

# Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data

Marcos Pérez-Losada<sup>a,\*</sup>, Emily B. Browne<sup>a</sup>, Aaron Madsen<sup>a</sup>, Thierry Wirth<sup>b</sup>,  
Raphael P. Viscidi<sup>c</sup>, Keith A. Crandall<sup>a,d</sup>

<sup>a</sup> Department of Integrative Biology, Brigham Young University, Provo, UT 84602, USA

<sup>b</sup> Department of Biology, Universitaetsstrasse 10, University Konstanz, D-78457 Konstanz, Germany

<sup>c</sup> Department of Pediatrics, Johns Hopkins Hospital, The Johns Hopkins Medical School, Baltimore, MD 21287, USA

<sup>d</sup> Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA

Received 3 June 2004; received in revised form 18 November 2004; accepted 14 February 2005

Available online 24 March 2005

## Abstract

The inference of population recombination ( $\rho$ ), population mutation ( $\Theta$ ), and adaptive selection is of great interest in microbial population genetics. These parameters can be efficiently estimated using explicit statistical frameworks (evolutionary models) that describe their effect on gene sequences. Within this framework, we estimated  $\rho$  and  $\Theta$  using a coalescent approach, and adaptive (or destabilizing) selection under heterogeneous codon-based and amino acid property models in microbial sequences from MLST databases. We analyzed a total of 91 different housekeeping gene regions (loci) corresponding to one fungal and sixteen bacterial pathogens. Our results show that these three population parameters vary extensively across species and loci, but they do not seem to be correlated. For the most part, estimated recombination rates among species agree well with previous studies. Over all taxa, the  $\rho/\Theta$  ratio suggests that each factor contributes similarly to the emergence of variant alleles. Comparisons of  $\Theta$  estimated under finite- and infinite-site models indicate that recurrent mutation (i.e., multiple mutations at some sites) can increase  $\Theta$  by up to 39%. Significant evidence of molecular adaptation was detected in 28 loci from 13 pathogens. Three of these loci showed concordant patterns of adaptive selection in two to four different species.

© 2005 Elsevier B.V. All rights reserved.

**Keywords:** Coalescent; Evolutionary models; Genetic diversity; Population structure; Recombination; Selection

## 1. Introduction

Maynard-Smith (1995) pointed out the need for population genetic insights when contemplating the evolutionary fate of infectious diseases. Population genetics is important in understanding the evolutionary history, epidemiology, and population dynamics of pathogens, the potential for and mode of the evolution of antibiotic resistance, and ultimately for public health control strategies. The key factors in the evolutionary response of pathogens to their environments can be measured by assessing the genetic diversity (and partitioning of that diversity within versus between populations), the impact of natural selection in shaping

that diversity, and the impact of recombination in redistributing that diversity, sometimes into novel combinations. Population studies of pathogens using multilocus sequencing typing (MLST) methods are generally aimed at inferring genetic diversity (usually estimated as the relative contribution of recombination and mutation per allele or per site), selection pressure, and population structure (Spratt and Maiden, 1999; Maynard-Smith et al., 2000; Dingle et al., 2001; Feil et al., 2003; Meats et al., 2003; Viscidi and Demma, 2003) to study the relative impact of genetic drift and natural selection on the evolutionary history of these pathogens.

Population parameters can be efficiently estimated using explicit statistical models of evolution, such as the coalescent approach, that describe their effect on gene sequences (Hudson, 1990; Nordborg, 2001; Felsenstein,

\* Corresponding author. Tel.: +1 801 422 9378; fax: +1 801 422 0090.  
E-mail address: [mp323@email.byu.edu](mailto:mp323@email.byu.edu) (M. Pérez-Losada).

2004). Consider, for example, recombination and mutation rates. They can be estimated separately using a standard coalescent approach that assumes large Fisher–Wright populations, nonoverlapping generations, constant population size, and no selection or migration (or recombination when estimating mutation rates). A model-based method such as this is almost certainly a simplification of reality, but the benefits gained are significant, namely the ease of comparison between genes or species, the ability to make predictions about the question of interest, and the potential to test whether the model of evolution is an adequate characterization of the underlying process (McVean et al., 2002).

In addition, in the case of recombination, the coalescent model can be used to test the presence of the parameter by comparing the likelihood of the data with and without recombination (Brown et al., 2001). Under “model-free” methods such as the index of association (Maynard-Smith et al., 1993) and the homoplasy test (Maynard-Smith and Smith, 1998), gene or species comparisons of recombination rates are problematic and there is little or no way of statistically testing whether data sets have different levels of recombination (Maynard-Smith et al., 2000; McVean et al., 2002).

When dealing with MLST sequence data it is important to have evolutionary models that accurately describe the process of DNA substitution (e.g., Yang et al., 1994; Yang, 1997; Kelsey et al., 1999; Posada and Crandall, 2001a). Accurate models can help clarify some of the most important processes of evolution (e.g., selection pressure) by the biological interpretation of their parameters, and provide more reliable estimates of other model-based statistics (e.g., coalescent estimates of recombination and mutation) (Goldman and Yang, 1994). The effect of natural selection on molecular sequence evolution is almost always calculated as an average over all codon (amino acid) sites in the gene and over the entire evolutionary time that separates the sequences (Yang et al., 2000a). But this criterion is a very stringent one for detecting positive selection, especially in conservative proteins such as those encoded by the housekeeping genes (Crandall et al., 1999). Conservative proteins present a high proportion of invariable amino acids and appear to be under purifying selection all the time (Li, 1997). Hence adaptive evolution, if present, is most likely punctual, that is, it will affect a few amino acid residue sites (e.g., Endo et al., 1996; Li, 1997; Yang et al., 2000a). Consequently, evolutionary models that do not allow for selection heterogeneity among sites, such as the one implemented by Nei and Gojobori (1986), will certainly not detect those few sites under positive selection. Several evolutionary models exist that account for site-specific differences on adaptive selection at the protein level (Nielsen and Yang, 1998; Yang et al., 2000a; McClellan and McCracken, 2001; Yang and Swanson, 2002), and their utility has been already demonstrated (e.g., Yang et al., 2000a,b; Haydon et al., 2001; Yang and Nielsen, 2002; McClellan et al., 2005); however, MLST data are not usually examined using these approaches.

MLST was proposed in 1998 (Maiden et al., 1998) as a general approach to provide accurate, portable data that were appropriate for bacterial epidemiological investigation and which also reflected their evolutionary and population biology (Urwin and Maiden, 2003). Since then, sequence data from 17 different prokaryotic and eukaryotic microbial pathogens and almost 100 housekeeping genes have been published and are currently available via the Internet. Now, several key questions concerning microbial population genetics can be addressed using these MLST databases: how do recombination, mutation, and selection pressure vary across species and loci? Are they correlated? Which is the major force generating genetic diversity? Are MLST housekeeping genes under adaptive selection? Our goal here is to answer these questions within an evolutionary-model framework using the approaches described above.

A logical concern in this study is the adequacy of the available MLST sequences for assessing these questions. The data retrieved from the databases, although representing the reported diversity of the organisms, are unstructured and are not necessarily representative of natural populations (Urwin and Maiden, 2003). Moreover, besides the particular case of the *Neisseria* database and, to some extent, the *Helicobacter pylori* database, all the other databases contain information from a limited number of isolates that do not represent the worldwide distribution of the species and rarely include the less pathogenic samples which frequently comprise the majority of the population (Spratt and Maiden, 1999). These caveats can obviously bias the estimates of the parameters of interest (recombination, mutation, and selection rates), although we think not to the extent of completely misleading the inferences deduced from them. Nevertheless, for comparative purposes, we will also analyze published subsets of the database sequence files including strains from asymptomatic carriage and local and worldwide collections

## 2. Materials and methods

### 2.1. Data sets

Our DNA sequence data sets consisted of 91 different loci corresponding to one yeast and fifteen bacterial pathogens (a total of 184 data sets; Table 1) downloaded from two MLST databases at <http://www.mlst.net> and <http://pubmlst.org/> (see also acknowledgements). Seventeen additional data sets for *Escherichia coli* and *Moraxella catarrhalis* were provided by one of us (TW) and can be accessed at <http://web.mpn-berlin.mpg.de/mlst>. We analyzed complete MLST allele sequences (as of January 2004) for each bacterial species in order to have a good representation of their population diversity. Additionally, for the following pathogens, we analyzed subsets of published data for comparison: *Haemophilus influenzae* (encapsulated and/or noncapsulated; Meats et al., 2003), *H. pylori* (Achtman et al., 1999), *Neisseria meningitidis* (Maiden et al., 1998),

Table 1

Population recombination rate ( $\Gamma$ ) and the probability of  $\Gamma = 0$  (indicated by asterisks) from the LPT test, population mutation rate using Watterson's method under an infinite-sites model ( $\Gamma_{Wi}$ ) and a finite-sites model ( $\Gamma_{Wf}$ ), per allele ratio of recombination to mutation ( $\Gamma/\Gamma_{Wf}$ ), and best-fit model of evolution per locus for every species

Locus	Alleles	Sites	$\Gamma$	$\Gamma_{Wi}$	$\Gamma_{Wf}$	$\Gamma/\Gamma_{Wf}$	Model
<i>Bacillus cereus</i>							
<i>glp</i>	40	381	11.9*	11.99	12.95	0.9	HKY + $\Gamma$
<i>gmk</i>	21	504	1.3	23.07	25.2	0.1	TrN + $\Gamma$
<i>ilv</i>	31	393	10.9**	20.28	22.79	0.5	HKY + $\Gamma$
<i>pta</i>	36	414	52.1*	14.71	15.73	3.3	TrN + $\Gamma$
<i>pur</i>	32	348	22.7*	19.12	21.58	1.1	TrN + $\Gamma$
<i>pyc</i>	36	363	5.8	20.98	23.96	0.2	TrN + $\Gamma$
<i>tpi</i>	26	435	55.3	8.12	8.27	6.7	TrN + $\Gamma$ + I
	21–40	348–504	1.3–55.3	8.12–23.7	8.27–25.2	0.1–6.7	
	<b>32</b>	<b>405</b>	<b>22.9</b>	<b>16.9</b>	<b>18.64</b>	<b>1.8</b>	
<i>Burkholderia pseudomallei</i>							
<i>ace</i>	11	519	0	15.02	15.57	0	TrN + $\Gamma$
<i>gltB</i>	19	522	0	10.3	10.44	0	K81uf + $\Gamma$
<i>gmhD</i>	23	468	0	12.46	13.1	0	HKY + $\Gamma$
<i>lepA</i>	16	486	1.4	11.15	11.66	0.1	TrN + $\Gamma$
<i>lipA</i>	17	402	0	10.35	10.85	0	HKY + $\Gamma$
<i>narK</i>	27	561	1.7	11.16	11.78	0.1	HKY + $\Gamma$ + I
	11–27	402–561	0–1.7	10.3–15.02	10.44–15.57	0–0.1	
	<b>19</b>	<b>493</b>	<b>0.5</b>	<b>11.74</b>	<b>12.24</b>	<b>0</b>	
<i>Candida albicans</i>							
<i>acc1</i>	22	407	–	2.04	2.04	–	F81
<i>adp1</i>	33	443	–	3.54	3.77	–	HKY + I
<i>gln4</i>	21	404	–	3.23	3.23	–	F81
<i>rpn2</i>	32	306	–	3.45	3.67	–	JC
<i>sya1</i>	35	391	–	2.65	2.74	–	F81uf + I
<i>vps13</i>	61	403	–	4.4	4.43	–	F81 + $\Gamma$
	21–61	306–443	–	2.04–3.54	2.04–4.43	–	
	<b>34</b>	<b>392</b>		<b>3.22</b>	<b>3.31</b>		
<i>Campylobacter jejuni</i>							
<i>aspA</i>	81	477	2.2**	18.13	20.03	0.1	TIM + $\Gamma$ + I
<i>glnA</i>	111	477	4.9*	23.56	27.19	0.2	TrN + $\Gamma$ + I
<i>gltA</i>	82	402	5.1	18.48	20.9	0.2	HKY + $\Gamma$ + I
<i>glyA</i>	117	507	0.1	32.45	39.55	0	TIM + $\Gamma$ + I
<i>pgm</i>	148	498	0	34.19	42.33	0	TrN + $\Gamma$ + I
<i>tkt</i>	118	459	1.5*	26.85	32.13	0	TrN + $\Gamma$ + I
<i>uncA</i>	67	489	1.8*	25.76	29.83	0.1	GTR + $\Gamma$ + I
	67–148	402–507	0–5.1	18.13–34.19	20.03–42.33	0–0.2	
	<b>103</b>	<b>473</b>	<b>2.2</b>	<b>25.63</b>	<b>30.28</b>	<b>0.1</b>	
<i>Escherichia coli</i>							
<i>adk</i> <sup>MA</sup>	72	536	0	28.27	32.7	0	TrNef + $\Gamma$ + I
<i>arcA</i> <sup>MA</sup>	22	564	0	18.38	19.74	0	TrNef
<i>aroE</i> <sup>MA</sup>	22	564	0	18.38	19.74	0	TrNef
<i>fumC</i> <sup>MA</sup>	83	465	0	32.06	39.53	0	TrNef + $\Gamma$ + I
<i>gyrB</i> <sup>MA</sup>	70	460	0	26.56	31.28	0	TrNef + $\Gamma$ + I
<i>icd</i> <sup>MA</sup>	76	516	0	22.86	25.8	0	TrN + $\Gamma$ + I
<i>icd</i>	22	1176	0	49.93	54.1	0	TrNef + $\Gamma$ + I
<i>mdh</i> <sup>MA</sup>	51	452	0	24.67	28.48	0	TrNef + $\Gamma$
<i>mdh</i>	22	846	0	40.6	44.84	0	K80 + $\Gamma$
<i>mltD</i>	22	1098	0	55.14	60.39	0	K80 + $\Gamma$
<i>pgi</i>	22	978	0	39.78	43.03	0	TrNef + $\Gamma$
<i>purA</i> <sup>MA</sup>	53	478	0	12.34	12.91	0	TrNef + $\Gamma$ + I
<i>recA</i> <sup>MA</sup>	60	510	0	21.87	24.48	0	TrNef + $\Gamma$ + I
<i>rpoS</i>	21	714	0	18.35	19.28	0	K80
	21–83	452–1176	0	12.34–55.14	12.91–60.39	0	
	<b>44</b>	<b>668</b>	<b>0</b>	<b>29.34</b>	<b>32.59</b>	<b>0</b>	
<i>Enterococcus faecium</i>							
<i>adk</i>	13	437	2.9	4.19	4.37	0.7	HKY
<i>atpA</i>	31	556	3.9	10.01	10.56	0.4	HKY + $\Gamma$

Table 1 (Continued)

Locus	Alleles	Sites	$\Gamma$	$\Gamma_{wi}$	$\Gamma_{wf}$	$\Gamma/\Gamma_{wf}$	Model
<i>ddl</i>	18	465	4	8.43	8.84	0.5	HKY + $\Gamma$
<i>gdh</i>	23	530	0	45.79	55.12	0	TrN + $\Gamma$
<i>gyd</i>	18	395	0	16.57	17.78	0	GTR
<i>pstS</i>	38	583	9.7 <sup>***</sup>	16.18	17.49	0.6	HKY + $\Gamma$ + I
<i>purK</i>	29	492	0	13.5	14.27	0	K80 + $\Gamma$
	13–38	395–583	0–9.7	4.19–45.79	4.37–55.12	0–0.7	
	<b>24</b>	<b>494</b>	<b>2.9</b>	<b>16.38</b>	<b>18.35</b>	<b>0.3</b>	
<i>Haemophilus influenzae</i>							
<i>adk</i>	32	477	73 <sup>***</sup>	16.14	17.17	4.3	TrN + $\Gamma$ + I
<i>adk</i> <sup>1</sup>	23	477	24 <sup>**</sup>	14.73	15.74	1.5	TrN + $\Gamma$ + I
<i>adk</i> <sup>1eca</sup>	13	477	15 <sup>***</sup>	15.41	16.22	0.9	TrN + $\Gamma$ + I
<i>adk</i> <sup>1nca</sup>	10	477	20 <sup>***</sup>	7.42	7.63	2.6	TrN + $\Gamma$ + I
<i>atpG</i>	33	447	6.7 <sup>*</sup>	9.61	9.83	0.7	HKY + $\Gamma$ + I
<i>atpG</i> <sup>1</sup>	26	447	10 <sup>**</sup>	9.17	9.39	1.1	HKY + $\Gamma$ + I
<i>atpG</i> <sup>1eca</sup>	13	447	4	8.06	8.49	0.5	HKY + $\Gamma$ + I
<i>atpG</i> <sup>1nca</sup>	13	447	4	9.02	9.39	0.4	HKY + $\Gamma$ + I
<i>frdB</i>	33	489	17.9 <sup>***</sup>	13.45	14.18	1.3	HKY + $\Gamma$ + I
<i>frdB</i> <sup>1</sup>	26	489	13 <sup>***</sup>	12.92	13.69	0.9	HKY + $\Gamma$ + I
<i>frdB</i> <sup>1eca</sup>	17	489	4 <sup>***</sup>	11.54	12.23	0.3	HKY + $\Gamma$ + I
<i>frdB</i> <sup>1nca</sup>	9	489	17 <sup>**</sup>	11.77	12.23	1.4	HKY + $\Gamma$ + I
<i>fucK</i>	25	345	0	9	9.32	0	HKY
<i>fucK</i> <sup>1</sup>	22	345	0	8.62	9.97	0	HKY
<i>fucK</i> <sup>1eca</sup>	12	345	0	9.60	10.01	0	HKY
<i>fucK</i> <sup>1nca</sup>	10	345	0	7.42	7.59	0	HKY
<i>mdh</i>	46	405	100 <sup>***</sup>	13.68	14.99	6.7	HKY + $\Gamma$ + I
<i>mdh</i> <sup>1</sup>	36	405	100 <sup>***</sup>	12.06	12.96	7.7	HKY + $\Gamma$ + I
<i>mdh</i> <sup>1eca</sup>	21	405	100 <sup>***</sup>	12.34	12.96	7.7	HKY + $\Gamma$ + I
<i>mdh</i> <sup>1nca</sup>	15	405	34 <sup>*</sup>	13.52	14.18	2.4	HKY + $\Gamma$ + I
<i>pgi</i>	41	468	100 <sup>***</sup>	20.68	22.93	4.4	HKY + $\Gamma$ + I
<i>pgi</i> <sup>1</sup>	32	468	100 <sup>***</sup>	19.71	21.53	4.6	HKY + $\Gamma$ + I
<i>pgi</i> <sup>1eca</sup>	20	468	69 <sup>**</sup>	20.29	22.0	3.1	HKY + $\Gamma$ + I
<i>pgi</i> <sup>1nca</sup>	12	468	78 <sup>***</sup>	17.55	18.72	4.2	HKY + $\Gamma$ + I
<i>recA</i>	29	426	3 <sup>**</sup>	17.32	18.74	0.2	HKY + $\Gamma$ + I
<i>recA</i> <sup>1</sup>	23	426	12 <sup>***</sup>	9.75	10.22	1.2	HKY + $\Gamma$ + I
<i>recA</i> <sup>1eca</sup>	14	426	8 <sup>*</sup>	10.38	10.65	0.8	HKY + $\Gamma$ + I
<i>recA</i> <sup>1nca</sup>	9	426	12 <sup>*</sup>	6.99	7.24	1.7	HKY + $\Gamma$ + I
	25–46	345–489	0–100	9–20.68	9.32–22.93	0–6.7	
	<b>34</b>	<b>437</b>	<b>42.9</b>	<b>14.27</b>	<b>15.31</b>	<b>2.51</b>	
<i>Helicobacter pylori</i>							
<i>atpA</i>	310	627	100 <sup>***</sup>	21.83	23.83	4.2	GTR + $\Gamma$ + I
<i>atpA</i> <sup>2</sup>	19	627	100 <sup>***</sup>	17.79	20.06	5	GTR + $\Gamma$ + I
<i>efp</i>	303	410	100 <sup>***</sup>	18.74	21.32	4.7	TIM + $\Gamma$ + I
<i>efp</i> <sup>2</sup>	19	410	100 <sup>***</sup>	15.41	16.4	6.1	TIM + $\Gamma$ + I
<i>mutY</i>	324	420	100 <sup>***</sup>	26.27	31.92	3.1	GTR + $\Gamma$ + I
<i>mutY</i> <sup>2</sup>	19	420	100 <sup>***</sup>	24.61	27.72	3.6	GTR + $\Gamma$ + I
<i>ppa</i>	317	398	100 <sup>***</sup>	15.26	17.11	5.8	TIM + $\Gamma$ + I
<i>ppa</i> <sup>2</sup>	19	398	100 <sup>***</sup>	11.34	11.94	8.4	TIM + $\Gamma$ + I
<i>trpC</i>	322	456	100 <sup>***</sup>	33.61	42.41	2.4	GTR + $\Gamma$ + I
<i>trpC</i> <sup>2</sup>	19	456	100 <sup>***</sup>	32.90	37.85	2.6	GTR + $\Gamma$ + I
<i>urel</i>	334	585	100 <sup>***</sup>	18.35	19.89	5	TrN + $\Gamma$ + I
<i>urel</i> <sup>2</sup>	19	585	100 <sup>***</sup>	19.17	20.48	4.9	TrN + $\Gamma$ + I
<i>vacA</i>	338	444	93.9 <sup>***</sup>	24.53	28.86	3.3	GTR + $\Gamma$ + I
<i>vacA</i> <sup>2</sup>	19	444	100 <sup>***</sup>	24.61	27.53	3.6	GTR + $\Gamma$ + I
<i>yphC</i>	332	510	100 <sup>***</sup>	25.69	29.58	3.4	GTR + $\Gamma$ + I
<i>yphC</i> <sup>2</sup>	19	510	100 <sup>***</sup>	27.47	30.6	3.3	GTR + $\Gamma$ + I
	303–338	398–627	93.9–100	15.26–33.61	17.11–42.41	2.4–5.8	
	<b>323</b>	<b>481</b>	<b>99.2</b>	<b>23.04</b>	<b>26.86</b>	<b>4</b>	
<i>Moraxella catarrhalis</i>							
<i>abcZ</i>	49	429	0	30.95	37.32	0	K80 + $\Gamma$ + I
<i>adk</i>	37	471	0	19.64	21.67	0	TrNef + $\Gamma$ + I
<i>efp</i>	27	414	4.3 <sup>*</sup>	18.16	19.87	0.2	K80 + $\Gamma$
<i>fumC</i>	30	465	20.5 <sup>**</sup>	16.66	18.14	1.1	TrN + $\Gamma$ + I
<i>glyB</i>	60	537	0	24.02	26.85	0	TrN + $\Gamma$ + I

Table 1 (Continued)

Locus	Alleles	Sites	$\Gamma$	$\Gamma_{Wi}$	$\Gamma_{Wf}$	$\Gamma/\Gamma_{Wf}$	Model
<i>mutY</i>	50	426	0	22.1	25.13	0	TVM + $\Gamma$ + I
<i>ppa</i>	40	393	26.3 <sup>***</sup>	20.45	23.19	1.1	TrNef + $\Gamma$ + I
<i>trpE</i>	16	372	5.2 <sup>*</sup>	13.56	14.51	0.4	K80 + $\Gamma$
	16–60	372–537	0–26.3	13.56–30.59	14.51–37.32	0–1.1	
	<b>39</b>	<b>438</b>	<b>7.04</b>	<b>20.69</b>	<b>23.34</b>	<b>0.35</b>	
<i>Neisseria gonorrhoeae</i>							
<i>abcZ</i>	5	884	11.1 <sup>*</sup>	2.4	2.65	4.2	K80
<i>aroE</i>	5	796	37.4 <sup>**</sup>	1.92	1.59	23.5	F81 + I
<i>gdh</i>	12	861	100 <sup>***</sup>	2.65	2.58	38.8	HKY + $\Gamma$ + I
<i>glnA</i>	14	1356	19.2 <sup>*</sup>	9.75	9.49	2	TrN + $\Gamma$ + I
<i>gnd</i>	14	1446	100 <sup>***</sup>	4.09	4.34	23	HKY + I
<i>gpdh</i>	13	1023	100 <sup>***</sup>	3.22	3.07	32.6	HKY + $\Gamma$ + I
<i>gpdhC</i>	6	992	100 <sup>***</sup>	2.63	2.98	33.6	F81
<i>pdhC</i>	4	498	79.2 <sup>***</sup>	1.09	1	79.2	F81
<i>pgi1</i>	4	954	100 <sup>***</sup>	1.09	0.95	105.3	F81
<i>pgi2</i>	13	1618	2	2.58	3.24	0.6	F81
<i>pilA</i>	18	944	100 <sup>***</sup>	8.43	8.5	11.8	HKY + $\Gamma$ + I
<i>pip</i>	12	826	75.2 <sup>**</sup>	3.64	3.3	22.8	F81 + I
<i>ppk</i>	8	906	100 <sup>***</sup>	2.7	2.72	36.8	F81 + I
<i>pyrD</i>	9	1005	76.8 <sup>**</sup>	2.94	3.02	25.4	HKY
<i>serC</i>	9	1104	2	27.6	28.7	0.1	HKY
	4–18	498–1618	2–100	1.09–27.6	0.95–28.7	0.1–105.3	
	<b>10</b>	<b>1014</b>	<b>66.9</b>	<b>5.12</b>	<b>5.21</b>	<b>29.3</b>	
<i>Neisseria meningitidis</i>							
<i>abcZ</i>	221	433	88.9 <sup>***</sup>	26.07	31.18	2.9	TrN + $\Gamma$ + I
<i>abcZ</i> <sup>3</sup>	15	433	38 <sup>***</sup>	23.07	25.55	1.5	TrN + $\Gamma$ + I
<i>adk</i>	157	465	47.4 <sup>***</sup>	21.63	24.65	1.9	SYM + $\Gamma$ + I
<i>adk</i> <sup>3</sup>	12	465	30 <sup>**</sup>	12.58	13.02	2.3	SYM + $\Gamma$ + I
<i>fumC</i>	262	465	100 <sup>***</sup>	17.38	19.53	5.1	TrN + $\Gamma$ + I
<i>gdh</i>	263	501	3.8 <sup>*</sup>	26.27	30.56	0.1	GTR + $\Gamma$ + I
<i>gdh</i> <sup>3</sup>	16	501	2.4 <sup>*</sup>	8.44	8.52	0.3	GTR + $\Gamma$ + I
<i>pdhC</i>	251	480	49.4 <sup>***</sup>	21.44	24.48	2	TrN + $\Gamma$ + I
<i>pdhC</i> <sup>3</sup>	24	480	32 <sup>***</sup>	21.42	23.52	1.4	TrN + $\Gamma$ + I
<i>pgm</i>	257	450	62.6 <sup>***</sup>	28.59	34.65	1.8	TrN + $\Gamma$ + I
<i>pgm</i> <sup>3</sup>	21	450	37 <sup>***</sup>	21.40	23.4	1.6	TrN + $\Gamma$ + I
	157–263	433–501	3.8–100	17.38–47.51	19.53–66.15	0.1–5.1	
	<b>235</b>	<b>466</b>	<b>58.6</b>	<b>23.56</b>	<b>28.93</b>	<b>2.3</b>	
<i>Streptococcus agalactiae</i>							
<i>adhP</i>	31	498	68.7 <sup>**</sup>	7.26	7.47	9.2	HKY + $\Gamma$
<i>atr</i>	24	501	80 <sup>***</sup>	6.96	7.01	11.4	HKY
<i>glcK</i>	21	459	6.1	6.67	6.89	0.9	HKY
<i>glnA</i>	24	498	100 <sup>***</sup>	5.62	5.98	16.7	K81uf
<i>pheS</i>	15	501	18.2 <sup>*</sup>	4.92	5.01	3.6	HKY
<i>sdhA</i>	23	519	9.1 <sup>*</sup>	6.5	6.75	1.3	HKY + $\Gamma$
<i>tkf</i>	12	480	2	4.64	4.8	0.4	K81uf
	12–31	459–519	2–100	4.64–7.26	4.8–7.47	0.4–16.7	
	<b>21</b>	<b>494</b>	<b>40.6</b>	<b>6.08</b>	<b>6.27</b>	<b>6.2</b>	
<i>Staphylococcus aureus</i>							
<i>arcC</i>	49	456	0	19.51	21.43	0	HKY + $\Gamma$
<i>arcC</i> <sup>4</sup>	17	456	0	5.62	5.93	0	HKY + $\Gamma$
<i>aroE</i>	80	456	12.9 <sup>*</sup>	14.54	15.96	0.8	HKY + $\Gamma$
<i>aroE</i> <sup>4</sup>	17	456	19 <sup>*</sup>	6.63	6.84	2.8	HKY + $\Gamma$
<i>glpF</i>	51	465	0	19.11	20.93	0	K81uf + $\Gamma$
<i>glpF</i> <sup>4</sup>	11	465	0	5.12	5.12	0	K81uf + $\Gamma$
<i>gmk</i>	46	429	0	13.2	14.16	0	HKY + $\Gamma$
<i>gmk</i> <sup>4</sup>	11	429	0	4.44	4.72	0	HKY + $\Gamma$
<i>pta</i>	53	474	1.6	35.04	42.66	0	HKY + $\Gamma$
<i>pta</i> <sup>4</sup>	15	474	3	5.54	5.69	0.5	HKY + $\Gamma$
<i>tpi</i>	70	402	21.4 <sup>*</sup>	34.45	44.62	0.5	GTR + $\Gamma$
<i>tpi</i> <sup>4</sup>	14	402	22 <sup>**</sup>	5.66	5.63	3.9	GTR + $\Gamma$
<i>yqiL</i>	60	516	0	22.3	24.77	0	HKY + $\Gamma$
<i>yqiL</i> <sup>4</sup>	16	516	0	5.73	5.68	0	HKY + $\Gamma$

Table 1 (Continued)

Locus	Alleles	Sites	$\Gamma$	$\Gamma_{wi}$	$\Gamma_{wF}$	$\Gamma/\Gamma_{wF}$	Model
	46–80	402–516	0–21.4	13.2–35.04	14.16–44.62	0–0.8	
	<b>58</b>	<b>457</b>	<b>5.1</b>	<b>22.59</b>	<b>26.36</b>	<b>0.2</b>	
<i>Staphylococcus epidermidis</i>							
<i>arcC</i>	9	465	3.7*	7.36	7.44	0.5	HKY
<i>aroE</i>	9	420	0	6.62	6.72	0	F81
<i>glpK</i>	11	468	0	11.61	12.17	0	HKY
<i>gmk</i>	9	465	0	7.73	7.91	0	HKY + $\Gamma$
<i>pta</i>	7	477	0	9.38	9.54	0	F81
<i>tpiA</i>	9	408	4.5	3.68	3.67	1.2	F81
<i>yqiL</i>	5	474	0	7.68	7.58	0	F81
	5–11	408–477	0–4.5	3.68–11.61	3.67–12.17	0–1.2	
	<b>8</b>	<b>454</b>	<b>1.2</b>	<b>7.72</b>	<b>7.86</b>	<b>0.2</b>	
<i>Streptococcus pneumoniae</i>							
<i>aroE</i>	59	405	6.4*	11.41	12.15	0.5	HKY + $\Gamma$
<i>aroE</i> <sup>5</sup>	6	405	26*	2.19	2.19	12.8	HKY + $\Gamma$
<i>ddl</i>	141	441	91***	26.65	31.75	2.9	TrN + $\Gamma$ + I
<i>ddl</i> <sup>5</sup>	9	441	100**	4.05	4.05	25.2	TrN + $\Gamma$ + I
<i>gdh</i>	83	460	30***	17.23	18.86	1.6	HKY + $\Gamma$
<i>gdh</i> <sup>5</sup>	10	460	13*	4.24	4.24	3.1	HKY + $\Gamma$
<i>gki</i>	96	483	94.7***	20.05	22.7	4.2	TrN + $\Gamma$ + I
<i>gki</i> <sup>5</sup>	9	483	47**	11.04	11.59	4.1	TrN + $\Gamma$ + I
<i>recP</i>	62	450	98.6**	14.69	15.75	6.3	TrN + $\Gamma$
<i>recP</i> <sup>5</sup>	8	450	100**	3.09	3.15	31.7	TrN + $\Gamma$
<i>spi</i>	91	474	100***	20.66	23.23	4.3	TrN + $\Gamma$ + I
<i>spi</i> <sup>5</sup>	11	474	34***	7.17	7.17	4.8	TrN + $\Gamma$ + I
<i>xpt</i>	131	486	79.5***	22.21	25.27	3.1	HKY + $\Gamma$
<i>xpt</i> <sup>5</sup>	14	486	31**	5.35	5.35	5.8	HKY + $\Gamma$
	59–141	405–486	6.4–100	11.41–26.65	12.15–31.75	0.5–6.3	
	<b>95</b>	<b>457</b>	<b>71.5</b>	<b>18.99</b>	<b>21.39</b>	<b>3.3</b>	
<i>Streptococcus pyogenes</i>							
<i>gki</i>	86	498	10.8***	21.09	23.9	0.5	TrN + $\Gamma$
<i>gtr</i>	64	450	0	24.32	27.9	0	TrN + $\Gamma$
<i>murI</i>	57	438	73.5**	10.63	11.39	6.5	HKY + $\Gamma$
<i>mutS</i>	46	405	1.8	11.83	12.56	0.1	HKY + $\Gamma$
<i>recP</i>	73	459	59.7***	15.64	16.98	3.5	TrN + $\Gamma$ + I
<i>xpt</i>	57	450	74.4***	12.58	13.5	5.5	HKY + $\Gamma$
<i>yqiL</i>	53	434	43.2**	11.24	12.15	3.6	HKY + $\Gamma$
	46–86	405–498	0–74.4	10.63–24.32	11.39–27.9	0–6.5	
	<b>62</b>	<b>448</b>	<b>37.6</b>	<b>15.33</b>	<b>16.91</b>	<b>2.8</b>	
<i>Vibrio vulnificus</i>							
<i>dtdS</i>	46	417	10.1*	12.74	13.76	0.7	TrNef + $\Gamma$ + I
<i>glp</i>	38	480	19.2***	10.95	11.52	1.7	TIMef + $\Gamma$ + I
<i>gyrB</i>	31	459	31.3**	8.51	8.72	3.6	TrNef + $\Gamma$ + I
<i>lysA</i>	41	465	22.2***	18.23	20	1.1	TrNef + $\Gamma$ + I
<i>mdh</i>	29	489	16.2*	7.64	7.82	2.1	K80 + $\Gamma$ + I
<i>metG</i>	31	429	8.1**	9.26	9.87	0.8	K80 + $\Gamma$ + I
<i>pntA</i>	32	396	5.1**	8.69	9.11	0.6	TrNef + $\Gamma$ + I
<i>purM</i>	28	444	9.1**	10.02	10.66	0.9	K80 + $\Gamma$
<i>pyrC</i>	35	423	9.1**	12.14	13.11	0.7	K80 + $\Gamma$ + I
<i>tnaA</i>	32	324	7.1	10.43	11.02	0.6	K80 + $\Gamma$
	28–46	324–489	5.1–31.3	7.64–18.23	7.82–20	0.6–3.6	
	<b>34</b>	<b>433</b>	<b>13.8</b>	<b>10.86</b>	<b>11.56</b>	<b>1.3</b>	

$\Gamma$  could not be estimated in *C. albicans* because of nucleotide ambiguities. Range and mean (bold) estimates are also indicated for each parameter based only on the database sequences.  $\Gamma$  and  $\Gamma$  estimates per site can be obtained dividing both parameters by the number of sites.

<sup>1</sup> Meats et al. (2003); <sup>2</sup> Achtman et al. (1999); <sup>3</sup> Maiden et al. (1998); <sup>4</sup> Enright et al. (2000); <sup>5</sup> Hanage et al. (2004). <sup>eca</sup> encapsulated; <sup>nca</sup> noncapsulated. Model abbreviations in alphabetical order: F81 (Felsenstein, 1981), GTR (Tavaré, 1986), HKY (Hasegawa et al., 1985), JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), K81 (Kimura, 1981), K81uf (K81 unequal base frequencies; Posada and Crandall, 1998), SYM (Zharkikh, 1994), TIM (Tavaré, 1986), TIMef (TIM with equal base frequencies; Posada and Crandall, 1998), TrN (Tamura and Nei, 1993), TrNef (TrN equal base frequencies; Posada and Crandall, 1998), and TVM (Tavaré, 1986).  $\Gamma$ : shape parameter of the gamma distribution;  $I$ : proportion of invariable sites.

\*  $P < 0.05$ .

\*\*  $P < 0.01$ .

\*\*\*  $P < 0.001$ .

*Staphylococcus aureus* (Enright et al., 2000), and *Streptococcus pneumoniae* (Hanage et al., 2004). Although most of the isolates analyzed here were collected worldwide, others actually represent local populations (*Neisseria gonorrhoeae*, *S. aureus*, *S. pneumoniae*) and one is from asymptomatic carriage (*S. pneumoniae*).

Sequences were aligned in Clustal X (Thompson et al., 1997) and then translated into amino acids using the universal reading frame in MacClade 4.05 (Maddison and Maddison, 2000). Haplotypes including stop codons were deleted from the analysis (e.g., the *ndh* locus from *Burkholderia pseudomallei*).

Models of nucleotide and codon substitution were assessed using the maximum likelihood approach described by Huelsenbeck and Crandall (1997) and Posada and Crandall (1998). Likelihood scores for each model were estimated in PAUP\* 4.0b10 (Swofford, 2003) and then compared through a series of hierarchical likelihood ratio tests (LRT) to determine the best-fit model. When two models are nested, twice the log-likelihood difference will be compared with a  $\chi^2$  distribution with the degrees of freedom  $\nu$  equal to the difference in the number of parameters between the two models. Recent simulation studies have shown that this approach performs very well at recovering the true underlying model of evolution (Yang et al., 2000a; Posada, 2001; Posada and Crandall, 2001a; Anisimova et al., 2001).

## 2.2. Genetic analysis

Population recombination ( $\rho$ ), population mutation ( $\Theta$ ), and molecular adaptive selection were estimated independently for each gene region and species.

### 2.2.1. Population recombination rate ( $\rho$ )

Within each gene region  $\rho$  was estimated using the standard likelihood coalescent approach implemented in the LDhat package (McVean et al., 2002). Within this framework,  $\rho$  can be expressed as  $4N_e r$  in diploid organisms (crossing-over model), where  $N_e$  is the inbreeding effective population size and  $r$  is the recombination rate per locus per generation, or as  $8N_e \bar{c} t$  in haploid organisms (gene-conversion model), where  $\bar{c}$  is the per base rate of initiation of gene conversion and  $t$  is the average gene conversion tract length. This method has the desirable property of relaxing the infinite-sites assumption (typically violated by many empirical data sets (Posada et al., 2002)) and accommodates different models of molecular evolution (including, importantly, rate heterogeneity). LDhat implements a composite-likelihood estimate of  $\rho$ , which has the advantage of being more computationally efficient relative to full-likelihood methods, but without summarizing the data in a single statistic (Hudson, 2001). In addition, LDhat includes a powerful likelihood permutation test (LPT) to test the hypothesis of no recombination ( $\rho = 0$ ). This method has proven to be more powerful than previous permutation-

based methods for detecting recombination (McVean et al., 2002), thus we will also apply it in our analyses.

### 2.2.2. Population mutation rate

A coalescent estimate (no recombination) of  $\Theta$  for haploids ( $2N_e\mu$ ) and diploids ( $4N_e\mu$ ) where  $\mu$  is the mutation rate per locus per generation was calculated using the statistical method of Watterson (1975) as implemented in LDhat. This program generates an estimate of  $\Theta$  based on the number of segregating sites in the sequences assuming an infinite-sites (i.e., mutations only occur once per site in a population) or a finite-sites model. By comparing both estimates, we will be able to draw inferences about the mutational process (e.g., lower estimates of  $\Theta$  under the infinite-sites model will indicate occurrence of multiple mutations at some sites). Other more powerful maximum likelihood approaches to estimate  $\Theta$  have been proposed (Kuhner et al., 1995, 1998), but these methods require a bifurcating phylogenetic tree, are computationally intense, and are more easily affected by the presence of recombination in the data (M.K. Kuhner, personal communication). Moreover, Fu and Li (1993) and Felsenstein (2004) have shown that Watterson's estimator, although less efficient than maximum likelihood, is remarkably good.

### 2.2.3. Adaptive selection

The effect of natural selection is usually studied by comparing the fixation rates of nonsynonymous (amino acid-altering) and synonymous (silent) mutations within a maximum likelihood phylogenetic framework (Yang et al., 2000a). A measure that has featured prominently in such studies is the nonsynonymous/synonymous substitution rate ratio ( $\omega = d_N/d_S$ ) or acceptance rate (Miyata and Yasunaga, 1980).  $\omega$  measures the selective pressure at the protein level, with  $\omega = 1$  meaning neutral mutations,  $\omega < 1$  purifying selection, and  $\omega > 1$  diversifying positive selection. We initially estimated  $\omega$  per site for all data sets using the codon-based nested models M1 (neutral), M2 (selection) and M3 (discrete) of Yang et al. (2000a). Those genes under positive selection were then examined under models M7 (beta) and M8 (beta and  $\omega$ ). Model likelihood scores were compared using a LRT as described before. M2 (3 parameters) and M3 (5 parameters) are more general than model M1 (1 parameter) and can be compared with M1. Similarly, M7 (2 parameters) is a special case of model M8 (4 parameters) and can be compared the same way. When  $\omega > 1$  in M2, M3, or M8 positively selected sites are inferred from the data. We also applied the empirical Bayesian approach implemented by Nielsen and Yang (1998) to identify the potential sites under diversifying selection as indicated by a posterior probability ( $pP$ )  $> 0.95$ . Sites where  $pP$  is lower than this value will not be reported. Finally, for comparative purposes, we also estimated  $\omega$  per gene using the Goldman and Yang (1994) model. All of the previous analyses were carried out in PAML 3.14b3 (Yang, 1997) and were performed under

initial  $\omega$  values  $>1$  and  $<1$ , as recommended by the author. If positive selection was detected, we reran PAML several times to check convergence. Here, we reported the estimates obtained under the best likelihood scores.

Maximum likelihood and Bayesian inferences under codon-substitution models relies on the phylogenetic relationships among the sequences and do not account for the presence of recombination. Empirical results reported by Yang et al. (2000a) and simulations by Anisimova et al. (2001, 2002) indicated that the LRTs and the inference of sites under positive selection do not seem to be sensitive to the assumed tree topology (a neighbor-joining tree in our analyses), even if a star tree is used. Hence, presumably, our results are not biased by whichever phylogenetic process (clonal, epidemic, or panmictic) drives the population structure of the studied pathogens. Nevertheless, to test this hypothesis, values of  $\omega > 1$  were re-estimated using alternative maximum parsimony trees generated in PAUP\*. High levels of recombination, however, seem to affect dramatically the accuracy of the LRT test and often recombination is mistaken as evidence of positive selection (simulations by Anisimova et al., 2003; although see Urwin et al., 2002 for a different opinion). Anisimova et al. (2003) showed that LRTs of M0–M3 and M1–M2 are heavily affected, but LRT of M7–M8 is much less (positive selection was falsely detected in only 20% of replicates at  $\alpha = 5\%$ ). Identification of sites under positive selection using the Bayesian approach appears to be less influenced by high levels of recombination. The Bayesian method predicted incorrectly  $\sim 25\%$  of the sites for M3,  $\sim 9\%$  for M8, and  $\sim 5\%$  for M2. However, when data were simulated at high levels of positive selection ( $\omega = 6$ ), Bayes's site prediction becomes more accurate and powerful (concrete values are not reported).

McClellan et al. (2005) have recently shown that  $d_N/d_S$  ratios are less sensitive to detecting single adaptive amino acid changes than methods that evaluate positive selection in terms of the amino acid properties, which comprise protein phenotypes that selection at the molecular level may act upon. Hence, in addition to estimating adaptive selection under codon-substitution models M2, M3, and M8, we also estimated adaptive selection in terms of 31 quantitative biochemical properties using the model of McClellan and McCracken (2001) as implemented in TreeSAAP 3.2 (Woolley et al., 2003). No study has shown how tree topology and recombination affect the performance of the amino acid-property-based models implemented in TreeSAAP. For the case of recombination, intuitively one could expect TreeSAAP to be less affected than PAML since the former infers selection at the phenotype level, hence its accuracy is independent of the force generating molecular change (mutation or recombination), and what really matters is if that physicochemical change is fixed or not (D.A. McClellan, personal communication). We will test all data sets under positive selection according to PAML using the protein model implemented in TreeSAAP. Based on a phylogenetic tree, this model establishes first a chronology

of observable molecular evolutionary events. The frequency of these events are then analyzed in order to identify (1) amino acid properties that may have radically changed more often than expected by chance (presumably due to selection promoting the occurrence of radical amino acid replacements) and (2) amino acid sites associated with selection, thus establishing a correlation between the sites of positive selection and the structure and function of the protein. We followed the general procedure outlined in McClellan et al. (2005). In this study, we are particularly interested in detecting molecular adaptation, selection that results in radical structural or functional shifts in local regions of the protein. To this end, the range of possible changes in an amino acid property was divided into eight magnitude categories, with numbers 6, 7, and 8 denoting radical changes. An amino acid property is said to be affected by adaptive selection (referred to as positive-destabilizing selection) when the frequency of changes in magnitude categories 6–8 significantly exceed the frequency (or frequencies) expected by chance, as indicated by  $z$ -scores  $> 2.326$  ( $P < 0.01$ ). Particular amino acid residue sites affecting those properties were then also identified by  $z$ -scores  $> 2.326$ .

### 3. Results and discussion

#### 3.1. Species comparisons

Evolutionary models chosen by the LRT, population recombination rates per locus ( $\rho$ ) and the probability of  $\rho = 0$  (indicated by asterisks) from the LPT, population mutation rates per locus using Watterson's method under infinite- ( $\Theta_{WT}$ ) and finite-sites models ( $\Theta_{WT}$ ), and ratio of recombination to mutation ( $\rho/\Theta_{WT}$ ), for every species and locus are presented in Table 1. No single available model in Modeltest best fit all the data and almost all possible models were chosen as most appropriate for one or more data sets. HKY (Hasegawa et al., 1985) and TrN (Tamura and Nei, 1993) models were chosen more often, but highly diverse data sets (large  $\rho$  and  $\Theta$ ) such as those of *H. pylori* required more complex models (TIM and GTR) to accommodate the observed variation. Most data sets presented rate heterogeneity (i.e., the evolutionary process exhibits site-to-site variation) as accounted for by the  $\Gamma$  distribution, and a fraction of invariable sites (sites incapable of accepting substitutions). Hence, both parameters should be incorporated as part of the evolutionary model for inferring phylogenetic relationships when using model-based tree-building methods such as neighbor-joining, maximum likelihood or Bayesian inference. Violation of this assumption can have devastating consequences. Different models fit the same gene in different species; however, the same model fit multiple genes in some pathogens (e.g., *Bacillus cereus*, *E. coli*, *H. influenzae*, and *H. pylori*).

As expected population recombination and population mutation rates and levels of adaptive selection varied greatly

between and within taxa, but some general trends can be observed. In the next section, we will describe them separately.

### 3.1.1. Population recombination rate ( $\rho$ )

*H. pylori*, *N. gonorrhoeae*, *N. meningitidis*, and *S. pneumoniae* showed high mean levels ( $\rho > 50$ ) of intragenic recombination across loci, which supports prior conclusions (e.g., Maynard-Smith et al., 1993; Suerbaum et al., 1998; Feil et al., 1999, 2000a, 2001). *B. cereus*, *H. influenzae*, *Streptococcus agalactiae*, and *Streptococcus pyogenes* showed moderate levels of recombination ( $15 < \rho \leq 50$ ). Interestingly, this second species group contained some gene regions that recombine frequently whilst others do not. This could be due to variable selective pressures on the genome and/or temporal/geographical structuring generated by random genetic drift, which would not be surprising considering the wide distribution and temporal dispersion of the isolates analyzed. These data support previous conclusions for low rates of recombination for *B. cereus* (Vilas-Boas et al., 2002), *S. pyogenes* (Enright et al., 2001; Feil et al., 2001) and *S. agalactiae* (Jones et al., 2003). Finally, *B. pseudomallei* (and closely related species), *M. catarrhalis*, *Staphylococcus epidermidis*, *Vibrio vulnificus*, *Campylobacter jejuni*, *Enterococcus faecium*, *E. coli*, and *S. aureus* showed consistently low mean levels of  $\rho$  ( $\leq 15$ ). Little information has been published on the first four of these species, but clonal (low recombination) and epidemic (sexual but superficially clonal) population structures have been proposed for *C. jejuni* (Suerbaum et al., 2001) and *E. faecium* (Homan et al., 2002). The frequency of recombination in *E. coli*, *S. aureus*, and *H. influenzae* is still debated: some studies suggest low rates or clonal structures (Whittam, 1995; Feil et al., 2001, 2003), while others indicate the opposite (Feil et al., 1999, 2001; Meats et al., 2003). Our results show low mean  $\rho$  rates for *E. coli* and *S. aureus* and a moderate rate for *H. influenzae*. It was surprising to find a value of  $\rho = 0$  for all gene regions in *E. coli* (Table 1). LDhat estimates intragenic recombination and will estimate  $\rho = 0$  if break points are distributed between the gene regions. Other commonly used approaches, however, are aimed to detect both intragenic recombination and allele replacement (Feil et al., 1999) or allele replacement (Maynard-Smith and Smith, 1998); hence, rate differences between our study and previous work (e.g., Feil et al., 1999) could be expected. Furthermore, all these methods differ significantly in their relative abilities to detect recombination, which may give them high false positive rates (Posada and Crandall, 2001b). A more detailed comparison among and within clonal complexes seems necessary to assess the role of recombination in *E. coli*.

We investigated whether our results depend on sample size by analyzing multiple subsets of data from five species including high, medium, and low recombinant taxa. These analyses yielded comparable mean  $\rho$  values within each species, indicating that LDhat estimates of this parameter

are not strongly affected by sample size (see other examples by Jolley et al., 2000; Maggi-Solcà et al., 2001; Feil et al., 2003; Viscidi and Demma, 2003). However, many MLST data sets represent biased samples that are concentrated on disease isolates and confirmation of our results with more population-based samples is desirable.

Based on the observed levels of population recombination, we could tentatively categorize the population structure of the studied pathogens as follows: the first and second groups of highly and moderate recombinant taxa, respectively, would conform to a panmictic or nonclonal model. We note that for almost all loci with  $\rho > 5$ , LDhat significantly rejected the alternative hypothesis of no recombination. The third group of species does not recombine or recombine only rarely; these taxa conform to a clonal (or almost clonal) model. Within a phylogenetic framework, the population structure of the panmictic group might be best described by a network approach (e.g., Posada and Crandall, 2001c). In contrast, a bifurcating tree could be used for the clonal species. The structure of the panmictic species including recombinant and nonrecombinant loci could be also assessed using a tree-based approach if the recombinant loci are excluded from the analysis. Genes with low levels of recombination according to LDhat could be concatenated prior to a phylogenetic analysis under a single model of evolution. Alternatively, gene-specific substitution models (i.e., mixed models) could be used for each gene region using a Bayesian approach in order to maximize the phylogenetic signal in the data. As an example, we have compared the minimum evolution trees obtained by Meats et al. (2003) using a K80 model after concatenating seven genes from encapsulated (eca) and nonencapsulated (nca) *H. influenzae* isolates with the results under the best fit model (HKY +  $\Gamma$  + *I*) after excluding the two genes (*mdh* and *pgi*) with the highest recombination rates ( $\rho > 65$ ). Nodal support using 1000 bootstrap replicates (Felsenstein, 1985) was higher for both data sets with trees based on the five concatenated genes (Fig. 1) and different relationships were indicated.

### 3.1.2. Population mutation rate ( $\Theta$ )

Overall, species with higher average number of alleles ( $n_a$ ) also showed higher average  $\Theta$  values ( $r \approx 0.59^*$ ), but this correlation is clearly altered by the amount of recombination in the data. For example, *H. pylori*, *N. meningitidis*, and *S. pneumoniae*, which have high mean  $n_a$  (95–323) and also high mean  $\rho$  ( $> 52$ ), showed similar average  $\Theta$  values to other species with clearly less mean  $n_a$  such as *E. coli* (44 alleles) or *S. aureus* (58 alleles), but also low mean  $\rho$  ( $< 6$ ). The correlation between mean  $n_a$  and  $\Theta$  increased significantly if these three species are deleted from the comparison ( $r \approx 0.69^{**}$ ). This indicates that in the former three species punctual mutation is not the major evolutionary force generating allelic variation (see below). Subsets of isolates with a worldwide distribution showed similar  $\Theta$  values to their corresponding full data sets.

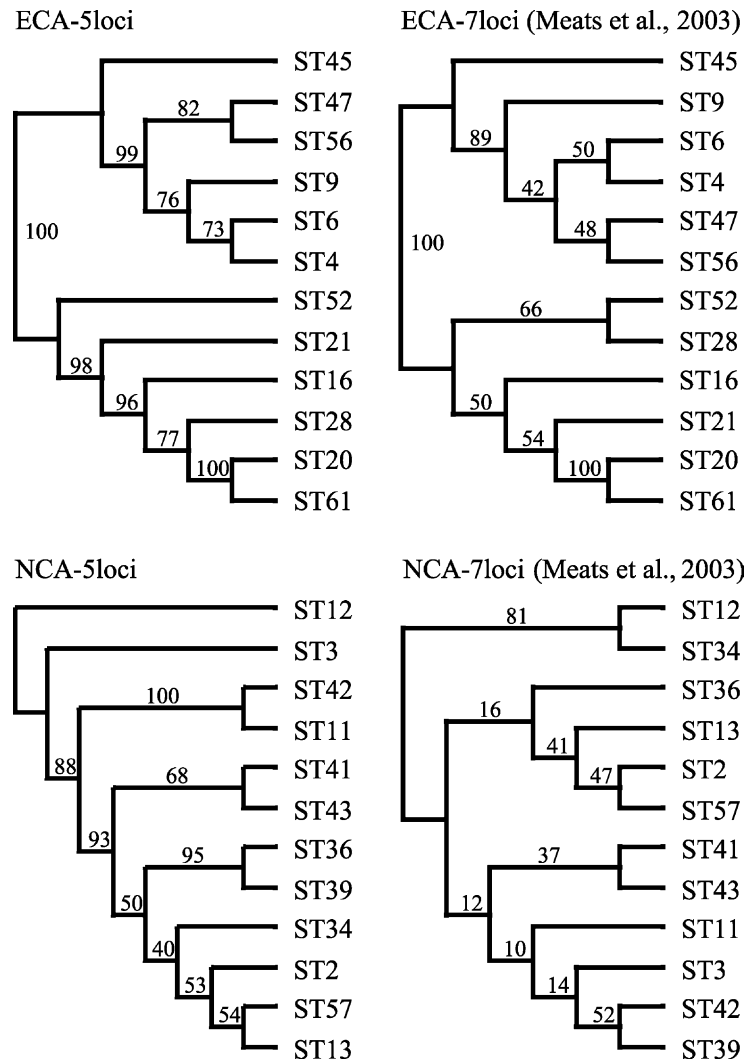


Fig. 1. Minimum evolution (ME) trees of encapsulated (eca) and noncapsulated (nca) isolates of *Haemophilus influenzae* using five low recombinant ( $\rho \leq 20$ ) genes. ME trees from Meats et al. (2003) using five low and two highly recombinant ( $\rho > 65$ ) genes are also depicted for comparison. ST: sequence type.

However, as expected, local isolates from *S. aureus* and *S. pneumoniae* showed lower  $\Theta$  values presumably due to a more homogeneous environment and recent shared evolutionary history. For these same two species,  $\rho$  rates between locally and widely dispersed isolates varied less. *S. aureus* seems to be an almost clonal taxa, thus differences in  $\rho$  rates were not expected, but in the case of *S. pneumoniae* this last result can be explained based on the molecular differences existing between both evolutionary processes. Recombination reshuffles existing variation generated by mutation and can potentially create new variants without novel mutations. Thus, at the local population level where events are more recent in evolutionary history high levels of  $\rho$  can be seen even with little variation in  $\Theta$ .

Encapsulated and noncapsulated *H. influenzae* isolates (Table 1) did not show notable variation in average  $\Theta$  or  $\rho$  rates:  $\Theta_{\text{Weca}} = 13.22$  versus  $\Theta_{\text{Wnca}} = 11$  and  $\rho_{\text{eca}} = 28.6$  versus  $\rho_{\text{nca}} = 23.6$ . However, ME phylogenetic trees (Fig. 1) of noncapsulated isolates were more weakly supported than

those of encapsulated trees (even after removing *pgi* and *mdh*), suggesting that the impact of recombination may be greater in the former than in the latter group, as reported by Meats et al. (2003).

All  $\Theta$  estimates under the finite-sites model ( $\Theta_{\text{Wf}}$ ) were higher than those generated under the infinite-sites model ( $\Theta_{\text{Wi}}$ ). Differences varied based on the amount of genetic variation, but in some loci such as *gdh* from *N. meningitidis* recurrent mutation (i.e., some sites experiencing multiple mutations in the history of the sample) increased  $\Theta$  by up to 39%. This stresses the need for using evolutionary models that relax the infinite-site assumption, such as those incorporated in LDhat, because recurrent mutation can generate patterns of genetic variability that resemble the effects of recombination (McVean et al., 2002).

The ratio between recombination and mutation is indicative of the contribution of each factor to the emergence of variant alleles (Feil et al., 1999, 2000a,b). Our results, as indicated by the mean  $\rho/\Theta_{\text{Wf}}$  ratio, showed that recombina-

tion generates more divergence than mutation in nine taxa (mean  $\rho/\Theta_{WF} > 1.0$ ) and less in seven cases (mean  $\rho/\Theta_{WF} < 1.0$ ). As expected, taxa with moderate or high levels of recombination showed greater  $\rho/\Theta_{WF}$  values, but results varied among loci ranging from 0 to  $\sim 17$ . Nevertheless, we note that  $\rho = 100$  was chosen as a cutoff as it is the limit for which likelihoods were estimated. This means that  $\rho > 100$  could be expected for those loci with  $\rho = 100$ . Consequently, the extent of the differences between the contribution of recombination and mutation to diversity may be greater than reported for those species with high levels of  $\rho$  (close to 100), but over all taxa, one factor does not seem to prevail over the other.

We can test the hypothesis that recombination has a major impact in leading to genetic diversity across species and loci by examining the correlation of genetic diversity (as measured by the  $\Theta_{WF}$  estimator) and recombination rate. By looking at Table 1 we observe that species with similar mean  $\Theta_{WF}$  values (independently of  $n_a$ ) such as *B. cereus* and *H. influenzae* or *H. pylori* and *S. aureus* differ in their mean  $\rho$  values. Furthermore, over all taxa or loci,  $\Theta_{WF}$  and  $\rho$  are clearly not correlated ( $r = -0.16$  and  $-0.02$ , respectively) as shown in Fig. 2a. Hence, in general, we can conclude that genetic diversity and recombination are not correlated, which supports our previous conclusion that recombination does not prevail over mutation in generating diversity.

### 3.1.3. Adaptive selection

Values of the  $d_N/d_S$  ratio per gene ( $\omega_{M0}$ ) were  $< 1$  for all species and loci except locus *abcZ* in *N. gonorrhoeae* (data

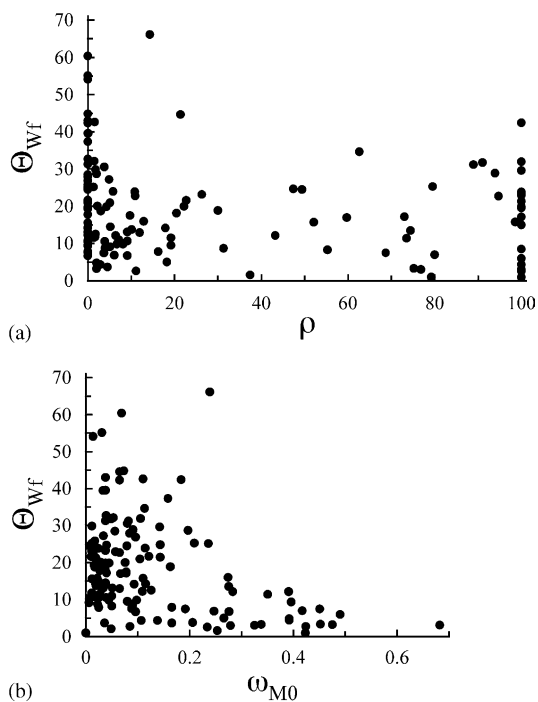


Fig. 2. Scattergrams of population recombination rates (a) and acceptance rates (b) and population mutation rates per locus. The locus *abcZ* from *N. gonorrhoeae* is not included in the scattergram (b).

not shown). Hence, on average, most loci and species seem to be under purifying selection. This has been confirmed in almost every genetic analysis of MLST sequences (e.g., Dingle et al., 2001; Feil et al., 2003; Meats et al., 2003). No apparent connection seems to exist between and  $\Theta_{WF}$  rates, as reflected by the observed low correlation ( $r = -0.29$ ) between both parameters (Fig. 2b). Thus there seems to be minimal impact of selection on genetic diversity due to the general lack of positive selection. Most variation within genes that encode essential metabolic enzymes, such as the housekeeping genes, is likely to be selectively neutral or deleterious (Li, 1997; Feil et al., 2000a). Adaptive evolution, if present, must be punctual. Hence, the criterion that this average  $\omega_{M0} > 1$  is a very stringent one for detecting adaptive selection (Crandall et al., 1999). Analyses of 91 housekeeping gene regions using models that account for  $\omega$  heterogeneity among sites have identified 13, 33, and 28 loci under significant positive selection as indicated by the LRTs of M1–M2, M1–M3, and M7–M8, respectively, and number of potential sites  $n_M \neq 0$  (Table 2). Under LRTs of M7–M8 (the most conservative model), all the species but *B. pseudomallei*, *C. jejuni*, *S. epidermidis*, and *V. vulnificus* seem to experience adaptive selection for one (e.g., *B. cereus*) to seven (*N. gonorrhoeae*) loci. The number of potential sites under diversifying selection ( $n_M$ ), as identified by the Bayesian approach, ranged from one (e.g., *pta* locus from *B. cereus*) to nine (*gpdh* from *N. gonorrhoeae*). All these sites were also found by TreeSAAP ( $n_{TS}$ ; Table 2) using a completely different procedure. Moreover, for most of the genes, additional sites under positive selection were found, which confirms that  $d_N/d_S$  ratios are not very sensitive to detecting adaptive selection in genes under low or moderate levels of diversifying selection (McClellan et al., 2005).

Acceptance rates and detected number of sites ( $n_M$ ) under positive selection diminished in the subsets compare to the full data sets. TreeSAAP, in contrast, still showed evidence of significant ( $P < 0.01$ ) destabilizing selection ( $n_{TS}$ ) in almost all of the same gene regions, although at a lower level (Table 2). This difference again reaffirms the higher sensitivity of the evolutionary model of McClellan and McCracken (2001) for detecting adaptive selection. As reported before by Anisimova et al. (2001, 2002), both power and accuracy of the LRT and Bayes tests decrease as sample size diminishes, especially when the sequences are highly similar. Both encapsulated and nonencapsulated isolates of *N. influenzae* showed evidence of adaptive selection, although no clear differences in selective pressure between them were observed. Interestingly, the amino acid sites and physicochemical properties under destabilizing selection (TreeSAAP) varied between both groups (Table 3).

Simulations by Anisimova et al. (2003) questioned the efficiency of  $d_N/d_S$  for detecting positive selection under high levels of recombination, such as those observed in some of our data sets (e.g., *B. cereus*), since this force may inflate  $\omega$  and  $n_M$  estimates. Nevertheless, in some of the MLST

Table 2

Acceptance rate per site ( $\omega_{M2}$ ,  $\omega_{M3}$ , and  $\omega_{M8}$ ) and proportion of sites ( $p_{M2}$ ,  $p_{M3}$ , and  $p_{M8}$ ) under models M2 (selection), M3 (discrete), and M8 (beta and  $\omega$ ) with a  $\omega > 1$ , and number of sites under positive (or destabilizing) selection with a posterior probability  $> 0.95$  ( $n_{M2}$ ,  $n_{M3}$ , and  $n_{M8}$ ) and a z-score  $> 2.326$ ,  $P < 0.01$ , ( $n_{TS}$ )

Locus	$\omega_{M2}$	$p_{M2}$ (%)	$n_{M2}$	$\omega_{M3}$	$p_{M3}$ (%)	$n_{M3}$	$\omega_{M8}$	$p_{M8}$ (%)	$n_{M8}$	$n_{TS}$
<i>Bacillus cereus</i>										
<i>pta</i>	–	–	–	1.46**	0.8	1	1.46***	0.8	1	3
<i>Candida albicans</i>										
<i>adp1</i>	6.93*	3.8	3	6.93	3.8	3	6.93*	3.8	3	2
<i>gln4</i>	18.42*	1.4	1	18.53	1.6	1	18.2*	1.6	1	4
<i>vps13</i>	–	–	–	5.05*	7.6	3	–	–	–	5
<i>Campylobacter jejuni</i>										
<i>pgm</i>	–	–	–	1.45***	2.8	4	–	–	–	4
<i>Escherichia coli</i>										
<i>adk<sup>MA</sup></i>	–	–	–	1.94***	0.6	1	1.97*	0.6	1	4
<i>mltD</i>	–	–	–	8.68***	0.9	2	8.53***	0.9	3	7
<i>pgi</i>	–	–	–	2.81*	1.8	2	2.85*	1.7	2	6
<i>Enterococcus faecium</i>										
<i>pstS</i>	–	–	–	1.97*	1.7	1	1.94*	1.7	1	6
<i>Haemophilus influenzae</i>										
<i>adk</i>	–	–	–	1.33*	3.4	4	1.33*	3.4	4	6
<i>adk<sup>1</sup></i>	–	–	–	1.29	2.8	3	1.35	2.5	2	2
<i>adk<sup>1eca</sup></i>	–	–	–	1.88	0.6	2	5.67*	0.7	1	2
<i>adk<sup>1nca</sup></i>	1.39	3.9	3	1.39	3.9	3	1.39	3.9	3	3
<i>atpG</i>	–	–	–	1.4*	3.1	1	1.41	3.1	1	5
<i>atpG<sup>1</sup></i>	–	–	–	–	–	–	–	–	–	3
<i>atpG<sup>1eca</sup></i>	2.4	5.9	6	2.4	5.9	6	2.4	5.9	6	2
<i>atpG<sup>1nca</sup></i>	–	–	–	–	–	–	–	–	–	2
<i>Helicobacter pylori</i>										
<i>vacA</i>	10.73	29.8	2	2.39***	1.4	2	2.4*	1.3	2	3
<i>vacA<sup>2</sup></i>	–	–	–	1.88**	8.9	7	1.9***	6.5	7	12
<i>Moraxella catarrhalis</i>										
<i>adk</i>	–	–	–	2.9*	3.7	3	–	–	–	5
<i>fumC</i>	3.78	1.0	–	2.09	2.2	3	3.51*	1.4	2	4
<i>mutY</i>	–	–	–	2.67***	7.6	7	3.04***	6.1	5	14
<i>Neisseria gonorrhoeae</i>										
<i>abcZ</i>	106.37*	2.4	3	103.32	2.4	3	106.2*	2.4	3	2
<i>gnd</i>	13.44*	1.0	3	13.44	1.0	3	13.51*	1.0	3	2
<i>gpdh</i>	22.29**	5.7	9	43.97*	2.2	2	22.4***	5.5	2	8
<i>pgi2</i>	45.15***	0.89	4	98.69***	0.3	1	49.15***	0.9	4	1
<i>pip</i>	6.52*	7.2	6	6.52	7.2	6	6.52*	7.2	6	4
<i>ppk</i>	13.38*	3.5	4	13.38	3.5	4	13.45*	3.4	4	1
<i>serC</i>	3.23**	10.2	2	3.27*	10.1	2	3.25*	10.2	2	2
<i>Neisseria meningitidis</i>										
<i>adk</i>	–	–	–	1.21***	2.6	3	1.3**	2.2	2	7
<i>adk<sup>3</sup></i>	–	–	–	–	–	–	–	–	–	2
<i>pdhC</i>	–	–	–	1.11***	5.2	7	–	–	–	9
<i>pdhC<sup>3</sup></i>	–	–	–	–	–	–	–	–	–	–
<i>Streptococcus agalactiae</i>										
<i>adhP</i>	21.43*	0.75	1	19.62	8.1	1	19.65*	0.8	1	6
<i>Staphylococcus aureus</i>										
<i>aroE</i>	21.78***	1.3	1	11.17*	1.9	2	11.2*	1.8	2	6
<i>aroE<sup>4</sup></i>	–	–	–	9.07	1.5	1	9.09	1.5	1	1
<i>glpF</i>	–	–	–	3.06***	0.83	1	–	–	–	2
<i>glpF<sup>4</sup></i>	–	–	–	–	–	–	–	–	–	3
<i>gmk</i>	–	–	–	1.02***	3.6	1	–	–	–	3
<i>gmk<sup>4</sup></i>	–	–	–	–	–	–	–	–	–	–
<i>yqiL</i>	–	–	–	1.77***	4.2	2	–	–	–	6
<i>yqiL<sup>4</sup></i>	–	–	–	6.37	0.8	1	6.41	0.8	1	5

Table 2 (Continued)

Locus	$\omega_{M2}$	$P_{M2}$ (%)	$n_{M2}$	$\omega_{M3}$	$P_{M3}$ (%)	$n_{M3}$	$\omega_{M8}$	$P_{M8}$ (%)	$n_{M8}$	$n_{TS}$
<i>Streptococcus pneumoniae</i>										
<i>aroE</i>	–	–	–	3.79*	7.5	2	4.21*	6.1	2	9
<i>aroE</i> <sup>5</sup>	–	–	–	–	–	–	–	–	–	–
<i>gdh</i>	–	–	–	2.55***	2.2	2	–	–	–	4
<i>gdh</i> <sup>5</sup>	–	–	–	–	–	–	–	–	–	2
<i>gki</i>	–	–	–	3.19***	0.7	1	–	–	–	9
<i>gki</i> <sup>5</sup>	8.80	0.6	1	8.53	0.7	1	8.45*	0.7	1	2
<i>xpt</i>	–	–	–	1.81***	6.3	3	1.80***	6.3	2	7
<i>xpt</i> <sup>5</sup>	8.26*	3.5	1	8.6	3.4	1	8.58*	3.4	1	1
<i>Streptococcus pyogenes</i>										
<i>gtr</i>	–	–	–	1.97***	5.3	5	–	–	–	5
<i>murI</i>	–	–	–	3.1*	7.6	4	–	–	–	5
<i>xpt</i>	6.81*	1	1	5.44**	1.3	1	5.46***	1.3	1	6
<i>yqiL</i>	–	–	–	3.26**	4.1	2	–	–	–	6
<i>Vibrio vulnificus</i>										
<i>glp</i>	–	–	–	2.47*	7	1	–	–	–	3

Significant differences in the LRTs between models M3 or M2 and model M1 and model M8 and model M7 are indicated with asterisks after the  $\omega$  values.  $\omega$  and  $P_M$  estimates under purifying selection ( $\omega < 1$ ) or neutral selection ( $\omega = 0$ ) are not reported.

<sup>1</sup> Meats et al. (2003); <sup>2</sup> Achtman et al. (1999); <sup>3</sup> Maiden et al. (1998); <sup>4</sup> Enright et al. (2000); <sup>5</sup> Hanage et al. (2004). <sup>cca</sup> encapsulated; <sup>nca</sup> noncapsulated.

\*  $P < 0.05$ .

\*\*  $P < 0.01$ .

\*\*\*  $P < 0.001$ .

genes analyzed here, the observed values of  $\omega$  and  $n_M$  are so high that it is hard to believe that LRTs are completely misleading in their conclusions, especially for M7–M8 comparisons (e.g., *N. gonorrhoeae* and *S. agalactiae*). Moreover, it has been shown that LRTs are conservative (Anisimova et al., 2001, 2003; Yang et al., 2000a), so genes inferred by the test to undergo positive selection are most likely true cases of adaptation rather than an artifact of the method, as proven in most of the published studies (e.g., Bishop et al., 2000; Peek et al., 2001; Yang et al., 2000b; Yang and Swanson, 2002). Besides, we have adopted an even more conservative approach since we are not considering the loci under significant positive selection for which positively selected sites were not identified. Furthermore, gene regions and sites undergoing adaptive selection under the models implemented in PAML were also verified by TreeSAAP using a completely different amino acid-based approach which potentially, is less affected by recombination. Therefore, in conclusion, we think that all of the previous evidence indicates that microbial MLST housekeeping genes are experiencing molecular adaptation. We find this quite surprising, since these genes were essentially selected as candidates for population genetic studies because of their lack of selection as inferred by the

average  $\omega$  ratio. Previous studies reporting lack of diversifying selection in these genes must be interpreted cautiously. Moreover, one should be aware of their lack of neutrality when used for population or molecular evolutionary studies. Nevertheless, we do not think that our findings invalidate the use of these molecular markers for typing purposes; we agree with Cooper and Feil (2004) that “the exclusion of genes that do not conform to classical housekeeping criteria is an ill-afforded luxury”.

The finding of selection in housekeeping loci raises important evolutionary questions such as: how do these adaptive changes affect the phenotypes (proteins)? Using TreeSAAP and PAML we have first identified the sites responsible for adaptive change, providing the initial information required to understand the changes in the form and function of proteins over evolutionary time (Anisimova et al., 2002). Specific hypotheses can then be formulated using this information, for example, to propose coevolutionary patterns between host and parasite (e.g., Bishop et al., 2000), study how pathogens escape the immune system (e.g., Haydon et al., 2001), or determine which structural and biochemical amino acid properties drive the evolution of proteins (e.g., McClellan et al., 2005). As an example of the latter, we have used TreeSAAP to detect

Table 3

Amino acid (AA) sites and physicochemical properties under destabilizing selection ( $z$ -score  $> 2.326$ ,  $P < 0.01$ ; TreeSAAP) for encapsulated and noncapsulated isolates of *Haemophilus influenzae* in *adk* and *atpG*

Locus	Encapsulated		Noncapsulated	
	AA sites	AA properties	AA sites	AA properties
<i>adk</i>	42,155 42	Partial specific volume, short and medium range nonbonded energy	45,95,133 133	Compressibility, polar requirement
<i>atpG</i>	147 30,147	Power to be at the N-terminal, refractive index	50 30	Composition, refractive index

amino acid properties under strong levels of destabilizing selection ( $z$ -scores  $> 2.326$ ;  $P < 0.01$ ) in *adk* and *atpG* for encapsulated and nonencapsulated isolates of *H. influenzae* (Table 3). Using this approach we were able to identify a total of four and three different potential properties driving protein evolution of *adk* and *atpG*, respectively. Then, following McClellan et al. (2005), future studies using protein structure models could explore how these property changes may affect the conformation and function of *adk* and *atpG* and look into their interconnections with the epidemiology and pathogenesis of both typeable and nontypeable *H. influenzae*.

### 3.2. Locus comparisons

Tables 1 and 2 show how population recombination, population mutation, and adaptive selection rates per locus vary within and between species. As we have shown this information can be used to identify appropriate candidate loci for phylogenetics and population genetics, study protein evolution, target potentially useful MLST gene regions in other species, examine the evolution of antibiotic resistance, and explore the population dynamics of species. Another interesting angle to look at these two tables is comparing how these three parameters change among species for the same locus -is there any observable pattern of gene evolution? Our data sets consist of 91 loci of which 65 were screened for only a particular species and 27 were screened for two to five species, hence, the number of data sets per locus to compare is not very large. Nevertheless, it seems like  $\rho$  and  $\Theta_{wf}$  vary arbitrarily between taxa, so no obvious gene-based pattern could be established. This is not completely surprising considering that these population parameters are driven by the particular biological and ecological characteristics of each species, although as mentioned before, natural selection and population structuring can also act upon particular genes. Adaptive selection, while influenced by biological and ecological factors, is mostly a reflection of the selection pressure operating at the protein level. Convergent evolutionary responses to similar diversifying selective regimes could result in concordant patterns of adaptive selection between species for a particular locus. Our scarce data indicate that most loci seem to be under nonconcordant patterns of adaptive selection pressure. However, *aroE* and *xpt* in two species and *adk* in four species showed significant  $\omega > 1$  under M8 or M3 (under nonsignificant values of  $\rho$ ) and  $n_M$  and  $n_{TS} \neq 0$ , which may suggest a common pattern of positive selection for each of these genes. Further analyses including more species and loci are needed to confirm this hypothesis.

## 4. Summary

Model-based statistical methods are of great utility for inferring and testing a wide variety of evolutionary

parameters and hypotheses. Here we have provided a robust example of their utility for inferring population recombination, population mutation, and selection rates and building consistent phylogenetic hypotheses of relationships using a large database of multilocus sequence typing sequence data from infectious microbial agents. Within this framework, important evolutionary questions within microbial genetics have been assessed and new ones have been proposed. We hope that the outcomes of our work will stimulate further research in the evolution of infectious diseases using statistical methodology.

## Acknowledgements

This publication made use of the following MLST websites: *Bacillus cereus* (<http://pubmlst.org/bcereus>), *Burkholderia pseudomallei* (<http://bpseudomallei.mlst.net>), *Candida albicans* (<http://calbicans.mlst.net>), *Campylobacter jejuni* (<http://pubmlst.org/campylobacter>), *Enterococcus faecium* (<http://efaecium.mlst.net>), *Haemophilus influenzae* (<http://haemophilus.mlst.net>), *Helicobacter pylori* (<http://pubmlst.org/helicobacter>), *Neisseria* (<http://pubmlst.org/neisseria>), *Streptococcus agalactiae* (<http://sagalactiae.mlst.net>), *Staphylococcus aureus* (<http://saureus.mlst.net>), *Staphylococcus epidermidis* (<http://sepidermidis.mlst.net>), *Streptococcus pneumoniae* (<http://spneumoniae.mlst.net>), *Streptococcus pyogenes* (<http://spyogenes.mlst.net>), and *Vibrio vulnificus* (<http://pubmlst.org/vvulnificus>).

We thank Mark Achtman and three anonymous referees for their suggestions to improve this manuscript. We gratefully acknowledge support from the National Institutes of Health grants R01 AI50217 (RPV, KAC) and GM66276 (KAC) and from the Brigham Young University Office of Research and Creative Activities.

## References

- Achtman, M., Azuma, T., Berg, D.E., Ito, Y., Morelli, G., Pan, Z.J., Suerbaum, S., Thompson, S.A., van der Ende, A., van Doorn, L.J., 1999. Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.* 32, 459–470.
- Anisimova, M., Bielawski, J.P., Yang, Z., 2001. Accuracy and power of the likelihood ratio test to detect adaptive molecular evolution. *Mol. Biol. Evol.* 18, 1585–1592.
- Anisimova, M., Bielawski, J.P., Yang, Z., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* 19, 950–958.
- Anisimova, M., Nielsen, R., Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164, 1229–1236.
- Bishop, J.G., Dean, A.M., Mitchell-Olds, T., 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant–pathogen coevolution. *Proc. Natl. Acad. Sci. U.S.A.* 97, 5322–5327.
- Brown, C.J., Garner, E.C., Dunker, A.K., Joyce, P., 2001. The power to detect recombination using the coalescent. *Mol. Biol. Evol.* 18, 1421–1424.

- Cooper, J.E., Feil, E.J., 2004. Multilocus sequence typing—what is resolved? *Trends Microbiol.* 12, 373–377.
- Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C., Salzman, N.P., 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* 16, 372–382.
- Dingle, K.E., Colles, F.M., Wareing, D.R.A., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J.L., Urwin, R., Maiden, M.C.J., 2001. Multilocus sequence typing for *Campylobacter jejuni*. *J. Clin. Microbiol.* 39, 14–23.
- Endo, T., Ikeo, K., Gojobori, T., 1996. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13, 658–690.
- Enright, M.C., Day, N.P., Davies, C.E., Peacock, S.J., Spratt, B.G., 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol.* 38, 1008–1115.
- Enright, M.C., Spratt, B.G., Kalia, A., Cross, J.H., Bessen, D.E., 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. *Infect. Immun.* 69, 2416–2427.
- Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Maynard-Smith, J., Murphy, M., Spratt, B.G., Moore, C.E., Day, N.P.J., 2003. How clonal is *Staphylococcus aureus*? *J. Bacteriol.* 185, 3307–3316.
- Feil, E.J., Enright, M.C., Spratt, B.G., 2000a. Estimating the relative contribution of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. *Res. Microbiol.* 151, 465–469.
- Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., Zhou, J., Spratt, B.G., 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U.S.A.* 98, 182–187.
- Feil, E.J., Maiden, M.C.J., Achtman, M., Spratt, B.G., 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol. Biol. Evol.* 16, 1496–1502.
- Feil, E.J., Maynard-Smith, J., Enright, M.C., Spratt, B.G., 2000b. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* 154, 1439–1450.
- Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.
- Felsenstein, J., 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.
- Fu, Y.-X., Li, W.-H., 1993. Maximum likelihood estimation of population parameters. *Genetics* 134, 1261–1270.
- Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736.
- Hanage, W.P., Auranen, K., Syrjanen, R., Herva, E., Makela, P.H., Kilpi, T., Spratt, B.G., 2004. Ability of pneumococcal serotypes and clones to cause acute otitis media: implications for the prevention of otitis media by conjugate vaccines. *Infect. Immun.* 72, 76–81.
- Hasegawa, M., Kishino, K., Yano, T., 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174.
- Haydon, D.T., Bastos, A.D., Knowles, N.J., Samuel, A.R., 2001. Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics* 157, 7–15.
- Homan, W.L., Tribe, D., Poznanski, S., Li, M., Hogg, G., Spalburg, E., Van Embden, J.D., Willems, R.J., 2002. Multilocus sequence typing scheme for *Enterococcus faecium*. *J. Clin. Microbiol.* 40, 1963–1970.
- Hudson, R.R., 1990. Gene genealogies and the coalescent process. In: Futuyma, D., Antonovics, J. (Eds.), *Oxford Surveys in Evolutionary Biology*, vol. 7. Oxford University Press, Oxford, pp. 23–36.
- Hudson, R.R., 2001. Two-locus sampling distributions and their application. *Genetics* 159, 1805–1817.
- Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28, 437–466.
- Jolley, K.A., Kalmusova, J., Feil, E.J., Gupta, S., Musilek, M., Kriz, P., Maiden, M.C., 2000. Carried meningococci in the Czech Republic: a diverse recombining population. *J. Clin. Microbiol.* 38, 4492–4498.
- Jones, N., Bohnsack, J.F., Takahashi, S., Oliver, K.A., Chan, M.S., Kunst, F., Glaser, P., Rusniok, C., Crook, D.W., Harding, R.M., Bisharat, N., Spratt, B.G., 2003. Multilocus sequence typing system for group B streptococcus. *J. Clin. Microbiol.* 41, 2530–2536.
- Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.M. (Ed.), *Mammalian Protein Metabolism*. Academic Press, New York, NY, pp. 21–132.
- Kelsey, C.R., Crandall, K.A., Voevodin, A.F., 1999. Different models, different trees: the geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* 13, 336–347.
- Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.
- Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. U.S.A.* 78, 454–458.
- Kuhner, M.K., Yamato, J., Felsenstein, J., 1995. Estimating effective population size and mutation from sequence data using Metropolis-Hastings sampling. *Genetics* 140, 1421–1430.
- Kuhner, M.K., Yamato, J., Felsenstein, J., 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149, 429–434.
- Li, W.-H., 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Maddison, D.R., Maddison, W.P., 2000. *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA.
- Maggi-Solcà, N., Bernasconi, M.V., Valsangiacomo, C., Van Doorn, L.J., Piffaretti, J.C., 2001. Population genetics of *Helicobacter pylori* in the southern part of Switzerland analysed by sequencing of four house-keeping genes (*atpD*, *glnA*, *scoB* and *recA*), and by *vacA*, *cagA*, *iceA* and *IS605* genotyping. *Microbiology* 147, 1693–1707.
- Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus-sequencing typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145.
- Maynard-Smith, J., 1995. Do bacteria have population genetics? In: Baumberg, J.P., Young, W., Saunders, J.R., Wellington, E.M.H. (Eds.), *Population Genetics of Bacteria*. Society for General Microbiology, Symposium 52. Cambridge University Press, London, pp. 1–12.
- Maynard-Smith, J., Feil, E.J., Smith, N.H., 2000. Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays* 22, 1115–1122.
- Maynard-Smith, J., Smith, N.H., 1998. Detecting recombination from gene trees. *Mol. Biol. Evol.* 15, 590–599.
- Maynard-Smith, J., Smith, N.H., O'Rourke, M., Spratt, B.G., 1993. How clonal are bacteria? *Proc. Natl. Acad. Sci. U.S.A.* 90, 4384–4388.
- McClellan, D.A., McCracken, K.G., 2001. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domain. *Mol. Biol. Evol.* 18, 917–925.
- McClellan, D.A., Palfreyman, E.J., Smith, M.J., Moss, J.L., Christensen, R.G., Sailsbery, J.K., 2005. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome *b* proteins. *Mol. Biol. Evol.* 22, 437–455.
- McVean, G., Awadalla, P., Fearnhead, P., 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160, 1231–1241.
- Meats, E., Feil, E.J., Stringer, S., Cody, A.J., Goldstein, R., Kroll, J.C., Popovic, T., Spratt, B.G., 2003. Characterization of encapsulated and nonencapsulated *Haemophilus influenzae* and determination of phyloge-

- netic relationships by multilocus sequence typing. *J. Clin. Microbiol.* 41, 1623–1636.
- Miyata, T., Yasunaga, 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and nonsynonymous amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.* 16, 23–36.
- Nei, M., Gojobori, T., 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426.
- Nielsen, R., Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936.
- Nordborg, M., 2001. Coalescent theory. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), *Handbook of Statistical Genetics*. John Wiley and Sons Ltd., Chichester, pp. 179–212.
- Peck, A.S., Souza, V., Eguiarte, L.E., Gaut, B.S., 2001. The interaction of protein structure, selection, and recombination on the evolution of the type 1 fimbrial major submit (fimA) from *Escherichia coli*. *J. Mol. Evol.* 52, 193–204.
- Posada, D., 2001. The effect of branch length variation on the selection of models of molecular evolution. *J. Mol. Evol.* 52, 434–444.
- Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Posada, D., Crandall, K.A., 2001a. A comparison of different strategies for selecting models of DNA substitution. *Syst. Biol.* 50, 580–601.
- Posada, D., Crandall, K.A., 2001b. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13757–13762.
- Posada, D., Crandall, K.A., 2001c. Intraspecific gene genealogies: trees grafting into networks. *TREE* 16, 37–45.
- Posada, D., Crandall, K.A., Holmes, E.C., 2002. Recombination in evolutionary genomics. *Annu. Rev. Genet.* 36, 15–91.
- Spratt, B.G., Maiden, M.C.J., 1999. Bacterial population genetics, evolution and epidemiology. *Philos. Trans. R. Soc. Lond. B* 354, 701–710.
- Suerbaum, S., Lohrengel, M., Sonnevend, A., Ruberg, F., Kist, M., 2001. Allelic diversity and recombination in *Campylobacter jejuni*. *J. Bacteriol.* 183, 2553–2559.
- Suerbaum, S., Smith, J.M., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dyrek, I., Achtman, M., 1998. Free recombination within *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* 95, 12619–12624.
- Swofford, D.L., 2003. *Phylogenetic Analysis Using Parsimony (PAUP and other methods)*. Sinauer Associates, Sunderland, MA.
- Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526.
- Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura, R.M. (Ed.), *Some Mathematical Questions in Biology—DNA Sequence Analysis*. American Mathematical Society, Providence, RI, pp. 57–86.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The clustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24, 4876–4882.
- Urwin, R., Holmes, E.C., Fox, A.J., Derrick, J.P., Maiden, M.C.J., 2002. Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen porB. *Mol. Biol. Evol.* 19, 1686–1694.
- Urwin, R., Maiden, M.C.J., 2003. Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.* 11, 479–487.
- Vilas-Boas, G., Sanchis, V., Lereclus, D., Lemos, M.V., Bourguet, D., 2002. Genetic differentiation between sympatric populations of *Bacillus cereus* and *Bacillus thuringiensis*. *Appl. Environ. Microbiol.* 68, 1414–1424.
- Viscidi, R.P., Demma, J.C., 2003. Genetic diversity of *Neisseria gonorrhoeae* housekeeping genes. *J. Clin. Microbiol.* 41, 197–204.
- Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276.
- Whittam, T.S., 1995. Genetic population structure and pathogenicity in enteric bacteria. In: Baumberg, S., Young, J.P.W., Wellington, E.M.H., Saunders, J.R. (Eds.), *Population Genetics of Bacteria*. Cambridge University Press, pp. 217–245.
- Woolley, S., Johnson, J., Smith, M.J., Crandall, K.A., McClellan, D.A., 2003. TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics* 19, 671–672.
- Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556. <http://abacus.gene.ucl.ac.uk/software/paml.html>.
- Yang, Z., Goldman, N., Friday, A., 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11, 316–324.
- Yang, Z., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917.
- Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M.K., 2000a. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449.
- Yang, Z., Swanson, W.J., 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among sites classes. *Mol. Biol. Evol.* 19, 49–57.
- Yang, Z., Swanson, W.J., Vacquier, V.D., 2000b. Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* 17, 1446–1455.
- Zharkikh, A., 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* 39, 315–329.