Published in final edited form as:
Nat Genet. 2008 August ; 40(8): 987–993. doi:10.1038/ng.195.

High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi

Kathryn E Holt[1], **Julian Parkhill**[1], **Camila J Mazzoni**[2,3], **Philippe Roumagnac**[3,4], **François-Xavier Weill**[5], **Ian Goodhead**[1,8], **Richard Rance**[1], **Stephen Baker**[1,6], **Duncan J Maskell**[7], **John Wain**[1], **Christiane Dolecek**[6], **Mark Achtman**[2,3], and **Gordon Dougan**[1]

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK [2]Environmental Research Institute, University College Cork, Lee Road, Cork, Ireland [3]Max-Planck-Institut für Infektionsbiologie, Department of Molecular Biology, Charitéplatz 1, 10117, Berlin, Germany [4]Université Mixte de Recherche 6191 Centre National de la Recherche Scientifique - Commissariat à l'Énergie Atomique-Aix-Marseille Université, Commissariat à l'Énergie Atomique Cadarache, 13108 Saint Paul lez Durance, France [5]Institut Pasteur, Laboratoire des Bactéries Pathogènes Entériques, 28 rue du docteur Roux, 75724 Paris cedex 15, France [6]Oxford University Clinical Research Unit, Hospital for Tropical Diseases, 190 Ben Ham Tu, District 5, Ho Chi Minh City, Vietnam [7]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK

Abstract

Isolates of *Salmonella enterica* serovar Typhi (Typhi), a human-restricted bacterial pathogen that causes typhoid, show limited genetic variation. We generated whole-genome sequences for 19 Typhi isolates using 454 (Roche) and Solexa (Illumina) technologies. Isolates, including the previously sequenced CT18 and Ty2 isolates, were selected to represent major nodes in the phylogenetic tree. Comparative analysis showed little evidence of purifying selection, antigenic variation or recombination between isolates. Rather, evolution in the Typhi population seems to be characterized by ongoing loss of gene function, consistent with a small effective population size. The lack of evidence for antigenic variation driven by immune selection is in contrast to strong adaptive selection for mutations conferring antibiotic resistance in Typhi. The observed patterns of genetic isolation and drift are consistent with the proposed key role of asymptomatic carriers of Typhi as the main reservoir of this pathogen, highlighting the need for identification and treatment of carriers.

Typhoid fever, along with plague, cholera and smallpox, is one of the classical infectious diseases of humans. The disease, which is spread via oral ingestion of contaminated food or water, is caused by *Salmonella enterica* serovar Typhi (Typhi), a Gram-negative bacterium classified as a serovar of the species *S. enterica*1. *S. enterica* is a broad and promiscuous species with isolates able to cause gastroenteritis in a range of animals, including humans2. In contrast to most other *S. enterica* serovars, Typhi has forsaken the promiscuous lifestyle

Correspondence should be addressed to K.E.H. (kh2@sanger.ac.uk)..
[8]Present address: School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, UK

AUTHOR CONTRIBUTIONS

G.D., J.P., M.A., P.R. and J.W. designed the study; F.-X.W. and C.D. contributed isolates for analysis; I.G. and R.R. performed 454 and Solexa sequencing; K.E.H. and S.B. performed validation experiments; D.J.M. co-supervises the PhD studies of K.E.H. and contributed to experimental design; K.E.H. and C.J.M. analysed data and K.E.H., J.P., P.R. and G.D. wrote the manuscript.

Note: Supplementary information is available on the Nature Genetics website.

Europe PMC Funders Author Manuscripts

to become a human-restricted pathogen causing both acute systemic infections (typhoid fever) and chronic infections (asymptomatic carriers). A number of evolutionary processes have been implicated in the adaptation of Typhi to this specialized niche, including the horizontal acquisition of several *Salmonella* pathogenicity islands (SPIs)3,4 and extensive loss of gene function3.

Typhi, together with other human pathogens such as *Yersinia pestis, Bordetella pertussis* and *Bacillus anthracis,* is regarded as a monomorphic organism, as the genomes of individual Typhi isolates are highly conserved and clonally related. A recent study involving the DNA sequencing of 199 gene fragments from a global collection of 105 Typhi isolates detected only 82 SNPs5. Analysis of the SNP data resolved Typhi into a rooted, fully parsimonious phylogenetic tree defining 85 genetically distinct haplotypes (H1-H85, Supplementary Fig. 1 online). The availability of a robust phylogenetic tree proved to be a useful framework against which to investigate the recent evolution of phenotypic traits such as the acquisition of resistance to fluoroquinolones, a class of antibiotics used to treat typhoid fever5-7.

Because Typhi shows such low levels of genetic variation, further studies require a whole-genome approach. Complete genome sequences are available for two Typhi isolates, CT18 and Ty2 (refs. 3,8). However, several new sequencing technologies have been developed that make draft genome sequencing simpler and more cost effective9. In order to capture variation in Typhi at the whole-genome level and minimize sampling bias10, we sequenced an additional 17 Typhi isolates dispersed in the phylogenetic tree, using a combination of 454 (Roche) and Solexa (Illumina) sequencing technologies.

RESULTS

Choice of Typhi isolates for whole-genome sequencing

In order to capture as much information as possible about the distribution of genomic variation in the Typhi population, we prepared DNA from CT18, Ty2 and 17 other isolates for a combination of 454 and Solexa sequencing (see Methods and Table 1). To limit selection bias as much as possible, we chose Typhi isolates from central haplotype clusters together with selected isolates from radial haplotype groups and subjected these to 454 sequencing (Supplementary Fig. 1). To gain additional insight into SNP variation among recently expanding haplotypes, we used Solexa sequencing to generate sequence information from an additional six isolates from the H58 group, which has undergone recent clonal expansion in Southeast Asia5,7, and a second isolate from the H59 group, known to be circulating in Indonesia11. Three isolates, including one H58 and one H59 isolate, were sequenced using both 454 and Solexa, and the results were compared.

SNPs and phylogenetic analysis

We detected high-quality SNPs by mapping 454 contigs or Solexa reads to the finished CT18 sequence (see Methods). Our analysis focused on the nonrepetitive component of the genome, and we did not attempt to identify single-base insertions or deletions. Repetitive sequences, including VNTRs, exact repeats of ≥20 bp, >95% identical repeats of >50 bp and phage and insertion sequences (*IS*), account for 7.4% of the CT18 genome (Supplementary Table 1a online). Here, we excluded these classes of repetitive sequences from SNP analysis as (i) non-identical repeats can appear indistinguishable from SNPs, particularly with short sequencing reads (100-250 bp for 454, 25 bp for Solexa), (ii) assembly and mapping of short reads are unreliable in repetitive regions and (iii) repeated regions may be subject to different selective pressures compared to the rest of the genome, for example, recombination between repeat copies. After excluding these sequences, we identified a total of 1,964 SNPs

in the nonrepetitive genome, approximately 1 every 2,300 bp. Details of these SNPs are given in Supplementary Table 2 online.

We determined complete allele data for 1,787 SNPs (missing data was due to low coverage or deletion of SNP loci in one or more isolates). These SNPs traced the same phylogenetic tree as previously defined5 (Supplementary Fig. 1) but provided better estimates of branch lengths and greatly increased resolution, particularly within the H58 and H59 groups (Fig. 1). By comparing sequence data from 454, Solexa and published sequences, we determined cut-offs for quality filtering (Supplementary Fig. 2 online) and estimated a false-positive rate of 2.7% and SNP detection sensitivity of 82-99.7% for both sequencing platforms (see Methods and Supplementary Table 1b). This apparently high false-positive rate is due to the extreme paucity of true SNPs; the actual error rate of the sequencing technologies is very low (around 7 to 10 errors in 4.45×10^6 base pairs for Solexa resequencing on CT18 and Ty2; see Supplementary Methods online). Only ten SNPs (0.56%) did not fit the previously determined phylogenetic tree, two of which are confirmed examples of convergent evolution at sites under adaptive selection in *gyrA* (see below). Thus, we have little reason to suspect high error rates among allele assignments, or to doubt the phylogenetic tree structure shown in Figure 1.

Using the resulting rooted phylogenetic tree, we were able to group mutations into relative age groups: (i) recent mutations, furthest from the root and lying on intra-haplotype branches, (ii) intermediate mutations, lying on haplotype-specific branches, and (iii) older mutations, lying on branches closest to the root and shared by multiple haplotypes. The distribution of SNPs and other variants in each group is shown in Table 2.

d*N*/d*S* as a measure of stabilizing selection

The ratio of nonsynonymous to synonymous SNP rates, d*N*/d*S*, is a common measure of stabilizing selection. A d*N*/d*S* ratio close to 1 indicates no selection against nonsynonymous SNPs, whereas d*N*/d*S* close to 0 indicates strong stabilizing selection. The mean d*N*/d*S* of each isolate compared to the last common ancestor was 0.66 ± 0.053 (s.d.), indicating either a weak trend in the direction of stabilizing selection since the last common ancestor of Typhi, or a combination of stabilizing selection in some genes and diversifying selection in others. As there is little evidence of diversifying selection in any Typhi genes (see below and Supplementary Table 3a online), weak stabilizing selection is most likely. The weakness of the signal for stabilizing selection observed here may be due to too little time for selection to act, and/or genetic drift due to low effective population size. It has been previously shown that in closely related bacteria, the reciprocal of d*N*/d*S*, or 1/(d*N*/d*S*), is related to time12: simulations indicated that when population size was large, this relationship was linear, but when effective population size was small, genetic drift became more important and 1/(d*N*/d*S*) reached a plateau. The relationship of 1/(d*N*/d*S*) to time (measured by the number of intergenic SNPs) for the sequenced Typhi isolates was nonlinear (Fig. 2a). Intergenic SNPs serve as an approximation of time, as they are less likely to be under purifying selection than SNPs in coding regions. However, intergenic SNPs may have regulatory or other functions that may be under selection, so as an alternative measure, we also calculated d*N*/d*S* among SNPs of different relative ages, which confirmed the same trajectory (Fig. 2b). In the light of the previously described model12, these patterns are suggestive of genetic drift in Typhi due to a small effective population size, which seems likely, as Typhi has no known reservoir outside of humans. A small effective population size ($N_e = 2.3 \times 10^5$ - 1.0×10^6) has been calculated previously using Bayesian skyline plots based on 82 SNPs in 105 Typhi isolates5.

Potential signals of selection

We found very little evidence of adaptive selection in Typhi genes, which would be represented by an overabundance of nonsynonymous SNPs or independent changes in the same or nearby amino acid residues. We found that 72% of genes contained no SNPs and that the distribution of SNPs per gene followed a Poisson distribution in the range of 0-6 SNPs per gene (Fig. 3). However, there were a few exceptions (Supplementary Table 3a). Three genes (*yehU, tviE* and STY2875) contained more than six SNPs, which deviates from the Poisson model. STY2875 is an exceptionally large gene (3,625 bp compared to the genome mean of 910 bp), which may account for the high number of SNPs. However, *yehU* and *tviE* are small (562-579 bp) and thus the high number of SNPs may be evidence of diversifying selection in these genes, the second of which is encoded within SPI7 and is involved with Vi synthesis4. Ten SNPs did not fit the phylogenetic tree, which may indicate either recombination or convergent evolution, whereby the same mutation arose independently in different lineages. If the latter explanation is true, it would suggest the possibility of adaptive selection at these sites, which include nonsynonymous SNPs in three genes encoding membrane proteins (STY1204, *yadG* and *tsaC*) and two nonsynonymous SNPs in *gyrA* that are known to increase resistance to fluoroquinolones, a class of antibiotics used to treat typhoid fever1,5,13. Fifteen genes contained clusters of nonsynonymous SNPs, whereby two residues within five amino acids were mutated, which may indicate adaptive selection in localized regions of the encoded protein (Supplementary Table 3a).

Of the 26 genes in which we detected potential signals of adaptive selection, half encode proteins that are surface-exposed, exported or secreted, or affect synthesis of such proteins (Supplementary Table 3a). These weak signals may reflect selective pressures stemming from interactions with the human host14, including selection for more virulent mutants or those with antigenic variants that better evade the human immune system. The genes identified here as potentially under selection warrant further investigation, illustrating the value of this approach, which could potentially be adapted to genetic association studies in pathogenic bacteria similar to those carried out routinely in eukaryotes15. However, most genes whose products are secreted or are surface-exposed showed no evidence of adaptive evolution. For example, with the exception of *sifA*, which encodes a SPI1 effector protein (Supplementary Table 3a), no other genes encoding known secreted effector proteins showed evidence of immune selection.

Recombination

Other than the ten SNPs that do not fit the phylogenetic tree, which are potentially a result of convergent evolution, we found no evidence of recombination between Typhi isolates and very little evidence of recombination with other bacteria (see Methods, Supplementary Note and Supplementary Table 1c online). Imports from *S. enterica* serovar Typhimurium (Typhimurium) have been reported in two Typhi isolates including 404ty5, but reanalysis of the history of the isolates affected revealed that these were introduced in the laboratory during the production of *aroA*-knockout mutants and do not represent wild-type variation. Large-scale recombination has been proposed between Typhi and the human-restricted *S. enterica* serovar Paratyphi A16. However, this most likely occurred before the evolution of the common ancestor of extant Typhi, which seems to be genetically isolated.

IS elements, phage, pathogenicity islands and plasmids

As 454 reads were long enough to be assembled, DNA insertion events could be identified among 454-sequenced Typhi isolates and confirmed by PCR and capillary sequencing (see Methods). Known *IS*1 insertions in the CT18 genome were not present at the same sites in

any other isolates, although we detected an *IS*1 element at a different genomic site within H58 isolates (see Supplementary Note).

CT18 harbors seven well-defined prophage-like elements3,17, and while all sequenced isolates showed conservation in most of these, some isolates harbored new phages. Figure 1 shows the occurrence of phage insertion events in the phylogenetic tree, and the number of insertion events occurring in each relative age group is shown in Table 2. The new phages are discussed further in the Supplementary Note.

We also observed variation in the 6-kbp genomic island SPI15 (ref. 18). This region includes an integrase gene adjacent to four hypothetical genes and was inserted within tRNA-Gly, generating direct flanking repeats. The region seemed to exist in three forms among the sequenced Typhi: (i) CT18; (ii) J185SM, 404ty and E03-4983; and (iii) all other isolates. In each case, the insertion site and direct repeats were identical, but three distinct but related alleles were present for the integrase gene (95% amino acid identity between forms i and ii, 70% between all three forms). Each of the three forms contained a unique set of cargo genes. The function of these genes is unknown, with no matches to known protein domains in the Pfam database (accessed July 3, 2008). These genes merit further investigation because of their potential contribution to virulence.

Plasmids detected in seven of the sequenced Typhi isolates (see Methods) fell into three classes (Table 1 and Fig. 1), which are discussed in more detail in the Supplementary Note.

Genomic deletions

Genomic insertions were rare in the sequenced isolates, but deletions were twice as common and more conserved (Table 2). Note that in many comparative studies, insertions and deletions are indistinguishable, but we were able to separate these events by using the rooted phylogenetic tree. The deletions range in size from 60 bp to 6.5 kbp, and some correspond to variant regions previously identified using DNA microarrays19 (Supplementary Table 3b). Most of the deleted regions include protein-coding sequences, resulting in partial or total deletion of 42 Typhi genes.

In addition, SPI7, which harbors genes required for synthesis of the polysaccharide Vi capsule4, was missing from 404ty and 150(98)S. The isolate E98-3139 seemed to be a mixed population in regards to SPI7, as its coverage in both 454 and Solexa reads was ~25% that of genomic coverage (Supplementary Fig. 3 online). Note that the low mapping coverage in this region is most likely due to deletion of SPI7 rather than replacement with a similar island, as deletion is known to occur during culture20,21 and no alternative island could be assembled from 454 reads. No other SPIs were deleted from the sequenced Typhi, indicating that they are relatively stable in the genome (although we observed three variants of the 6-kbp SPI15, as described above).

Ongoing functional gene loss

In addition to identifying 42 genes affected by deletion events, we detected 55 nonsense SNPs that had occurred since the last common ancestor of Typhi. These introduce stop codons into protein-coding genes, thereby cutting short translation. Read-through of stop codons has been reported22; however, the described mechanism applies to only two of the nonsense SNPs we detected. There was evidence of selection against nonsense SNPs, with a lower rate of occurrence than nonsynonymous SNPs. Nevertheless, many nonsense SNPs were fixed, making up 2.9% of SNPs in the intermediate and oldest age groups (Table 2).

CT18 and Ty2 each contain ~200 pseudogenes3,8, defined as genes that are putatively inactivated by mutations including nonsense SNPs, frameshifts and truncation by deletion or

rearrangement. This constitutes 4.5% of Typhi genes, much higher than the frequency in Typhimurium (0.9%) or *Escherichia coli* K12 (0.7%). High pseudogene frequencies are associated with host restriction in a variety of bacteria23-26, presumably as certain genes required for infection in a broad range of hosts are not required in the preferred host. This is potentially also attributable to high rates of mutation fixation resulting from accelerated genetic drift caused by evolutionary bottlenecks associated with host adaptation.

By mapping the deletions and nonsense SNPs to the phylogenetic tree, we found that 92 new pseudogenes have accumulated among the sequenced Typhi isolates since their last common ancestor (Supplementary Table 3c), which itself harbored ~180 pseudogenes3,8. Many of these genes fall into gene categories (metabolism, cobalamin utilization, peptide or sugar transport, fimbriae) previously associated with pseudogenes in host-restricted pathogens23 (Supplementary Table 3c). Figure 4 shows the rate of accumulation of inactivating mutations in each branch of the phylogenetic tree. Nearly all of these genes showed evidence of expression in Typhi according to microarray data accessible at the NCBI GEO database (Supplementary Methods and Supplementary Table 3c), thus most of the nonsense and deletion mutations we observed probably result in true inactivation of previously functional genes. Because the losses have occurred independently in different lineages, Typhi isolates at different points in the phylogenetic tree have varying complements of functional genes and may have different pathogenic potential. This may contribute to the differences observed in clinical manifestations of typhoid fever in different regions1. Of note, different lineages show variation in the relative rates of accumulation of SNPs and inactivating mutations (line slopes in Figure 4). This may be attributable to variation in mutation rates or different selective pressures for or against pseudogene formation in particular lineages.

As only 3% of possible SNPs in the Typhi genome are nonsense SNPs, we expect only 1-2 false nonsense SNP calls overall (3% of the estimated total of 53 false SNP calls). This constitutes ~2% of genes inactivated by nonsense or deletion mutation and thus would make little difference to conclusions regarding the continuous accumulation of pseudogenes. In addition, we did not attempt to analyze frameshift mutations, as single-base insertions or deletions are currently difficult to detect reliably from 454 and Solexa sequence data. However, most of the genes identified as differentially inactivated between CT18 and Ty2 were due to frameshift mutations (20 frameshifts versus 4 nonsense SNPs and 2 deletions)3,8, thus we hypothesize that many more pseudogenes may have accumulated in the Typhi population than those caused by nonsense SNPs or deletions. Therefore, although our analysis demonstrates that gene loss is ongoing in Typhi, we most likely underestimate the extent of this phenomenon.

DISCUSSION

Few whole-genome intraspecies comparisons of this scale exist for pathogenic bacteria27,28 and none at this level of subspecies resolution. In addition, the choice of isolates for sequencing is usually driven by clinical phenotype or simply availability, rather than unbiased sampling from reliable phylogenies. However, isolate selection is critically important for comparative analysis, which can only uncover mutations that differ between the sampled isolates. Sampling from one part of the phylogenetic tree will overlook much of the variation present in the population and collapse all isolates outside the sequenced subpopulation into a single type10. For example, when SNPs detected between CT18 and Ty2 were typed in a larger Typhi collection29, most isolates were assigned to the same genotype even though they were probably far more variable than the scheme suggests. By sequencing isolates from major nodes in the previously defined phylogenetic tree (Supplementary Fig. 1), we expect to have captured much of the variation present in the

Typhi population. We also anticipate that the SNPs we have detected among these sequences will serve as genotypic markers providing phylogenetic information at high resolution in future genotyping studies.

Our whole-genome analysis supports the proposals of small population size and genetic drift in Typhi. Although we detected signals of selection in *gyrA*, we did not detect signals of the same magnitude in other Typhi genes, suggesting that this level of selection is exceptional in Typhi. Furthermore, our whole-genome comparisons provide the opportunity to gain broad insight into the spectrum of genetic variation in Typhi, including SNPs, insertions, deletions and recombinations as well as plasmid and phage content (although we did not analyze insertion or deletion of single nucleotides). The patterns of genome-wide variation we detected demonstrate that pseudogene formation is ongoing in Typhi (Fig. 4) and support the hypothesis that evolution in this host-restricted pathogen is dominated by genetic drift and loss of gene function rather than by diversifying selection or gain of function through point mutation, recombination or acquisition of new sequences. Although gain of function seems to be rare, it may be occurring in a few genes through point mutations.

The lack of evidence for adaptive selection in general is in contrast with the known adaptive selection for mutations in *gyrA* associated with fluoroquinolone resistance. We detected the signal of selection in *gyrA* as clustered, homoplasic nonsynonymous SNPs in neighboring codons 83 and 87. Three other genes contained homoplasic nonsynonymous SNPs, one of which (*yadG*) encodes the membrane component of an efflux protein in *E. coli*30 and may therefore be associated with antibiotic resistance in Typhi (efflux proteins can act as pumps to remove antibiotics from the bacterial cell31). However, no genes besides *gyrA* contained multiple homoplasic SNPs, and few contained multiple nonsynonymous SNPs at all, consistent with the hypothesis of genetic drift in the Typhi genome. The adaptive mutations evident in the *gyrA* gene highlight the strong selective pressure on the Typhi genome associated with antibiotic use in the human population. This is not particularly surprising, as the fitness advantage associated with increased antibiotic resistance is likely to be very strong. However, the lack of similar evidence for other adaptive mutations suggests that Typhi is under relatively little selective pressure from its host or the environment in general.

The limited evidence of selection in Typhi gene sequences is particularly notable when compared to patterns observed among other human bacterial pathogens, which show a variety of mechanisms for antigenic variation. For example, antigenic variation is achieved by extensive recombination in the *Helicobacter pylori* and *Chlamydia trachomatis* populations32,33, whereas in *Mycobacterium tuberculosis*, antigenic variation is associated with duplication and diversification of antigen-associated gene families34. In contrast, only 3 Typhi genes contained more than 6 SNPs, and 16 genes contained independent nonsynonymous SNPs in the same or neighboring amino acids (see Supplementary Table 3a). Although these may represent cases of antigenic variation, the level of variation is low, with most of the SNPs unique to a single haplotype and therefore most haplotypes sharing identical sequences. Similarly, although there was some evidence of import of small fragments of non-Typhi sequences (see Supplementary Table 1c), the only indication of possible recombination between Typhi isolates were ten SNPs that do not fit the phylogenetic tree (Supplementary Table 3a), which could equally be due to convergent evolution. The sparsity of direct sequence evidence for antigenic variation in Typhi suggests that this pathogen is not under strong selective pressures from the human immune system and may interact with its host in a different way, possibly favoring immune evasion and localization to immune privileged sites. However, it cannot be ruled out that Typhi may possess as yet unidentified mechanisms of generating antigenic diversity or that phage genes, which were excluded from SNP analysis in this study, may have a role.

It has long been suspected that human carriers provide the main reservoir driving the transmission of Typhi35,36. The bacterium is relatively difficult to isolate from water and the environment even in endemic regions37,38, and it is generally believed that Typhi has a limited survival time outside the human host39. If human carriers provide the main persistent reservoir for Typhi, this could account for the patterns of genetic drift and lack of recombination or gene acquisition we observed, as the human reservoir is likely to be small and physiologically isolated35,36. Furthermore, adaptive mutations arising during symptomatic typhoid infections may have no fitness advantage in the carrier state and may therefore not persist in the long-term Typhi population. All nodes of the Typhi phylogenetic tree shown in Supplementary Figure 1 were detected among a set of approximately 450 extant Typhi isolates5, suggesting that the Typhi population is not shaped by clonal replacement. These patterns are well described by the source-sink model of evolutionary dynamics, which distinguishes permanent ‘source’ and transient ‘sink’ populations and predicts that adaptive mutations arising in the sink (individuals with typhoid) may be short lived in the population if they provide no fitness benefit in the long-term source (carriers of Typhi)40. Similar dynamics may be occurring in other human-restricted bacterial populations.

An understanding of the evolutionary dynamics of the Typhi population has important implications for the control of typhoid. The SNP typing of individual Typhi isolates into distinct genotypes may lead to improved methods for tracking the spread of Typhi between human hosts11. Vaccination may be a crucial long-term strategy for disease control, as it could not only reduce the level of typhoid infections but also the level of asymptomatic Typhi carriage in the population, a key reservoir of typhoid infections.

METHODS

Bacterial strains

Details of Typhi isolates used in this study are provided in Table 1. Isolates were provided by the Oxford University Clinical Research Unit (CT18, J185SM, AG3); B. Holmes at the National Collection of Type Cultures (M223); the Wellcome Trust Sanger Institute (404ty, Ty2); and F.-X.W. (all other isolates).

DNA sequencing

We pelleted bacterial cells by centrifugation and prepared DNA using the Wizard Genomic DNA Kit (Promega) according to the manufacturer’s instructions. We sequenced eight Typhi isolates using a 454 Life Sciences GS-20 sequencer (Roche), and an additional two isolates (M223, E02-1180) using the 454 Life Sciences GS-FLX sequencer (Roche). Twelve isolates were sequenced using the Illumina/Solexa Genome Analyzer System according to the manufacturer’s specifications. In all cases, we generated singleend reads. Two isolates, E02-2759 and ISP-04-06979, were each sequenced over seven Solexa lanes during protocol optimization and thus have much higher coverage than other isolates, which were sequenced in one Solexa lane each.

We used Sanger sequencing of PCR products to confirm insertion and deletion sites. Primers used for PCR and sequencing are provided in Supplementary Table 3d. PCR was done in a 25 μl volume using PCR Supermix Taq Polymerase (Invitrogen) and cycled on an MJ Research thermal cycler. Products were checked on a 0.8% agarose gel and purified using QIAquick PCR Purification Kit (QIAGEN).

Europe PMC Funders Author Manuscripts

Plasmid identification

In order to verify the presence and size of plasmids within Typhi isolates, we prepared plasmid DNA from Typhi isolates, as described in Supplementary Methods. All plasmids detected in this way were represented in the sequence data for their host isolates and were identified by mapping to known plasmid sequences (using blastn for 454 contigs and Maq v0.6.0 for Solexa reads).

SNP detection from sequence data

We assembled 454 reads *de novo* into contigs (that is, without reference to any other sequence) using newbler (v1.1, Roche). We used MUMmer (v3.19, nucmer algorithm)41 to align contigs to the finished CT18 sequence and to generate primary SNP calls. Solexa reads were too short to be assembled effectively using current software and thus were mapped directly to the CT18 reference sequence using Maq v0.6.0, which was also used to generate primary SNP calls. We filtered SNP calls according to quality criteria determined by comparison of data from multiple sequencing platforms, as described in Supplementary Methods. We combined filtered SNP calls into a single list of SNP loci and determined the allele at each locus in each of the 19 Typhi sequences and additional *S. enterica* serovars (using fasta3 search for 454 contigs or finished sequences, and Maq consensus base calls for Solexa data). This allowed recovery of some SNPs that were initially rejected in one isolate because of low confidence but detected with high confidence in a second isolate. Nonsense SNPs were verified by manually inspecting multiple alignments of all 454 and Solexa reads mapping to each nonsense SNP locus.

Estimation of sensitivity and specificity of SNP calls

We estimated SNP detection accuracy and sensitivity for 454 and Solexa by comparing results from three isolates sequenced using both platforms, as described in Supplementary Methods. Additional estimates for Solexa data were determined by comparing Solexa data from CT18 and Ty2 to the published sequences (see Supplementary Methods).

Phylogenetic analysis

SNPs lying within recombined regions (see below) or within repeat regions were excluded from analysis, leaving 1,964 SNP calls. We checked alleles against an independent whole-genome multiple alignment of all 454 and published Typhi sequences generated using Kodon (Applied Maths). Alleles could be confirmed in all nineteen Typhi isolates for 1,787 (90%) SNPs. These support a single maximum parsimony tree, determined using the mix algorithm in the PHYLIP package (Fig. 1), consistent with the reference phylogenetic tree (Supplementary Fig. 1).

dN/dS calculations

We calculated dN/dS according to the formula (N/n)/(S/s), where N = sum of nonsynonymous SNPs, n = nonsynonymous sites in nonrepetitive protein-coding sequences, S = sum of synonymous SNPs, s = synonymous sites in nonrepetitive protein-coding sequences. The mean dN/dS since the last common ancestor was calculated by weighting dN/dS by 1/2 for H59 isolates, 1/7 for H58 isolates and 1 for all other isolates, so that each haplotype is represented equally. The error reported (0.053) is 1 s.d. of this weighted mean.

Detection of recombination events

We checked SNP calls from each Typhi isolate for SNP clusters (defined as >3 SNPs within 1,000 bp) and searched these regions against the EMBL database using blastn in order to identify potential recombination events (Supplementary Note).

Europe PMC Funders Author Manuscripts

URLs

Maq, http://maq.sourceforge.net; PHYLIP, http://evolution.genetics.washington.edu/phylip.html; mapped assemblies of all 454 and Solexa datasets, http://www.sanger.ac.uk/Projects/S_typhi; 454 and Solexa reads data, ftp://ftp.era.ebi.ac.uk/ERA000001; Enteritidis strain PT4 sequence, http://www.sanger.ac.uk/Projects/Salmonella.

Accession codes

EBI Whole Genome Shotgun database: raw sequence data (454 *de novo* assembled contigs) are available with accession codes CAAQ-CAAZ. European Short Read Archive: Solexa and 454 reads, ERA000001. GenBank: Typhi strain CT18, AL513382; Typhi strain Ty2, AE014613; Typhimurium strain LT2, AE006468; Paratyphi A strain, CP000026; Choleraesuis strain SC-B67, AE017220; *E. coli* K12, NC_000913; Shigella flexneri 5 strain 8401, CP000266; pHCM1, AL513383; pHCM2, AL513384; pBSSB1, AM419040; pAKU1, AM412236.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Wellcome Trust. M.A. and C.J.M. are supported in Ireland by grant 05/FE1/B882 from the Scientific Foundation Ireland and C.J.M. was supported in Berlin by a Wellcome Trust grant to J. Farrar. We gratefully acknowledge the support of the Sanger Institute core sequencing and informatics groups. Isolates were provided by the Oxford University Clinical Research Unit (CT18, J185SM, AG3); B. Holmes at the National Collection of Type Cultures (M223); the Wellcome Trust Sanger Institute (404ty, Ty2); and F.-X.W. (all other isolates).

References

1. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid fever. N. Engl. J. Med. 2002; 347:1770–1782. [PubMed: 12456854]
2. Coburn B, Grassl GA, Finlay BB. *Salmonella*, the host and disease: a brief review. Immunol. Cell Biol. 2007; 85:112–118. [PubMed: 17146467]
3. Parkhill J, et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature. 2001; 413:848–852. [PubMed: 11677608]
4. Pickard D, et al. Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. J. Bacteriol. 2003; 185:5055–5065. [PubMed: 12923078]
5. Roumagnac P, et al. Evolutionary history of *Salmonella* Typhi. Science. 2006; 314:1301–1304. [PubMed: 17124322]
6. Chau TT, et al. Antimicrobial drug resistance of *Salmonella enterica* serovar Typhi in Asia and molecular mechanism of reduced susceptibility to the fluoroquinolones. Antimicrob. Agents Chemother. 2007; 51:4315–4323. [PubMed: 17908946]
7. Le TA, et al. Clonal expansion and microevolution of quinolone-resistant *Salmonella enterica* serotype Typhi in Vietnam from 1996 to 2004. J. Clin. Microbiol. 2007; 45:3485–3492. [PubMed: 17728470]
8. Deng W, et al. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. 2003; 185:2330–2337. [PubMed: 12644504]
9. Hall N. Advanced sequencing technologies and their wider impact in microbiology. J. Exp. Biol. 2007; 210:1518–1525. [PubMed: 17449817]
10. Pearson T, et al. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proc. Natl. Acad. Sci. USA. 2004; 101:13536–13541. [PubMed: 15347815]

11. Baker S, et al. High-throughput genotyping of *Salmonella* Typhi allows geographical assignment of haplotypes and pathotypes within an urban district of Jakarta, Indonesia. J. Clin. Microbiol. 2008; 46:1741–1746. [PubMed: 18322069]
12. Rocha EPC, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. 2006; 239:226–235. [PubMed: 16239014]
13. Turner AK, Nair S, Wain J. The acquisition of full fluoroquinolone resistance in *Salmonella* Typhi by accumulation of point mutations in the topoisomerase targets. J. Antimicrob. Chemother. 2006; 58:733–740. [PubMed: 16895934]
14. Haraga A, Ohlson MB, Miller SI. Salmonellae interplay with host cells. Nat. Rev. Microbiol. 2008; 6:53–66. [PubMed: 18026123]
15. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006; 14:353–355. [PubMed: 16782339]
16. Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res. 2007; 17:61–68. [PubMed: 17090663]
17. Thomson N, et al. The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. J. Mol. Biol. 2004; 339:279–300. [PubMed: 15136033]
18. Vernikos GS, Parkhill J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics. 2006; 22:2196–2203. [PubMed: 16837528]
19. Boyd EF, Porwollik S, Blackmer F, McClelland M. Differences in gene content among *Salmonella enterica* serovar Typhi isolates. J. Clin. Microbiol. 2003; 41:3823–3828. [PubMed: 12904395]
20. Nair S, et al. *Salmonella enterica* serovar Typhi strains from which SPI7, a 134-kilobase island with genes for Vi exopolysaccharide and other functions, has been deleted. J. Bacteriol. 2004; 186:3214–3223. [PubMed: 15126484]
21. Bueno SM, et al. Precise excision of the large pathogenicity island, SPI7, in *Salmonella enterica* serovar Typhi. J. Bacteriol. 2004; 186:3202–3213. [PubMed: 15126483]
22. Bertram G, Innes S, Minella O, Richardson JP, Stansfield I. Endless possibilities: translation termination and stop codon recognition. Microbiology. 2001; 147:255–269. [PubMed: 11158343]
23. Thomson NR, et al. The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. PLoS Genet. 2006; 2:e206. [PubMed: 17173484]
24. Parkhill J, et al. Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. 2003; 35:32–40. [PubMed: 12910271]
25. Cole ST, et al. Massive gene decay in the leprosy bacillus. Nature. 2001; 409:1007–1011. [PubMed: 11234002]
26. Andersson JO, Andersson SGE. Genome degradation is an ongoing process in *Rickettsia*. Mol. Biol. Evol. 1999; 16:1178–1191. [PubMed: 10486973]
27. Hiller NL, et al. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J. Bacteriol. 2007; 189:8186–8195. [PubMed: 17675389]
28. Tettelin H, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. Proc. Natl. Acad. Sci. USA. 2005; 102:13950–13955. [PubMed: 16172379]
29. Octavia S, Lan R. Single nucleotide polymorphism typing and genetic relationships of *Salmonella enterica* serovar Typhi isolates. J. Clin. Microbiol. 2007; 45:3795–3801. [PubMed: 17728466]
30. Saurin W, Hofnung M, Dassa E. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. J. Mol. Evol. 1999; 48:22–41. [PubMed: 9873074]
31. Webber MA, Piddock LJ. The importance of efflux pumps in bacterial antibiotic resistance. J. Antimicrob. Chemother. 2003; 51:9–11. [PubMed: 12493781]
32. Suerbaum S, et al. Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA. 1998; 95:12619–12624. [PubMed: 9770535]

33. Gomes JP, et al. Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots. Genome Res. 2007; 17:50–60. [PubMed: 17090662]
34. Gey Van Pittius NC, et al. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol. Biol. 2006; 6:95. [PubMed: 17105670]
35. Vaishnavi C, et al. Epidemiology of typhoid carriers among blood donors and patients with biliary, gastrointestinal and other related diseases. Microbiol. Immunol. 2005; 49:107–112. [PubMed: 15722595]
36. Levine MM, Black RE, Lanata C. Precise estimation of the number of chronic carriers of *Salmonella* typhi in Santiago, Chile, an endemic area. J. Infect. Dis. 1982; 146:724–726. [PubMed: 7142746]
37. Lewis MD, et al. Typhoid fever: a massive, single-point source, multidrug-resistant outbreak in Nepal. Clin. Infect. Dis. 2005; 40:554–561. [PubMed: 15712078]
38. Sears SD, Ferreccio C, Levine MM. The use of Moore swabs for isolation of *Salmonella* typhi from irrigation water in Santiago, Chile. J. Infect. Dis. 1984; 149:640–642. [PubMed: 6373964]
39. Cho JC, Kim SJ. Viable, but non-culturable, state of a green fluorescence protein-tagged environmental isolate of *Salmonella* typhi in groundwater and pond water. FEMS Microbiol. Lett. 1999; 170:257–264. [PubMed: 9919676]
40. Sokurenko EV, Gomulkiewicz R, Dykhuizen DE. Source-sink dynamics of virulence evolution. Nat. Rev. Microbiol. 2006; 4:548–555. [PubMed: 16778839]
41. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

Europe PMC Funders Author Manuscripts

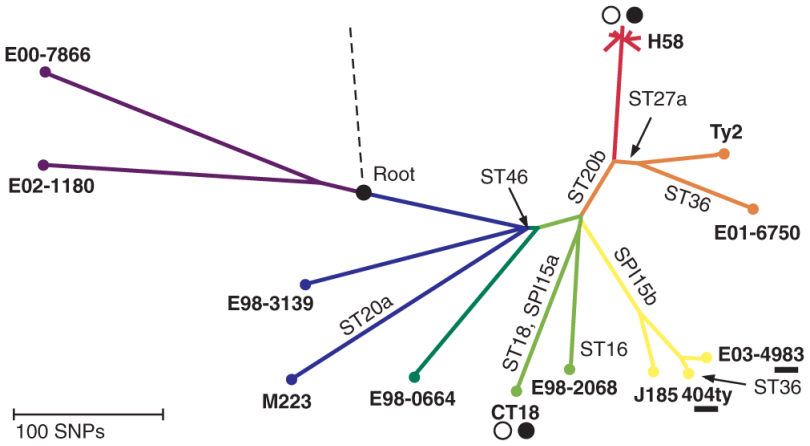

Figure 1.

Phylogenetic tree based on SNP data. Branch colors indicate different lineages of Typhi; branch lengths are measured in number of SNPs, scale as indicated. Central, small black circle indicates the ancestral root, dashed line represents *Salmonella* lineage; phage (ST) and SPI15 insertion events are shown along branches; plasmids detected in each isolate are indicated by filled circles (IncHI1 multidrug resistance plasmids), open circles (cryptic plasmid) and filled lines (linear plasmid carrying z66 flagella variant).

Europe PMC Funders Author Manuscripts

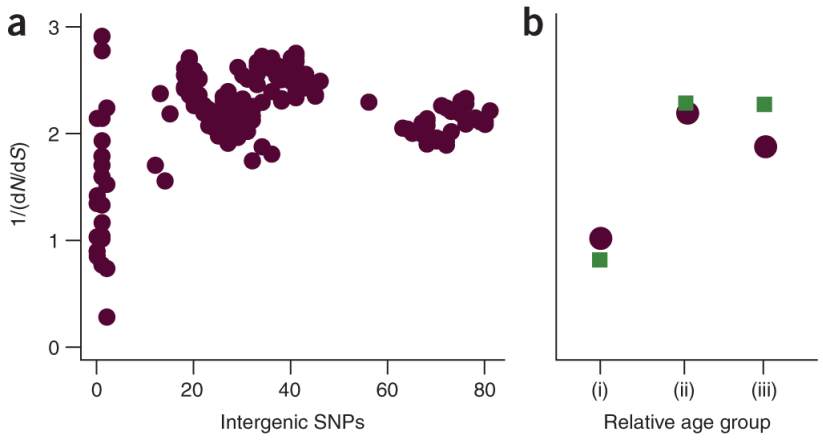

Figure 2.

Trajectory of d*N*/d*S* over time. *y* axis is the reciprocal of d*N*/d*S*, or 1/(d*N*/d*S*). (**a**) Pairwise d*N*/d*S* between 19 Typhi isolates. (**b**) dN/dS for SNPs in three relative age groups (i-iii, youngest-oldest), calculated from SNPs with complete allele data in 19 isolates (purple circles) and SNPs with incomplete allele data (green squares).

Europe PMC Funders Author Manuscripts

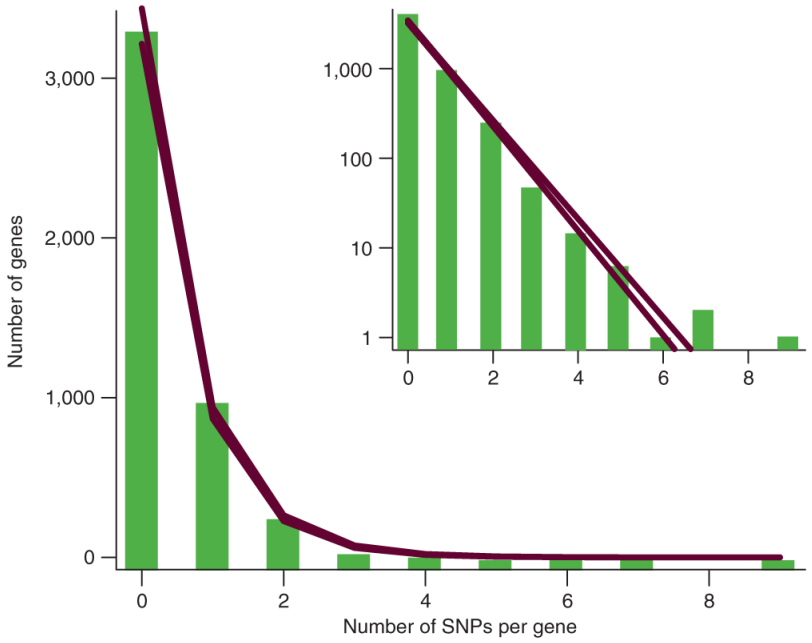

Figure 3.
Distribution of number of SNPs per gene. Lines indicate 95% confidence interval of mean predicted values under a Poisson distribution fitted to the data shown in green. Inset shows gene count on a log scale to better show deviation from the Poisson model at high numbers of SNPs per gene.

Europe PMC Funders Author Manuscripts

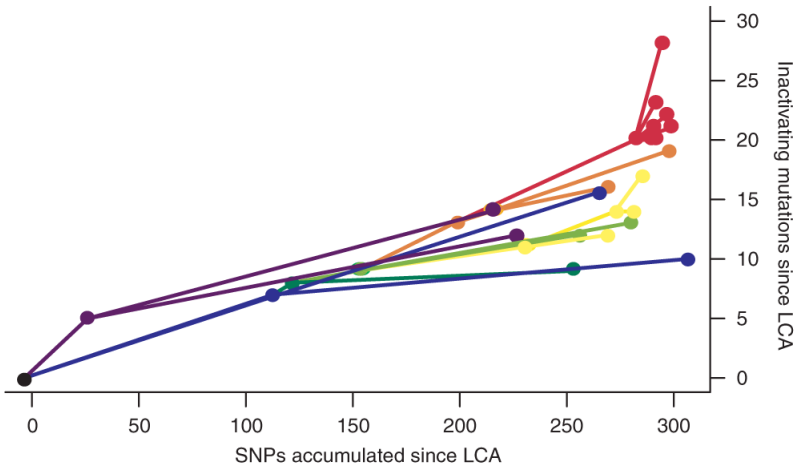

Figure 4.

Accumulation of gene-inactivating mutations in Typhi lineages. Points correspond to bifurcations in the phylogenetic tree in Figure 1; *y* axis shows the total number of genes inactivated by deletion or nonsense mutation up to that bifurcation. Each line represents the accumulation of mutations in a particular isolate; different lineages of Typhi are colored as in Figure 1. LCA, last common ancestor.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Typhi isolates sequenced in this study

Isolate	Country	Year	Haplotype	454 coverage	Solexa coverage	Plasmid
E00-7866	Morocco	2000	H46	10.5×	-	n.d.
E01-6750	Senegal	2001	H52	8.16×	-	n.d.
E02-1180	India	2002	H45	13.1×	-	n.d.
E98-0664	Kenya	1998	H55	10.8×	-	n.d.
E98-2068	Bangladesh	1998	H42	10.9×	-	n.d.
J185SM	Indonesia	1985	H85	13.5×	-	n.d.
M223	Unknown	1939	H8	11.1×	-	n.d.
404ty	Indonesia	1983	H59	8.49×	24.6×	pBSSB1
AG3	Vietnam	2004	H58	10.1×	13.1×	n.d.
E98-3139	Mexico	1998	H50	11.1×	5.40×	n.d.
150(98)S	Vietnam	1998	H63	-	8.60×	n.d.
8(04)N	Vietnam	2004	H58	-	13.1×	n.d.
CT18	Vietnam	1993	H1	-	9.80×	pHCM1, pHCM2
E02-2759	India	2002	H58	-	65.5×	pHCM2
E03-4983	Indonesia	2003	H59	-	7.42×	pBSSB1
E03-9804	Nepal	2003	H58	-	8.19×	pAKU1
ISP-03-07467	Morocco	2003	H58	-	7.87×	pAKU1
ISP-04-06979	Central Africa	2004	H58	-	72.9×	pAKU1
Ty2	Russia	1916	H10	-	8.60×	n.d.

Country and year of isolation are shown. Haplotypes correspond to those previously defined5. Coverage refers to oversampling in sequence data. n.d., none detected.

Europe PMC Funders Author Manuscripts

Table 2
Genetic variation detected in 19 Typhi genomes

	Intrahaplotype	Interhaplotype	Conserved	Total
Deletions	5	8	7	20
Phage insertions	n.a.	5	4	9
Plasmids	3	2	0	5
SNPs (complete allele data)	93	1,356	338	1,787
- Intergenic	6 (6.5%)	177 (13.1%)	44 (13.0%)	227
- Synonymous	21 (22.6%)	477 (35.2%)	106 (31.4%)	604
- Nonsynonymous	61 (65.6%)	663 (48.9%)	176 (52.1%)	900
- Nonsense	5 (5.4%)	39 (2.9%)	12 (3.6%)	56
- d*N*/d*S*	0.98	0.46	0.52	0.49
SNPs (incomplete allele data)	19	122	35	176
- Intergenic	4 (21.1%)	24 (19.7%)	6 (17.1%)	34
- Synonymous	3 (15.8%)	41 (33.6%)	12 (34.3%)	56
- Nonsynonymous	12 (63.2%)	57 (46.7%)	17 (48.6%)	86
- Nonsense	0 (0.0%)	0 (0.0%)	0 (0.0%)	0
- d*N*/d*S*	1.24	0.44	0.44	0.48

Frequencies of mutations according to age (recent, intrahaplotype; intermediate, interhaplotype; oldest, conserved). Percentages indicate the relative frequencies within each age group, separately for alleles that could be reliably determined for all isolates (complete allele data) and those that could not (incomplete allele data). n.a., not applicable