Received: 7 March 2019 | Revised: 21 July 2019 | Accepted: 26 August 2019

DOI: 10.1111/eva.12864

ORIGINAL ARTICLE

Evolutionary Applications Open Access WILEY

A novel taxon selection method, aimed at minimizing recombination, clarifies the discovery of a new sub-population of *Helicobacter pylori* from Australia

Binit Lamichhane[1] | Michael J. Wise[1,2] | Eng Guan Chua[1] | Barry J. Marshall[1] | Chin Yen Tay[1]

[1]Helicobacter pylori Research Laboratory, Marshall Centre for Infectious Disease Research and Training, School of Biomedical Sciences, University of Western Australia, Perth, WA, Australia

[2]Department of Computer Science and Software Engineering, University of Western Australia, Perth, WA, Australia

Correspondence
Michael J. Wise, Computer Science & Software Engineering, M002, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6008, Australia.
Email: Michael.Wise@uwa.edu.au

Funding information
This project was supported by the National Health and Medical Research Council (grant no. 572723) and the Vice Chancellor of the University of Western Australia.

Abstract

We present a novel method for taxon selection, the aim being to minimize problems arising from highly recombinant species such as *Helicobacter pylori*. *Helicobacter pylori* has accompanied modern-human migration out of Africa and is marked by a phylogeographic strain distribution, which has been exploited to add an extra layer of information about human migrations to that obtained from human sources. However, *H. pylori's* genome has high sequence heterogeneity combined with a very high rate of recombination, causing major allelic diversification across strains. On the other hand, recombination events that have become preserved in sub-populations are a useful source of phylogenetic information. This creates a potential problem in selecting representative strains for particular genetic or phylogeographic clusters and generally ameliorating the impact on analyses of extensive low-level recombination. To address this issue, we perform multiple population structure-based analyses on core genomes to select exemplar strains, called 'quintessents', which exhibit limited recombination. In essence, quintessent strains are representative of their specific phylogenetic clades and can be used to refine the current MLST concatenation-based population structure classification system. The use of quintessents reduces the noise due to local recombination events, while preserving recombination events that have become fixed in sub-populations. We illustrate the method with an analysis of core genome concatenations from 185 *H. pylori* strains, which reveals a recent speciation event resulting from the recombination of strains from phylogeographic clade hpSahul, carried by Aboriginal Australians, and hpEurope, carried by some of the people who arrived in Australia over the past 200 years. The signal is much clearer when based on quintessent strains, but absent from the analysis based on MLST concatenations.

KEYWORDS
core genomes, *Helicobacter pylori*, phylogeography, recombination, taxon selection

Binit Lamichhane and Michael J. Wise contributed equally to this work.

1 | INTRODUCTION

1.1 | *Helicobacter pylori* and its highly recombinant genome

Helicobacter pylori is a major gastric pathogen that infects about half of the human population in the world. The infection rate can be as high as 90% in developing countries, though it is usually less than 30% in developed countries (Covacci, Telford, Giudice, Parsonnet, & Rappuoli, 1999; Linz et al., 2007). The general prevalence of *H. pylori* in Australia has been reported as ranging from 15% to 30% (Lin, Lambert, Nicholson, Lukito, & Wahlqvist, 1998; Robertson, Cade, Savoia, & Clancy, 2003). However, the prevalence among Aboriginal Australians can reach as high as 76% (Windsor et al., 2005). *Helicobacter pylori* is believed to be transmitted between close family members via oral–oral, gastric–oral or faecal–oral routes, usually during childhood and usually from mother to children (Goh, Chan, Shiota, & Yamaoka, 2011). It has been coevolving with its human host for more than 120,000 years and migrated with us, which has led to the emergence of different phylogeographic genotypes over time. Therefore, *H. pylori* has also become a useful genetic marker, providing information about relationships between human ethnic populations and the human migration history (Linz et al., 2007; Tay et al., 2009). However, the sequence heterogeneity within *H. pylori* is very high (Tay et al., 2009), likely due to the lack of a proof-reading function in DNA polymerase I (Garcia-Ortiz et al., 2011) together with a very high recombination rate that facilitates the exchange of genes between genetically different isolates (Baltrus, Guillemin, & Phillips, 2007; Didelot et al., 2013).

1.2 | Studying *H. pylori* evolution and population structure

Multilocus sequence typing (MLST), based on concatenated fragments of 7 housekeeping genes, has long been used to reveal the evolutionary history of *H. pylori* and its correlation with the human Out-of-Africa migration hypothesis (Linz et al., 2007; Maixner et al., 2016; Moodley et al., 2009). On top of this, the program STRUCTURE, using Bayesian methods, has been widely used to deduce population structure based on MLST data. In particular, STRUCTURE has been used in many studies to determine the population structure of many human pathogens, including *H. pylori* (Falush et al., 2003; Maixner et al., 2016; Tay et al., 2009). To date, *H. pylori* has been classified into 7 distinct populations that are associated with particular geographic areas: hpAfrica2, hpAfrica1, hpNEAfrica, hpEurope, hpAsia2, hpEastAsia and hpSahul (Falush et al., 2003; Linz et al., 2007; Moodley et al., 2009). hpSahul is named after the ancient Sahul continent, that is mainland Australia, Tasmania and New Guinea, which were joined from 100 kya (100,000 years ago) to relatively recent times, some 31,000–37,000 years ago. hpSahul is only carried by Aboriginal Australians and is thought to have been split from the East Asia *H. pylori* population (Moodley et al., 2009) when Aboriginal Australians first migrated to Australia 65,000 years ago (Clarkson et al., 2017).

1.3 | Genomic and taxon selection issues when studying *H. pylori* and other species with highly recombinant genomes

In studying *H. pylori*, we became aware that previous studies of the population structure of *H. pylori* have faced unresolved methodological issues. Most studies based on *H. pylori* have, thus far, been based on MLST concatenations (Linz et al., 2007; Maixner et al., 2016; Moodley et al., 2009; Tay et al., 2009), though with the development of cost-effective, high-throughput whole-genome sequencing technology, it is possible to undertake phylogenetic studies using concatenations of core genes (Gressmann et al., 2005) or whole genomes (Kumar et al., 2015). However, even this may not give clear-cut information due to the highly recombinant nature of *H. pylori*. In particular, recombination events can be problematic for phylogenetic analyses as they break the assumption of clonal descent (Didelot & Falush, 2007). On the other hand, issues arising from the fact that different genes face different evolutionary pressures (Wise, 2013) can affect single-gene studies or studies based on concatenations of just a few genes.

In the phylogeographic literature based on *H. pylori*, there have been, in essence, two approaches to dealing with species whose genomes evidence a high level of recombination. The first approach has been to tacitly ignore the problem. This approach is evident in the numerous MLST gene fragment-based studies, but also in the whole-genome studies, which are discussed above. The other approach is to use applications such as ClonalFrameML (Didelot & Wilson, 2015) or Gubbins (Croucher et al., 2015) which are given, or compute, a phylogenetic tree and then remove parts of the input sequence multiple alignments which are not consonant with an assumption of clonal descent. Currently, both of these programs only apply to nucleotide sequence data.

For this study, we sequenced 177 stains and, together with 8 well-studied strains from the literature, then identified the core genomes of the total set of 185 *H. pylori* strains. Core genomes provide us with much better resolution and allow us to delimit strain diversity more precisely rather than have been possible with MLST-based studies. However, core genomes can still exhibit considerable levels of recombination. To deal with this, using the core genomes, we identified exemplar strains from each *H. pylori* sub-population that has limited, or no, recombination with the other sub-populations. We have called these exemplars quintessents (to connote a set of primary objects, from which secondary objects are obtained by combination). In other words, in this study we have taken a different approach than ClonalFrameML and Gubbins, viewing the problem of limiting recombination as a taxon selection issue. There has been a considerable history of debate about how best to select taxa for phylogenetic analyses (see the review Nabhan and Sarkar (2012), for example). What is proposed here is that freedom from recombination could be one criterion.

2 | MATERIALS AND METHODS

2.1 | Selecting *H. pylori* strains and obtaining DNA

One hundred and fifty-five *H. pylori* strains were isolated from patients attending Sir Charles Gairdner Hospital (Perth, Western Australia) for treatment of antibiotic-resistant *H. pylori* infection. All were informed about the nature of study, and written consent was obtained from those who wished to proceed. The protocols were approved by the hospital's Human Research Ethics Committee. A further 22 strains, classed as hpSahul by MLST, originally obtained during the Windsor et al. (2005) study, with MLST sequences reported in Moodley et al. (2009), were fully sequenced in this study. Of these, strain HPJ023, sequenced in this study, was also studied by Montano et al., who named it ausabrJ05 (Montano et al., 2015).

The 177 strains were grown on 5% horse blood agar (HBA) plates as previously described (Lu et al., 2014). Genomic DNA was extracted from each *H. pylori* strain using DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instruction. The quality of the DNA samples was checked using NanoDrop 2000 Spectrophotometer and Qubit Fluorometer. The samples were then stored at −20°C until they were analysed.

2.2 | DNA sequencing, genome assembly and determination of the core genome

1 ng of bacterial genomic DNA was used for genomic library preparation using Nextera XT protocol (Ver. September 2014). The libraries were then subjected to 250 bp paired-end sequencing on a MiSeq Sequencer running version 1.1.1 MiSeq Control Software (Illumina Inc.). The de novo assembly of raw reads was performed using St. Petersburg genome assembler (SPAdes, Ver.3.8.2) (Bankevich et al., 2012). Contigs with coverage less than 10 times and length less than 500 were filtered. In addition to the above 177 *H. pylori* strains, eight well-studied, publicly available complete genome sequences were also included in the analysis. FASTA format files of the strains 26695, G27, Sahul64, India7, Pecan4, SouthAfrica7, J99 and F57 sequences were obtained from NCBI via the genome browser (http://www.ncbi.nlm.nih.gov/genome/browse/). To ensure uniformity of approach, all 185 genomes, including the eight reference sequences, were annotated using Prokka (Ver. 1.11) (Seemann, 2014). Details of the respective genomes for the strains sequenced for this project can be found in Table S1; corresponding details for the strains from NCBI can be found in Table S2. The core genome of all 185 *H. pylori* strains was determined at the protein level using a best reciprocal BLAST heuristic implemented in the program Proteinortho v.5.1 (settings: -e = 1e−05, -p = blastp, -id = 50, -cov = 80, -conn = 0.1, -sim = 0.95) (Lechner et al., 2011). Genes present in 100% of the strains constituted the core genome, which consisted of 898 genes. Gene-by-gene alignments of single-copy orthologous core genes were performed using MAFFT version 7.271 (Katoh & Standley, 2013) with options –maxiterate = 1,000 and -localpair. The SNPs were extracted using SNP sites (Page et al., 2016).

2.3 | Determining population structures and the set of quintessents

The SNPs were subjected to STRUCTURE v.2.3.4 (Pritchard, Stephens, & Donnelly, 2000) analysis, which implements a Bayesian approach to deducing the population structure, based on an a priori fixed number of sub-populations. The Markov chain Monte Carlo (MCMC) simulation underpinning STRUCTURE was run for 100,000 iterations, following a burn-in of 100,000 iterations. A no-admixture population model was used, supported by a correlated frequency model for allele frequencies (Porras-Hurtado et al., 2013). To determine the number of sub-populations, K, STRUCTURE was run for K ranging from 4 to 12, and each run was repeated 12 times. Structure Harvester v0.6.94 (Earl & vonHoldt, 2012) was then used to determine the optimal value of K, which occurred for K = 11. For the K = 11 data set, a strain was considered to be quintessent, that is an exemplar of a particular sub-population, if it was assigned to that sub-population with probability of at least 0.75 in at least 60% of the runs. We tried a number of combinations of the probability and run-percentage cut-offs, and selected this particular combination because it maximized the number and size of clades while also minimizing recombination. With the combination of 0.75 probability in at least 60% of the runs, 93 strains were identified as being quintessents. The set of quintessent strains is noted with a '*' in Tables S1 and S2.

2.4 | Phylogenetic and recombination analysis

Four different sequence data sets were created for this study: nucleotide concatenations of the 898 core genes for each of the 185 strains (the set labelled all nt), protein concatenations of the corresponding 898 core proteins for each of the 185 strains (all aa), MLST concatenations for each of the 185 strains (mlst) and protein concatenations from 472 minimally recombinant genes from 93 quintessent strains (quint aa). The 472 minimally recombinant genes were those whose Phi statistics were greater than or equal to 0.1. (The Pairwise Homoplasy Index, Phi, is discussed below, together with the interpretation of the *p*-value threshold.) The MPI-based, genome-scale phylogenetic tree building application ExaML (Kozlov, Aberer, & Stamatakis, 2015) was used to create the trees for the different sequence data sets based on multiple sequence alignments created using Clustal Omega (Sievers et al., 2011). In each case, 500 bootstrapped trees were computed, with a gamma model for mutation rate heterogeneity across sites. A starting neighbour-joining tree was created using Molecular Evolutionary Genetics Analysis software version 7.0 (MEGA 7.0) (Kumar, Stecher, & Tamura, 2016) with default parameter settings. The best of the trees computed by ExaML was then annotated with bipartition data—effectively bootstrap percentages—using RAxML (Stamatakis, 2014), and the final trees were visualized using Figtree (v1.4.3) (https://github.com/rambaut/figtree/releases).

The extent of recombination was measured using the Pairwise Homoplasy Index, Phi (Bruen, Philippe, & Bryant, 2006). Phi is a *p*-value, related to the probability of rejecting the null hypothesis that there is no recombination in the set of aligned sequences being

tested in a sliding window. In our experiments, the Phi statistic was based on a window size of 20 for amino acids or 60 for nucleotides, with a permutation test used to compute statistical significance. The window size was reduced from 100 to prevent the metric becoming saturated, and every gene in this very recombinogenic organism thus appearing to be recombinant.

Finally, to assess the phylogenetic tree distance between the new hpEuropeSahul clade (see below) and the nearest hpEurope or hpSahul taxon for each of the core genes, PhyML (Guindon & Gascuel, 2003) was used to create an unbootstrapped tree, but with topology and branch-length optimization. A stand-alone Python program was then used to compute the shortest distance from each hpEuropeSahul taxon to the nearest hpEurope taxon, and also to the nearest hpSahul taxon. The code for the program can be found at https://github.com/mw263/clade to clade distance.

2.5 | Comparison of quintessents with ClonalFrameML

Comparison of quintessent selection with existing methods, for example ClonalFrameML, is complicated by the fact the methods are very different: ClonalFrameML only works with nucleotide sequences (genomes or gene concatenations), while optimal results for the quintessent method are obtained with amino acid data, though the method also works for nucleotide data. ClonalFrameML keeps the complete set of input sequences, but the length of the sequences has been reduced, while the quintessent method returns a reduced set of sequences, but with the sequence lengths unchanged. Therefore, in order to compare like with like, the data set all nt (see above) was input into ClonalFrameML to create the data set all nt cf, while the 93 strains, identified as quintessent, were taken from all nt to represent the quintessent data set. This data set was called quint nt. In other words, to the original suite of data sets described above: all nt, all aa, mlst and quint aa, were added all nt cf and quint nt. The starting tree created using MEGA 7.0 (described earlier) was the second input to ClonalFrameML.

As first comparison, cladistic information content, in the form of the dCITE metric (Wise, 2016), was computed for the two new data sets, all nt cf and quint nt, and compared with the correspondding data from the all nt data set. A second method for examining the two approaches was to compare the trees computed using ExaML (outlined above) for the six data sets—all against all—but focusing on all nt cf. The application TreeCmp (Bogdanowicz, Giaro, & Wróbel, 2012) was used to compare pairs of trees based on four metrics: Robinson–Foulds distance, Estabrook's quartet distance, Steel and Penny's path difference distance and the TreeCmp authors' own metric, matching split distance.

3 | RESULTS AND DISCUSSION

3.1 | General features and pan-genome of *H. pylori* genomes

We sequenced 177 *H. pylori* genomes isolated from Australia, which represented all 7 known MLST-based STRUCTURE clades. Eight complete genomes obtained from NCBI belonging to different MLST STRUCTURE clades were also included in the analysis. The 177 new genomes were sequenced with at least 100× coverage for each strain and assembled into between 19 and 86 contigs. In our study, *H. pylori* genome size ranged between 1.53 and 1.74 Mb. The genomes had an average G + C content of 38% and were predicted to encode between 1,443 and 1,658 genes. Summaries of genomes sequenced in this study are presented in Table S1. Orthologous group analysis resulted in 906 genes found to be present in all 185 strains and constituted the core genome. Among them, 898 were single-copy orthologous genes with the concatenated length of 848 kbp. The number of accessory genes ranged from 495 to 710. The number of core genes found in this study is much lower than the 1,111 core genes found by Gressmann et al. (2005), 1,223 found by Fischer et al. (2010), the 1,226 genes found by Kumar et al. (2015), and 1,187 found by Kumar, Albert, Abkal, Siddique, and Ahmed (2017). Given that *H. pylori* is a panmictic species, the reduced core genome suggests that our samples represented a greater breadth of *H. pylori* strains. However, the fact that these are draft genomes may also have had an impact on size of the core genome.

3.2 | Discrepancy in phylogenetic analysis based on MLST versus core genomes

The phylogenetic tree created from the concatenated MLST genes demonstrated a clustering pattern similar to those reported by many previous studies (Achtman et al., 1999; Falush et al., 2003; Linz et al., 2007; Moodley et al., 2009). Briefly, strains were clustered into 7 major *H. pylori* population types (Figure 1, in which each strain is coloured by the highest percentage STRUCTURE group). By contrast, the phylogenetic analysis using concatenated 898 core genes revealed clustering that is only partially consistent with that found using MLST; among other differences, five hpSahul strains namely, HP01140, HP01316, HP03127, HP01193 and Sahul64, were grouped together with the hpEurope strains (Figure 2, based on nucleotide data, and Figure 3, based on the corresponding protein sequence concatenations), compared with the MLST-based tree (Figure 1), where these strains shared an ancestral node with hpSahul. In particular, based on MLST assignment, Sahul64 has previously been used as a reference genome to represent the hpSahul *H. pylori* population (Lu et al., 2014). However, in this study, it was among the five strains that were grouped with hpEurope. Interestingly, all of these five strains were isolated from unrelated Aboriginal Australians, suggesting these strains are the result of recombination between hpSahul and hpEurope *H. pylori* populations that are now being stably inherited. Discrepancies between the phylogenetic trees derived from MLST gene fragment concatenations and core genome concatenations have also been reported in previous studies (Gressmann et al., 2005; Munoz-Ramirez et al., 2017), but this is the first time it has been reported to the extent where strains are grouped in a totally unrelated cluster.

Whole-genome sequencing has broadened the opportunity to include more genetic information, which increases the resolution in

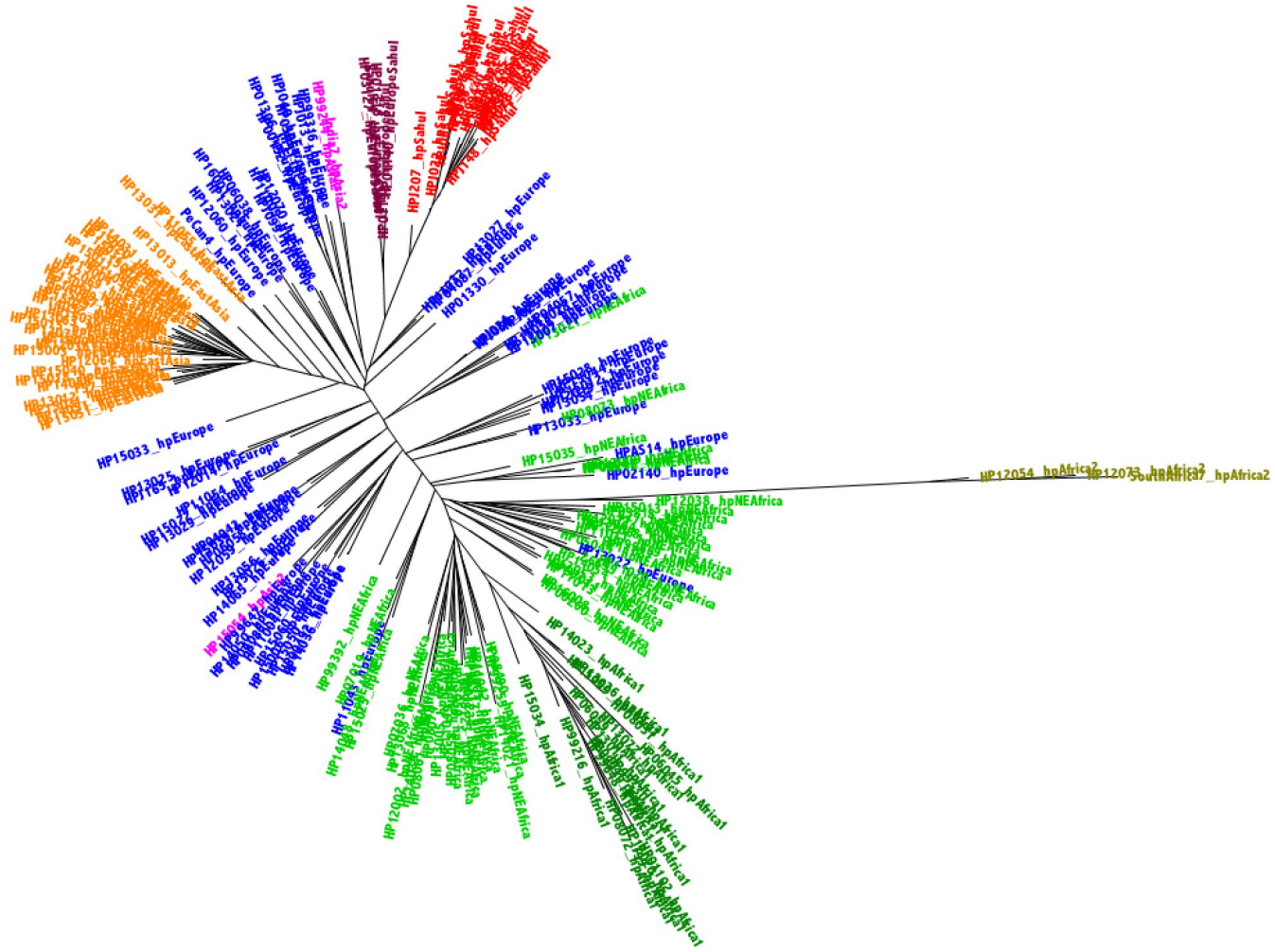

FIGURE 1 185 Strains MLST tree. A phylogenetic tree was created by ExaML, based on MLST gene fragment concatenation data from all 185 strains. A starting tree was created using Mega. Five hundred bootstrapped trees were computed, with the best (i.e., lowest absolute value log likelihood) shown here. The taxa have been labelled, both in the labels and by colour, according to the highest percentage STRUCTURE group, based on the best (lowest absolute value log likelihood) of 12 runs for k = 7 bins (based on suggestion from Moodley et al. (2009)). The clade colours are as follows: hpAfrica2 (olive), hpNEAfrica (bright green), hpAfrica1 (dark green), hpEurope (blue), hpAsia2 (pink), hpEastAsia (orange) and hpSahul (red). The sub-clade of hpSahul that we now discover is a new clade, hpEuropeSahul, is shown in maroon

population-based phylogenetic studies (Qin et al., 2016). Therefore, using a broader spectrum of genes should be preferred, rather than MLST concatenations of gene fragments, in order to increase the cladistic information (Wise, 2016) and thereby improve the resolution in phylogenetic analyses. However, use of whole-genome data is problematic in highly recombinant species such as *H. pylori*.

3.3 | Selection and testing of quintessents

The presence of a remarkable degree of genetic variability in *H. pylori* is driven by frequent recombination and a high mutation rate (Falush et al., 2001; Kersulyte, Chalkauskas, & Berg, 1999; Suerbaum & Josenhans, 2007; Suerbaum et al., 1998). As a result, *H. pylori* largely lacks a clonal structure—except due to founder effects—and has been characterized as 'weakly clonal' (Achtman et al., 1999). On the other hand, the high recombination rate and ability of *H. pylori* to undergo frequent mutation lead to only partial linkage disequilibrium between polymorphic loci, which can provide additional information for population genetic analysis (Didelot et al., 2013; Suerbaum & Josenhans, 2007). However, this may cause substantial problems in the selection of strains representing particular sub-populations for comparative genomic and phylogeographic studies.

In this study, to determine the set of strains that are exemplars of a particular sub-population—which we have called quintessents—we ran the program STRUCTURE on the 270,782 SNPs extracted from the core genomes using an admixture model, where the individuals have inherited some fraction of their genomes from ancestors in up to K = 11 sub-populations. Ninety-three strains were found to belong to a particular structure cluster with probability of at least 0.75 in 60% of the runs and were therefore considered to be quintessents. We believe it is necessary to distinguish between ancient recombinations that have been preserved in particular

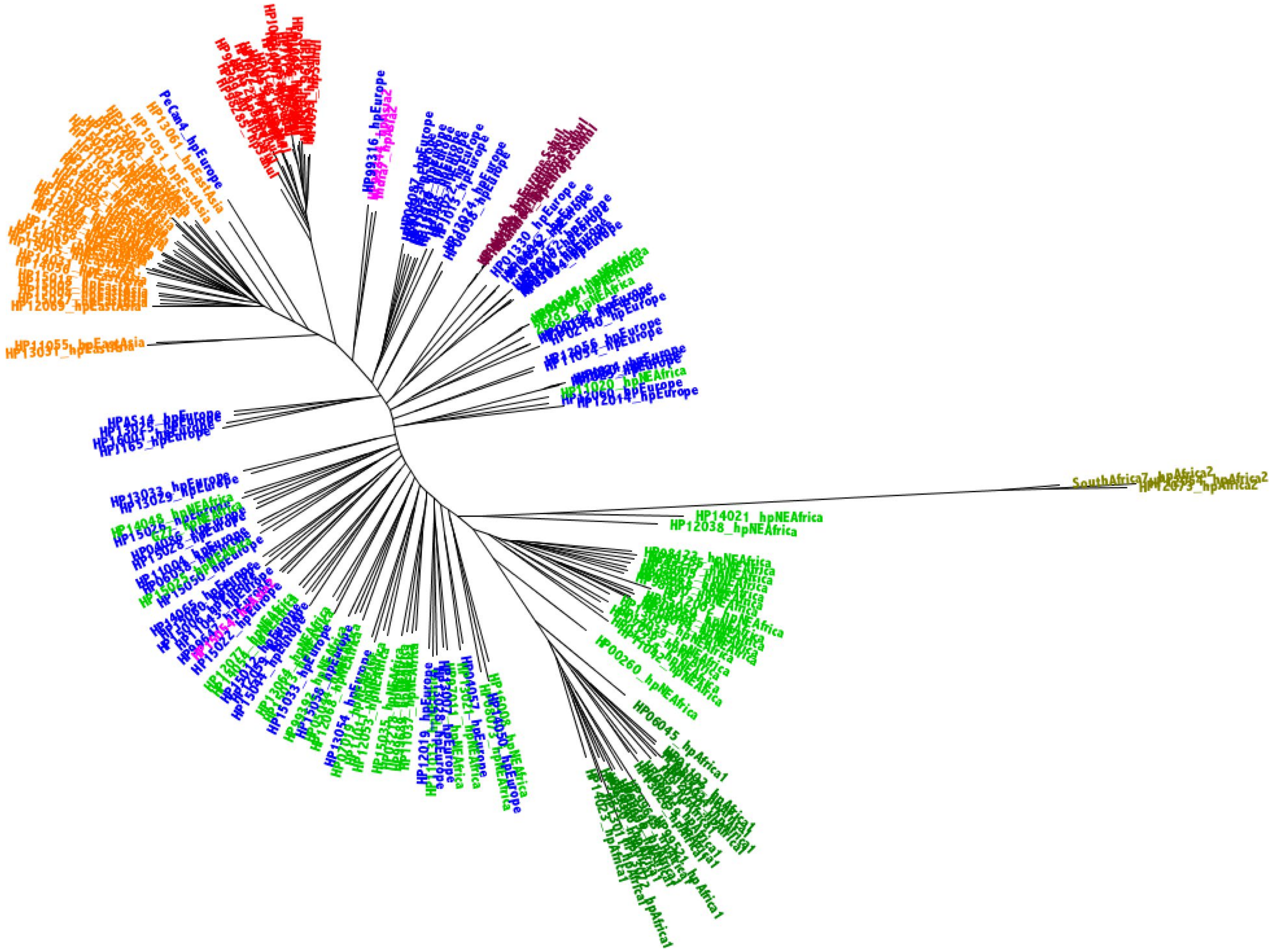

FIGURE 2 185 Strains core genome tree. A phylogenetic tree was created by ExaML, based on core genome concatenations of 898 genes from all 185 strains. A starting tree was created using Mega. Five hundred bootstrapped trees were computed, with the best (i.e., lowest absolute value log likelihood) shown here. The taxa have been labelled, both in the labels and by colour, according to the highest percentage STRUCTURE group, based on the best (lowest absolute value log likelihood) of 12 runs for k = 7 bins (based on suggestion from Moodley et al. (2009)). The clade colours are as follows: hpAfrica2 (olive), hpNEAfrica (bright green), hpAfrica1 (dark green), hpEurope (blue), hpAsia2 (pink), hpEastAsia (orange) and hpSahul (red). The sub-clade of hpSahul that we now discover is a new clade, hpEuropeSahul, is shown in maroon

sub-populations and local recent recombination noise. In this context, the significance of quintessents is that, with limited recombination evident, we presume the quintessents to be closer to the founder strains for their respective sub-populations. Viewed another way, quintessents are sub-populations of strains that have the same population structure.

To assess the impact of selecting quintessents strains on the extent of recombination, the Phi statistic was computed from multiple sequence alignments for each of the 898 core genes in the data sets: all nt, all aa and quint aa. The results are in Table 1. Each cell in Table 1 shows the number of genes/proteins (out of 898) that fail to reject the null hypothesis of nonrecombination (at the given Phi threshold). That is, they are presumed to be nonrecombinant.

It is clear from the data in Table 1 that simply moving from nucleotide to amino acid concatenations can significantly reduce the level of apparent recombination. For example, in the all nt data set, at the 0.1 Phi threshold, only 150 genes were found not to be recombinant (i.e., failing to reject the null hypothesis), compared with 439 of the corresponding proteins from the all aa data set. This is most likely due to the buffering provided by the redundancy in the amino acid codon translation table versus the input nucleotide sequences, particularly at the highly variable third codon positions.

Turning to the comparison of amino acid sequences from quintessent strains versus corresponding sequences from nonquintessent strains, use of quintessents further increased the number of nonrecombinant sequences to 482, which is a statistically significant increase (p = .0018, on a binomial distribution statistic). This suggests that much of the recombination evident at the nucleotide sequence level may be relatively recent recombination noise, and the move to amino acid sequences—and, particularly, amino acid sequences from quintessent strains—brings us closer to ancient recombinations

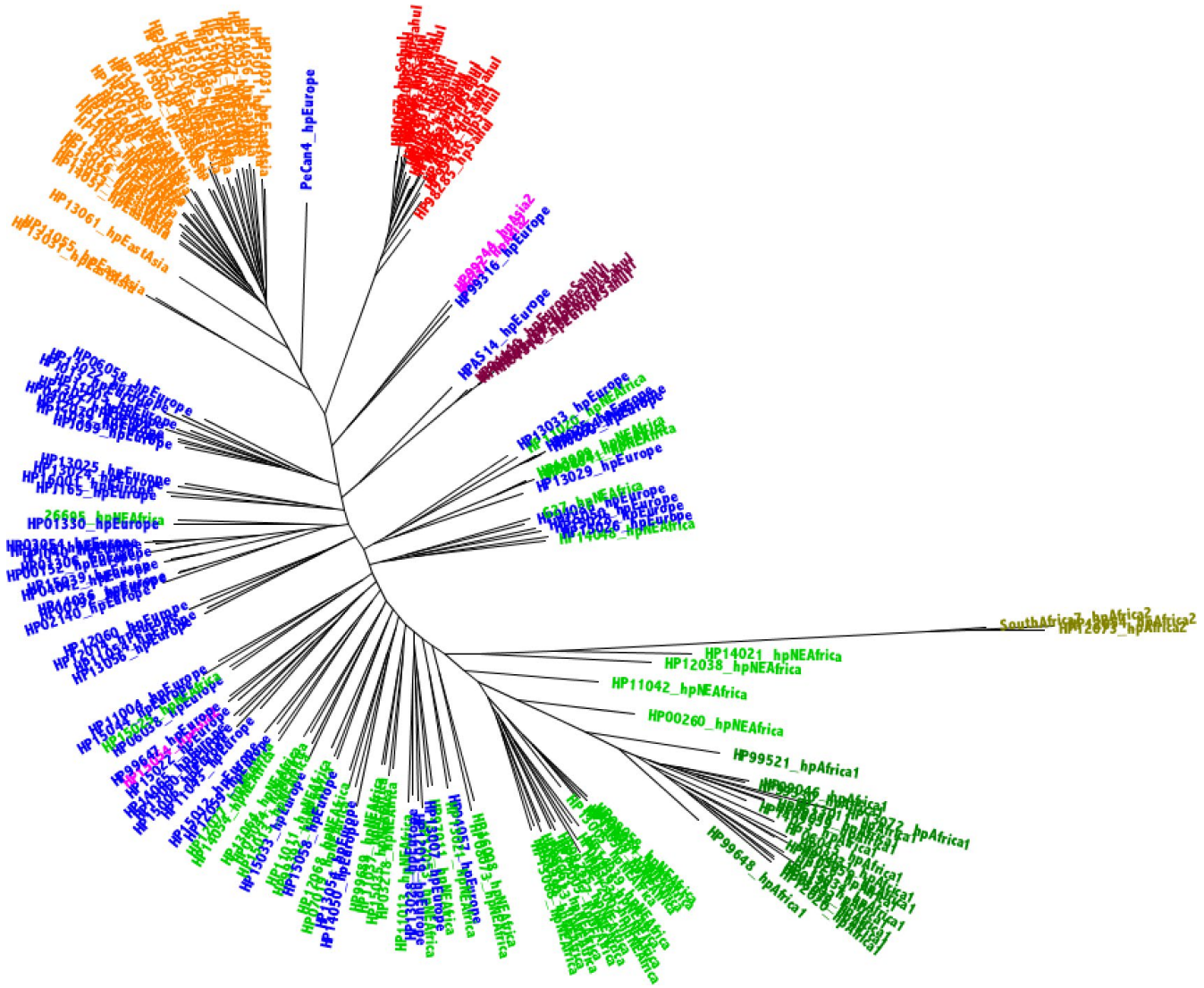

FIGURE 3 185 Strains core proteome tree. A phylogenetic tree was created by ExaML, based on core proteome concatenations of 898 protein sequences from all 185 strains. A starting tree was created using Mega. Five hundred bootstrapped trees were computed, with the best (i.e., lowest absolute value log likelihood) shown here. The taxa have been labelled, both in the labels and by colour, according to the highest percentage STRUCTURE group, based on the best (lowest absolute value log likelihood) of 12 runs for k = 7 bins (based on suggestion from Moodley et al. (2009)). The clade colours are as follows: hpAfrica2 (olive), hpNEAfrica (bright green), hpAfrica1 (dark green), hpEurope (blue), hpAsia2 (pink), hpEastAsia (orange) and hpSahul (red). The sub-clade of hpSahul that we now discover is a new clade, hpEuropeSahul, is shown in maroon

that have been preserved in the population due to founder effects. However, even after the selection of quintessent strains, certain genes/proteins will exhibit some recombination signal (at thresholds described above), so these have been omitted from the quintessent concatenations.

TABLE 1 Counts of genes/proteins whose Phi *p*-value is greater than designated threshold, thus failing to reject null hypothesis of recombination

Strains set	Count of genes (*N* = 898)			
	Phi > _0.1	Phi > _0.5	Phi > _0.01	≤1 Site (0 sites)
quint aa	482	562	693	10 (4)
all aa	439	518	633	4 (0)
all nt	150	187	255	0 (0)

In view of these results, it is worth turning our attention to the corresponding MLST concatenations, which have been the foundation of many previous analyses; the Phi *p*-value was 0, comprehensively rejecting the null hypothesis of nonrecombination. In other words, the MLST gene fragment concatenations are overwhelmingly recombinant.

3.4 | Using quintessents to build phylogenetic trees

The quintessent strains and their STRUCTURE groups are denoted using '*' in Tables S1 and S2. Of the 185 strains in the starting set,

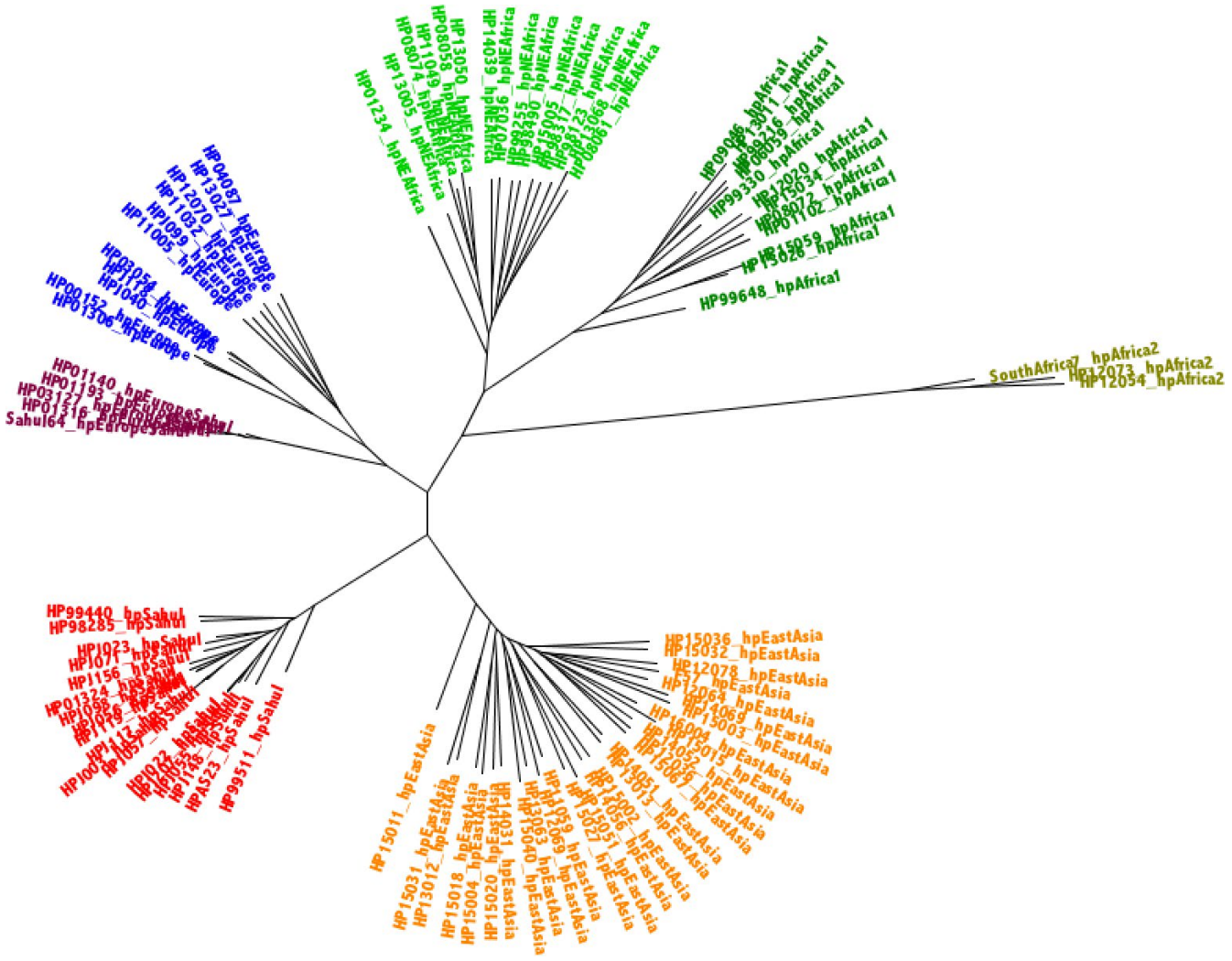

FIGURE 4 93 Quintessent strain tree. A phylogenetic tree was created by ExaML, based on concatenations of 472 minimally recombinant proteins from 93 strains that were found to be quintessents. A starting tree was created using Mega. Five hundred bootstrapped trees were computed, with the best (i.e., lowest absolute value log likelihood) shown here. The taxa have been labelled, both in the labels and by colour, according to the highest percentage STRUCTURE group, based on the best (lowest absolute value log likelihood) of 12 runs for k = 7 bins (based on suggestion from Moodley et al. (2009)). The clade colours are as follows: hpAfrica2 (olive), hpNEAfrica (bright green), hpAfrica1 (dark green), hpEurope (blue), hpEastAsia (orange) and hpSahul (red). The sub-clade of hpSahul that we now discover is a new clade, hpEuropeSahul, is shown in maroon

92 strains were removed as they are likely to be hybrid strains resulting from local recombinations between *H. pylori* strains, for which there was limited evidence of preservation in a significant sub-population. Thus, we lost a significant proportion of the hpEurope strains, where only 16 out of 89 strains were selected. We assume that this is because of the considerable history of human migration across Europe (Lazaridis et al., 2014). Similarly, both of the hpAsia2 strains sequenced for this study could not form a separate group with probability 0.75 in 60% of the runs and therefore were removed. Given that hpAsia2 has been identified as a sub-population in other studies, the disappearance of these two representatives may simply be due to both of these examples being recombinant. Similarly, the reference strains used in this study, 26695, G27, Pecan4, J99 and India7, also failed to be included as quintessents, suggesting they do not fully represent the sub-populations they are generally associated with, for example, hpEurope, in the case of 26695.

A phylogenetic tree created using the 472 minimally recombinant proteins from these 93 strains (Figure 4) showed a clustering pattern that is similar to trees obtained from the core genomes and core proteomes, but much more clearly delineated. The five quintessent strains isolated from the Aboriginal Australian individuals—originally characterized as hpSahul based on MLST data, but now assigned to a separate clade closer to hpEurope based on core genome and quintessent phylogenies—together suggest a speciation event where hpEurope strains have recombined with hpSahul strains to form a new sub-population. Therefore, based on the STRUCTURE and phylogenetic data, we have named this new clade hpEuropeSahul.

hpEuropeSahul strains are distinct from both of the parent clades, as their core genes are a mosaic of a majority of genes which are closer in phylogenetic distance to hpEurope and a smaller number of genes which are closer to hpSahul. Specifically, of the 898 core genes, for 662 genes hpEuropeSahul clade is closer to the nearest hpEurope than the nearest hpSahul. For 125 genes, the nearest hpSahul is closer, and for 111 genes, the hpEuropeSahul clade appears equidistant to the parent clades. Given that the strains were obtained from unrelated individuals, we can assume that this clade is being stably inherited. In terms of the evolutionary biology of hpEuropeSahul compared with hpSahul, it is clear that the speciation event that gave rise to hpSahul happened 65 kya, when Aboriginal Australians first migrated to the continent of Australia (then Sahul). What is less clear is when hpEuropeSahul arose, but we assume it is of recent origin, reflecting the colonization of Australia by European settlers around 200 years ago. There is some support for this in the locations where the quintessent strains from Aboriginal Australians were collected. Thirteen quintessent hpSahul strains were collected from the remote settlement of Jigalong (Windsor et al., 2005), one from Alice Springs and four from Perth. By contrast, all five hpEuropeSahul strains were collected from Perth.

3.5 | Comparing the quintessents method to ClonalFrameML

Table 2 summarizes the data related to cladistic information for the original nucleotide data set (all nt), which contains a considerable level of recombination, together with data from the same data set once it had been processed by ClonalFrameML (all nt cf) and 93 sequences drawn from all nt, corresponding to the strains whose core genomes were identified as being quintessents (quint nt). Unlike its parent data set, the all nt cf data set has a Phi *p*-value of 1.0, so is clearly nonrecombinant, but the sequences are now much shorter than the parent sequences (or those in quint nt), and the number of informative sites was decreased from 272,018 in the parent set to 224,461 in quint nt to 188,404 in all nt cf. The total entropy score (in bits) in fact increases slightly in quint nt over the parent, presumably because the splits between strains induced by the different sites are more balanced (i.e., half the strains have one nucleotide, while the other have a different nucleotide) in the quintessent data set, which reflects the smaller number of larger clades. On the other hand, when duplicate sites with the same split are removed, there is a significant drop in the dCITE score (Wise, 2016). This is not seen in the all nt cf set, which ends up having the same score as the parent. What the drop reflects is linkage disequilibrium, which we know exists in *H. pylori*, so the fact that the drop is far less in the all nt cf set suggests that that signal is being disrupted by the nonquintessent strains. It should also be noted that invariant sites have been removed from the all nt cf set, which precludes the use of substitution models involving a percentage of invariant sites. It also precludes use of codon-based models that are generally superior to single nucleotide substitution models (Shapiro, Rambaut, & Drummond, 2006).

TABLE 2 Informative sites and cladistic information content (measured by dCITE scores) for all nt data set, all nt processed by ClonalFrameML and quintessent nucleotide data set drawn from aa nt

		Strains set	
	Metric nt	all nt	all nt cf quint
Length (nt)	828,027	188,404	828,027
Informative sites	272,018	188,404	224,461
Total entropy (bits)	89,870	84,489	90,635
dCITE (bits)	82,526	82,526	75,312

TreeCmp was used to compare the trees produced by the ClonalFrameML-derived data set (all nt cf), the quintessent nucleotide tree (quint nt) and the amino acid quintessent data set involving minimally recombinant genes (quint aa). The results for four different tree-difference metrics are shown in Table 3, with visualizations of the trees available as Figures S1 and S2. The TreeCmp prune option was used to enable comparisons of trees with different counts of taxa—the quintessent sets have 93 taxa versus the other data sets' 185—so only the shared taxa are compared. Each of these trees was compared to all the others described above. The first thing that emerges from Table 3 is that, based on 93 common strains, the trees from the all nt cf and quint aa data sets are reasonably similar. However, viewed from the tree due to the quint aa data set, the quint nt data set is closer, despite having been based on concatenations of 898 nucleotide sequences rather than 472 amino acid sequences. The similarity is quite evident in the visualizations of the respective trees.

4 | CONCLUSIONS

Our study has described a new method for taxon selection based on taking the strains whose genomes have minimal evidence of recombination. These exemplar strains, which we have called quintessents, represent particular sub-populations of the input strains. In addition, by moving from nucleotide to amino acid data, and through use of quintessents, we have shown that recombination noise can be greatly reduced, exposing more clearly ancient recombination events that are evident as speciation events. As a demonstration of the new approach, from a starting set of 177 new *H. pylori* genomes plus 8 from the literature, we found 93 quintessents representing 7 *H. pylori* sub-populations, including a new Sahul sub-population that has arisen as a result of a recombination event involving hpEurope and hpSahul strains, that is now being stably inherited. Finally, this study has provided further evidence that, in order to get better resolution in phylogenetic analyses, one needs to include more genes than the conventional MLST concatenations, and those genes need to be minimally recombinant. For future work, a more rigorous method, perhaps using an

TABLE 3 All-against-all comparison, using TreeCmp, of trees produced all 6 data sets all nt, all aa, mlst and quint aa, all nt cf and quint nt, with the focus on the trees from the ClonalFrameML-derived data set (all nt cf)

			Tree-difference metrics			
Set1	Set2	Common taxa	R-F	MatchingSplit	PathDifference	Quartet
all nt cf	all nt cf	185	0	0	0	0
all nt cf	quint nt	93	24	100	103.9711	122,122
all nt cf	all nt	185	12	109	131.5979	210,794
all nt cf	quint aa	93	37	156	138.9676	223,553
all nt cf	all aa	185	88	436	386.3832	2,258,933
all nt cf	mlst	185	147	1,121	820.1683	12,265,590
quint nt	quint nt	93	0	0	0	0
quint nt	all nt cf	93	24	100	103.9711	122,122
quint nt	quint aa	93	34	118	136.8722	125,298
quint nt	all nt	93	25	105	106.7895	129,569
quint nt	all aa	93	37	206	162.8435	246,198
quint nt	mlst	93	67	372	281.0125	617,813
quint aa	quint aa	93	0	0	0	0
quint aa	quint nt	93	34	118	136.8722	125,298
quint aa	all nt cf	93	37	156	138.9676	223,553
quint aa	all nt	93	38	151	131.1106	223,594
quint aa	all aa	93	41	211	160.1187	251,317
quint aa	mlst	93	67	364	280.7526	604,579

Note: The quintessent data sets involve 93 strains, while the other data sets involve 185 strains. All the metrics are difference metrics, so a distance of 0 implies identical sequences.

information-theoretic metric, is required for selecting the quintessent bin probability and percentage of run values.

ACKNOWLEDGEMENTS

The authors would like to thank and acknowledge Ms Fanny Peters, who performed the DNA extraction and some of the sequencing, and Dr Mary Webberley, who commented on an early draft of this paper.

CONFLICT OF INTEREST

Barry J. Marshall is medical director of Tri-Med (http://www.trimed.com.au), a Perth company which distributes diagnostic tests for *Helicobacter pylori* ('PYtest' urea breath tests and 'CLOtest' biopsy rapid urease test) and marketing orphan drugs (bismuth subcitrate, tetracycline, furazolidone and rifaximin).

DATA AVAILABILITY STATEMENT

The genomes are available as NCBI Bioproject PRJNA374603. A spreadsheet with the STRUCTURE runs for K = 11 can be downloaded from the University of Western Australia Repository, https://doi.org/10.26182/5d64e7694a120 (Wise, 2019).

ORCID

Binit Lamichhane https://orcid.org/0000-0002-5808-8274
Michael J. Wise https://orcid.org/0000-0001-6559-4303
Barry J. Marshall https://orcid.org/0000-0003-4853-5015
Chin Yen Tay https://orcid.org/0000-0001-9705-4010

REFERENCES

Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., Pan, Z.-J., … van Doorn, L.-J. (1999). Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Molecular Microbiology*, *32*, 459–470. https://doi.org/10.1046/j.1365-2958.1999.01382.x

Baltrus, D. A., Guillemin, K., & Phillips, P. C. (2007). Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. *Evolution*, *62*, 39–49. https://doi.org/10.1111/j.1558-5646.2007.00271.x

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021

Bogdanowicz, D., Giaro, K., & Wróbel, B. (2012). Treecmp: Comparison of trees in polynomial time. *Molecular Microbiology*, *8*, 475–487. https://doi.org/10.4137/EBO.S9657

Bruen, T. C., Philippe, H., & Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics*, *172*, 2665–2681. https://doi.org/10.1534/genetics.105.048975

Clarkson, C., Jacobs, Z., Marwick, B., Fullagar, R., Wallis, L., Smith, M., … Pardoe, C. (2017). Human occupation of northern Australia by 65,000 years ago. *Nature*, *547*, 306–310. https://doi.org/10.1038/nature22968

Covacci, A., Telford, J. L., Giudice, G. D., Parsonnet, J., & Rappuoli, R. (1999). *Helicobacter pylori* virulence and genetic geography. *Science*, *284*, 1328–1333. https://doi.org/10.1126/science.284.5418.1328

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., … Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, *15*, e15. https://doi.org/10.1093/nar/gku1196

Didelot, X., & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, *175*, 1251–1266. https://doi.org/10.1534/genetics.106.063305

Didelot, X., Nell, S., Yang, I., Woltemate, S., van der Merwe, S., & Suerbaum, S. (2013). Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 13880–13885.

Didelot, X., & Wilson, D. J. (2015). Clonalframeml: Efficient inference of recombination in whole bacterial genomes. *PLoS Computational Biology*, *11*, e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Earl, D. A., & vonHoldt, B. M. (2012). Structure harvester: A website and program for visualizing structure output and implementing the Evanno method. *Conservation Genetics Resources*, *4*, 359–361. https://doi.org/10.1007/s12686-011-9548-7

Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M., & Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: Estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 15056–15061. https://doi.org/10.1073/pnas.251396098

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., … Suerbaum, S. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science*, *299*, 1582–1585. https://doi.org/10.1126/science.1080857

Fischer, W., Windhager, L., Rohrer, S., Zeiller, M., Karnholz, A., Hoffmann, R., … Haas, R. (2010). Strain-specific genes of *Helicobacter pylori*: Genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic Acids Research*, *38*, 6089–6101. https://doi.org/10.1093/nar/gkq378

Garcia-Ortiz, M.-V., Marsin, S., Arana, M. E., Gasparutto, D., Gueaccentbrois, R., Kunkel, T. A., & Radicella, J. P. (2011). Unexpected role for *Helicobacter pylori* DNA polymerase i as a source of genetic variability. *PLoS Genetics*, *7*, e1002152. https://doi.org/10.1371/journal.pgen.1002152

Goh, K.-L., Chan, W.-K., Shiota, S., & Yamaoka, Y. (2011). Epidemiology of *Helicobacterpylori*infectionandpublichealthimplications.*Helicobacter*, *16*, 1–9. https://doi.org/10.1111/j.1523-5378.2011.00874.x

Gressmann, H., Linz, B., Ghai, R., Pleissner, K.-P., Schlapbach, R., Yamaoka, Y., … Achtman, M. (2005). Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLoS Genetics*, *1*, e43. https://doi.org/10.1371/journal.pgen.0010043

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*, 696–704. https://doi.org/10.1080/10635150390235520

Katoh, K., & Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010

Kersulyte, D., Chalkauskas, H., & Berg, D. E. (1999). Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Molecular Microbiology*, *31*, 31–43. https://doi.org/10.1046/j.1365-2958.1999.01140.x

Kozlov, A. M., Aberer, A. J., & Stamatakis, A. (2015). Examl version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics*, *13*, 2577–2579. https://doi.org/10.1093/bioinformatics/btv184

Kumar, N., Albert, M. J., Abkal, H. A., Siddique, I., & Ahmed, N. (2017). What constitutes an Arabian *Helicobacter pylori*? Lessons from comparative genomics. *Helicobacter*, *22*, e12323.

Kumar, N., Mariappan, V., Baddam, R., Lankapalli, A. K., Shaik, S., Goh, K.-L., … Ahmed, N. (2015). Comparative genomic analysis of *Helicobacter pylori* from Malaysia identifies three distinct lineages suggestive of differential evolution. *Nucleic Acids Research*, *43*, 324–335. https://doi.org/10.1093/nar/gku1271

Kumar, S., Stecher, G., & Tamura, K. (2016). Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*, 1870–1874. https://doi.org/10.1093/molbev/msw054

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*, 409–413. https://doi.org/10.1038/nature13673

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, *12*, 124. https://doi.org/10.1186/1471-2105-12-124

Lin, S. K., Lambert, J. R., Nicholson, L., Lukito, W., & Wahlqvist, M. (1998). Prevalence of *Helicobacter pylori* in a representative anglo-celtic population of urban Melbourne. *Journal of Gastroenterology and Hepatology*, *13*, 505–510. https://doi.org/10.1111/j.1440-1746.1998.tb00677.x

Linz, B., Balloux, F., Moodley, Y., Manica, A., Liu, H., Roumagnac, P., … Achtman, M. (2007). An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, *445*, 915–991. https://doi.org/10.1038/nature05562

Lu, W., Wise, M. J., Tay, C. Y., Windsor, H. M., Marshall, B. J., Peacock, C., & Perkins, T. (2014). Comparative analysis of the full genome of *Helicobacter pylori* isolate sahul64 identifies genes of high divergence. *Journal of Bacteriology*, *196*, 1073–1083. https://doi.org/10.1128/JB.01021-13

Maixner, F., Krause-Kyora, B., Turaev, D., Herbig, A., Hoopmann, M. R., Hallows, J. L., … Zink, A. (2016). The 5300-year-old *Helicobacter pylori* genome of the iceman. *Science*, *351*(6269), 162–165. https://doi.org/10.1126/science.aad2545

Montano, V., Didelot, X., Foll, M., Linz, B., Reinhardt, R., Suerbaum, S., … Jensen, J. D. (2015). Worldwide population structure, long-term demography, and local adaptation of *Helicobacter pylori*. *Genetics*, *200*, 947–963.

Moodley, Y., Linz, B., Yamaoka, Y., Windsor, H. M., Breurec, S., Wu, J.-Y., … Achtman, M. (2009). The peopling of the pacific from a bacterial perspective. *Science*, *323*, 527–530. https://doi.org/10.1126/science.1166083

Muñoz-Ramírez, Z. Y., Mendez-Tenorio, A., Kato, I., Bravo, M. M., Rizzato, C., Thorell, K., … Torres, J. (2017). Whole genome sequence and phylogenetic analysis show *Helicobacter pylori* strains from Latin America have followed a unique evolution pathway. *Frontiers in Cellular and Infection Microbiology*, *7*, 50. https://doi.org/10.3389/fcimb.2017.00050

Nabhan, A. R., & Sarkar, I. N. (2012). The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy. *Briefings in Bioinformatics*, *13*, 122–134. https://doi.org/10.1093/bib/bbr014

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). Snp-sites: Rapid efficient extraction of snps

from multi-fasta alignments. *Microbial Genomics*, *2*, e000056. https://doi.org/10.1099/mgen.0.000056

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of structure: Applications, parameter settings, and supporting software. *Front Genet*, *4*, 98. https://doi.org/10.3389/fgene.2013.00098

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Qin, T., Zhang, W., Liu, W., Zhou, H., Ren, H., Shao, Z., … Xu, J. (2016). Population structure and minimum core genome typing of legionella pneumophila. *Scientific Reports*, *6*, 21356. https://doi.org/10.1038/srep21356

Robertson, M. S., Cade, J. F., Savoia, H. F., & Clancy, R. L. (2003). *Helicobacter pylori* infection in the Australian community: Current prevalence and lack of association with abo blood groups. *Internal Medicine Journal*, *33*, 163–167.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shapiro, B., Rambaut, A., & Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, *23*, 7–9. https://doi.org/10.1093/molbev/msj021

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, *7*, 539. https://doi.org/10.1038/msb.2011.75

Stamatakis, A. (2014). Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Suerbaum, S., & Josenhans, C. (2007). *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nature Reviews Microbiology*, *5*(441–452), 441–452. https://doi.org/10.1038/nrmicro1658

Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., … Achtman, M. (1998). Free recombination within *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 12619–12624. https://doi.org/10.1073/pnas.95.21.12619

Tay, C. Y., Mitchell, H., Dong, Q., Goh, K.-L., Dawes, I. W., & Lan, R. (2009). Population structure of *Helicobacter pylori* among ethnic groups in Malaysia: Recent acquisition of the bacterium by the Malay population. *BMC Microbiology*, *9*, 126. https://doi.org/10.1186/1471-2180-9-126

Windsor, H. M., Abioye-Kuteyi, E. A., Leber, J. M., Morrow, S. D., Bulsara, M. K., & Marshall, B. J. (2005). Prevalence of *Helicobacter pylori* in indigenous western Australians: Comparison between urban and remote rural populations. *Medical Journal of Australia*, *182*, 210–213.

Wise, M. J. (2013). Mean protein evolutionary distance: A method for comparative protein evolution and its application. *PLoS ONE*, *8*, e61276. https://doi.org/10.1371/journal.pone.0061276

Wise, M. J. (2016). Measuring necessary cladistic information can help you reduce polytomy artefacts in trees. *PLoS ONE*, *11*, e0166991.

Wise, M. J. (2019). 185_structure_2_k_11.xlsx. *University of Western Australia Profiles and Research Repository*, https://doi.org/10.26182/5d64e7694a120

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Lamichhane B, Wise MJ, Chua EG, Marshall BJ, Tay CY. A novel taxon selection method, aimed at minimizing recombination, clarifies the discovery of a new sub-population of *Helicobacter pylori* from Australia. *Evol Appl*. 2020;13:278–289. https://doi.org/10.1111/eva.12864