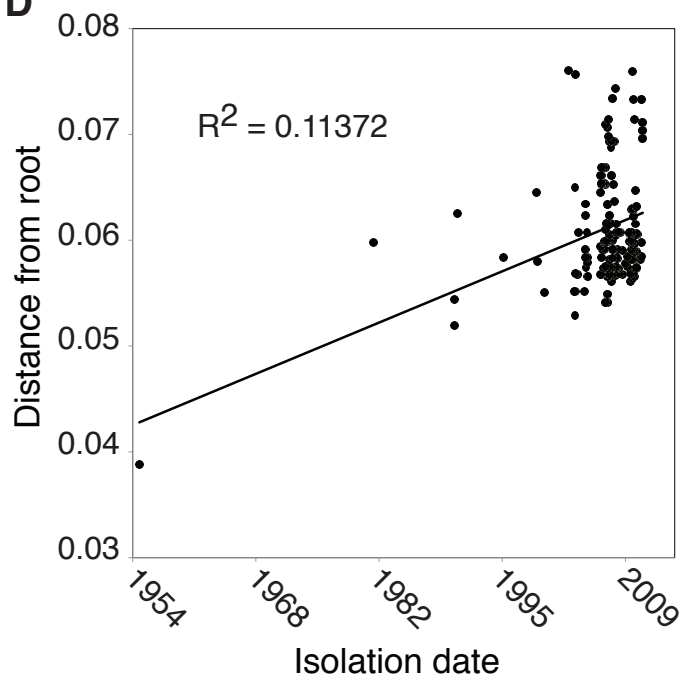
**B****C****D**

Figure 1

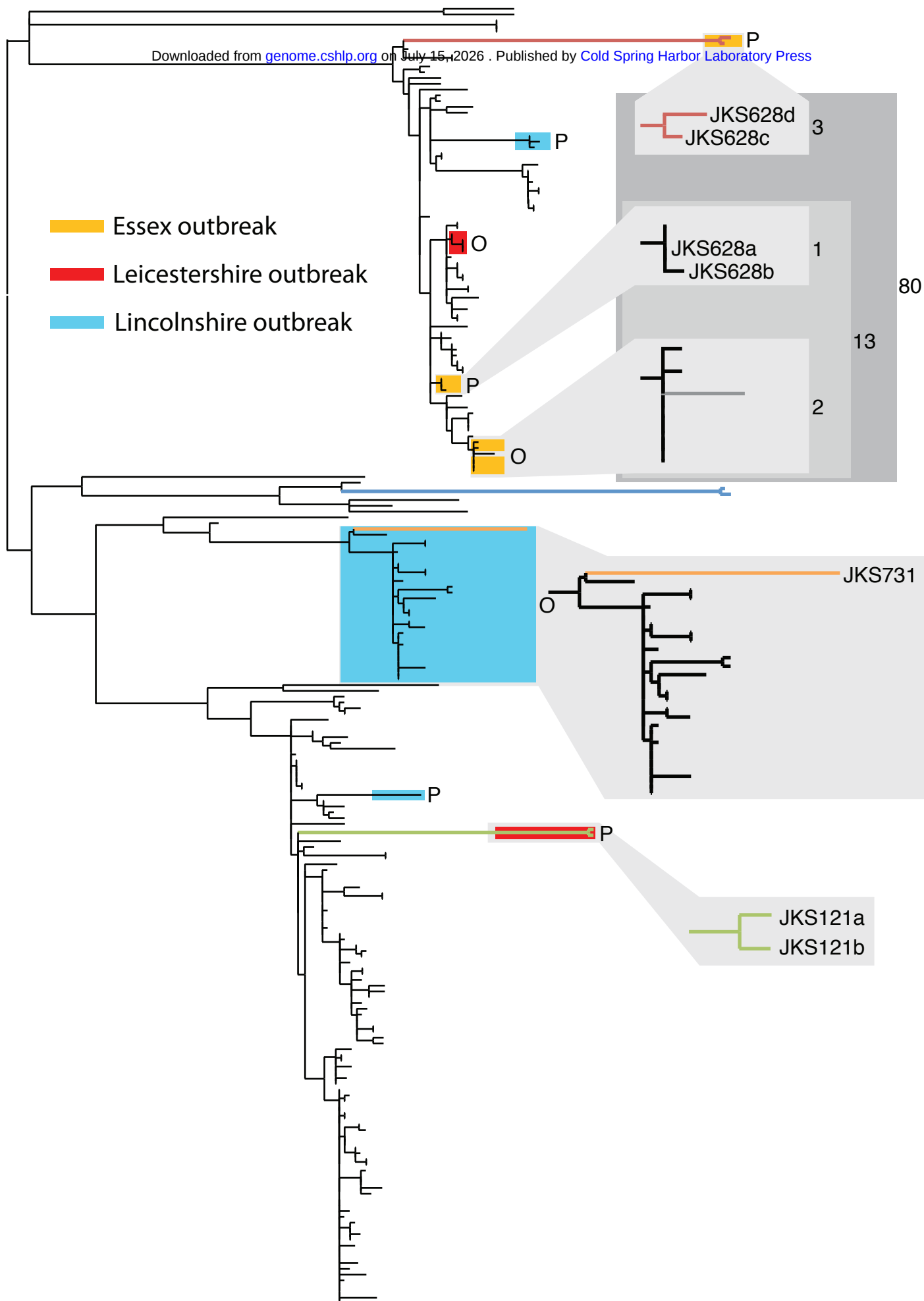


Figure 2

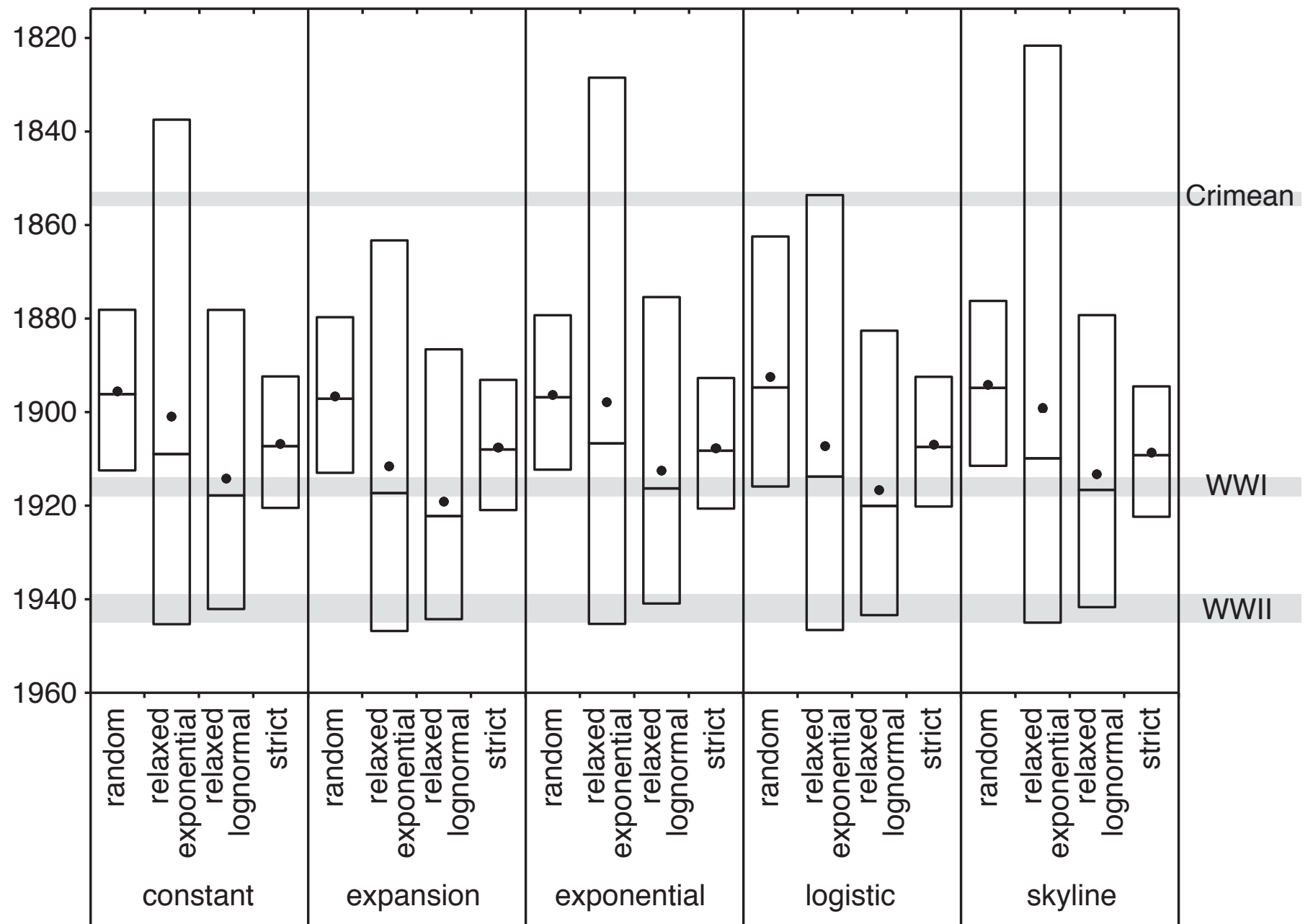


Figure 3

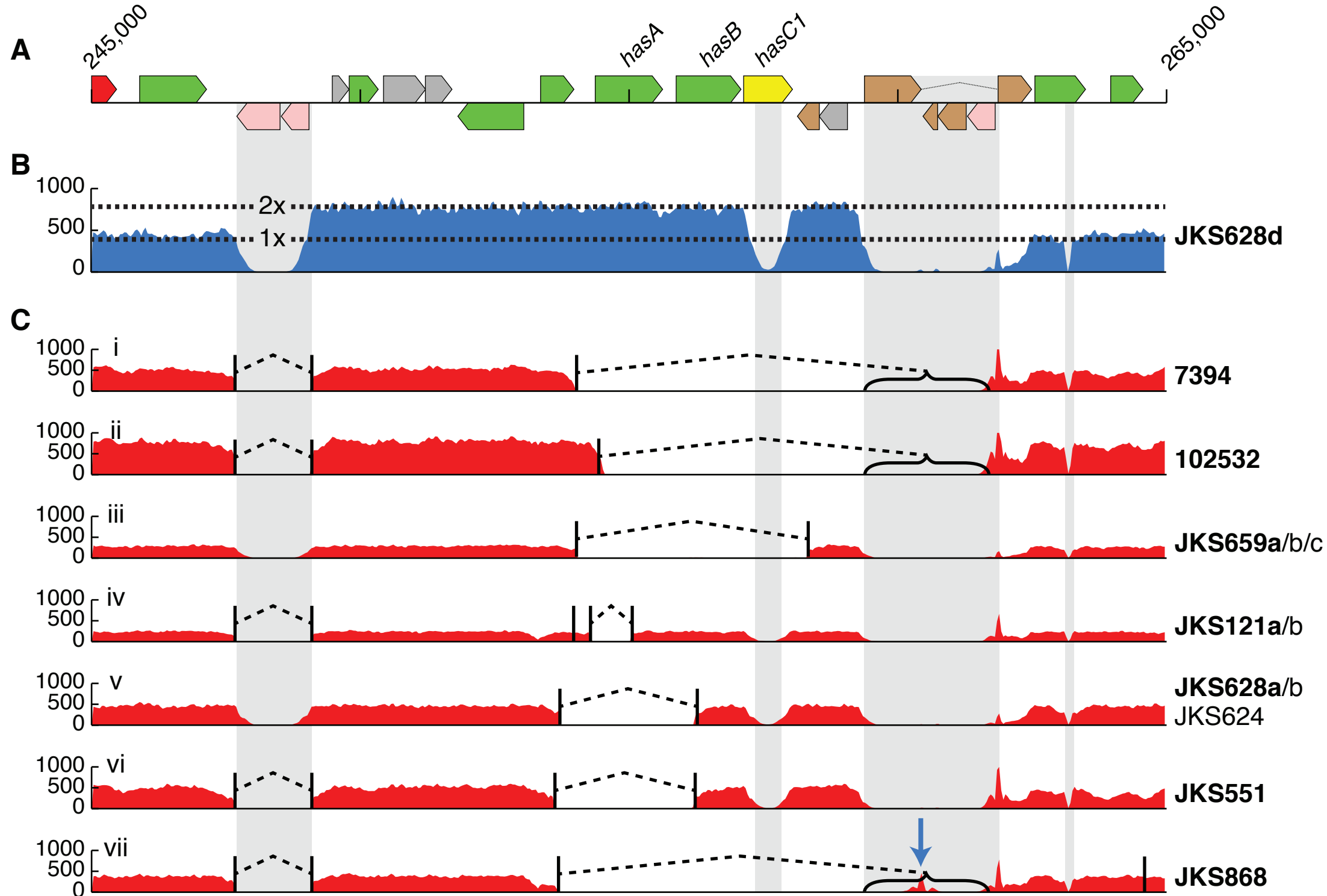


Figure 4

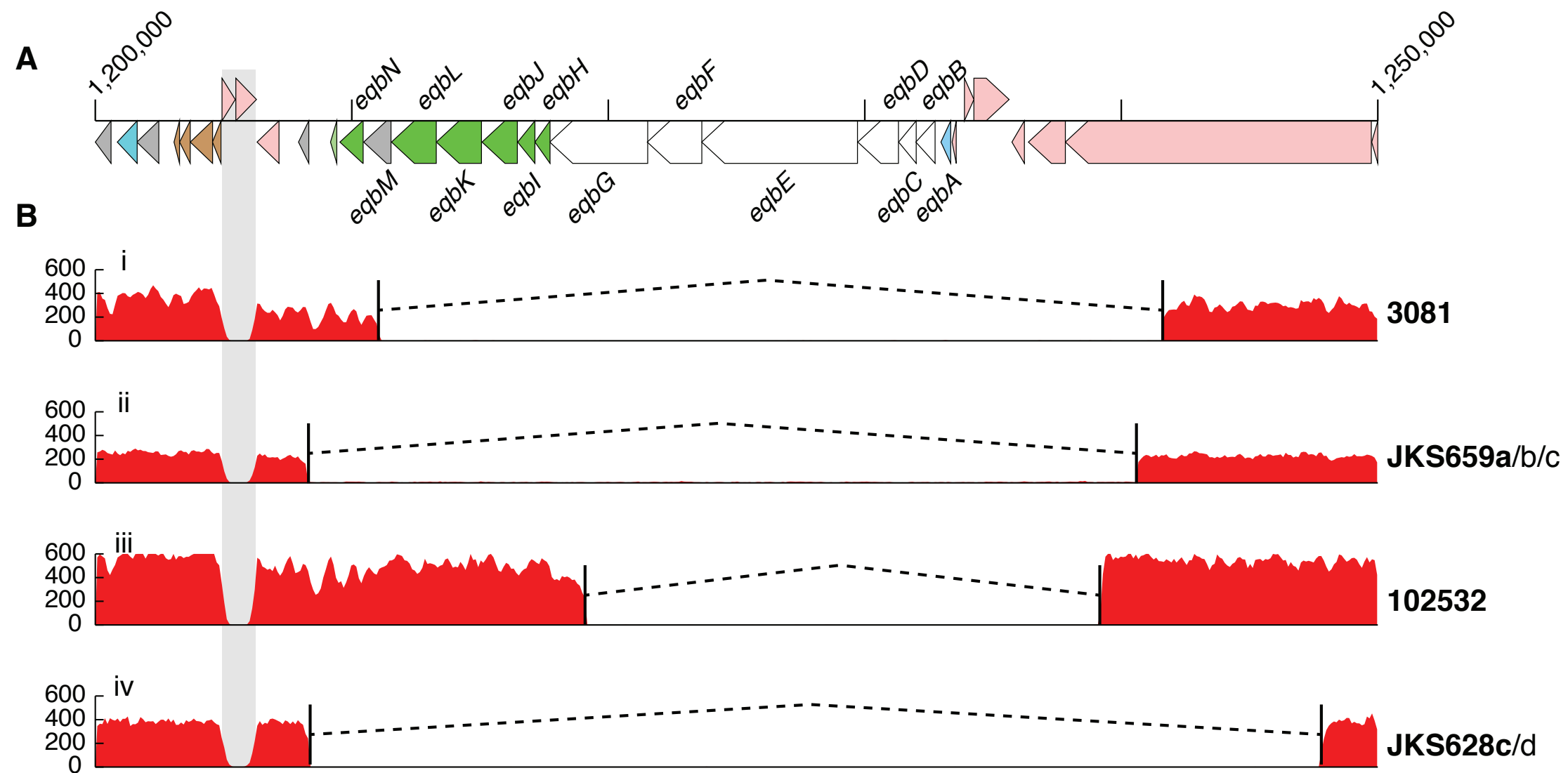


Figure 5

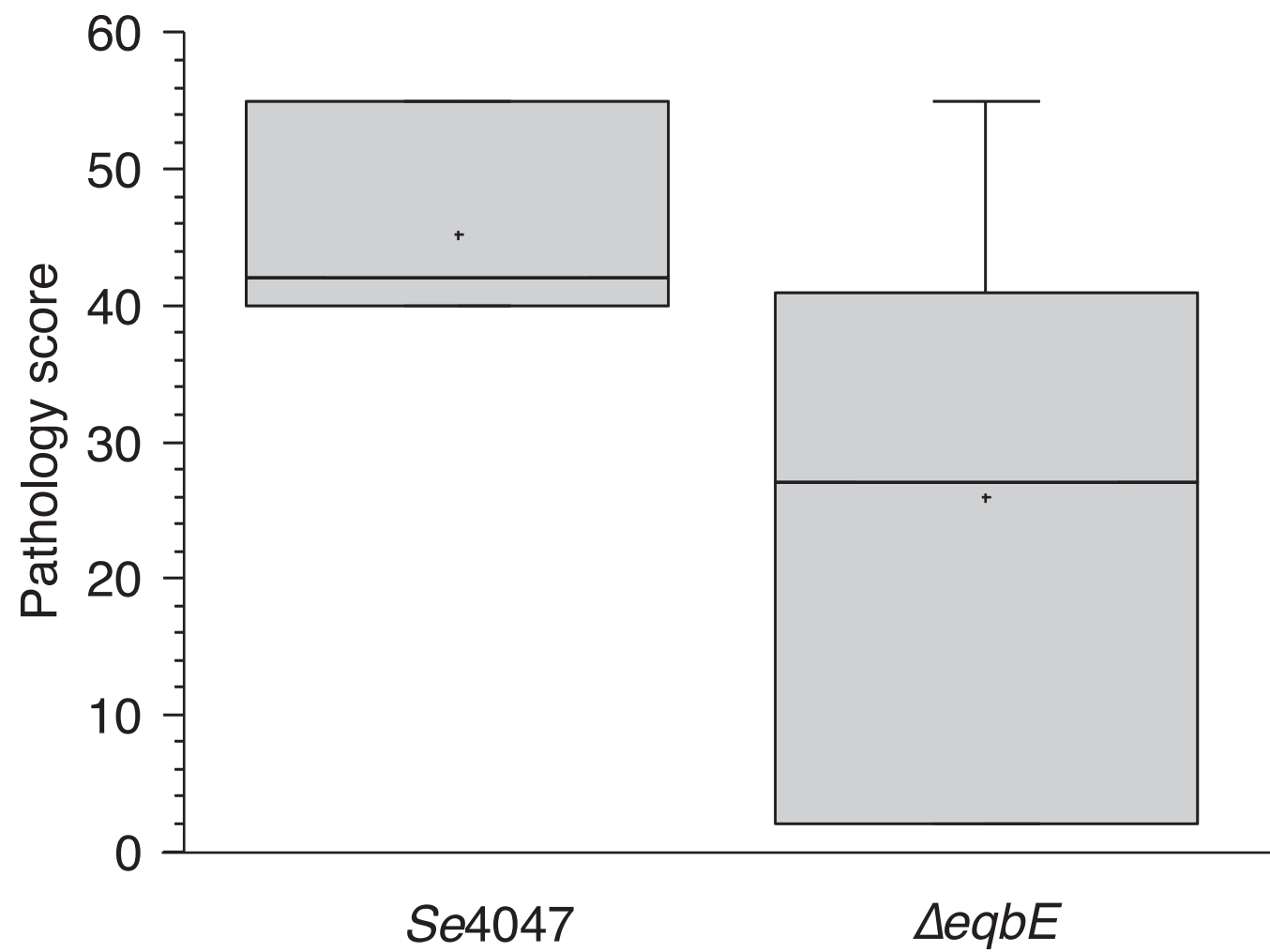


Figure 6