Infection, Genetics and Evolution 8 (2008) 593–602

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Rate heterogeneity in the evolution of *Helicobacter pylori* and the behavior of homoplastic sites

Richard J. Meinersmann [a,*], Judith Romero-Gallo [b], Martin J. Blaser [c]

[a] *USDA Agricultural Research Service, Athens, GA, United States*
[b] *Vanderbilt University, Nashville, TN, United States*
[c] *New York University School of Medicine and VA Medical Center, New York, NY, United States*

ARTICLE INFO

Article history:
Received 7 February 2008
Received in revised form 9 April 2008
Accepted 11 April 2008
Available online 18 April 2008

Keywords:
Homoplasy
Recombination
Transitions
Transversions
Helicobacter pylori

ABSTRACT

Helicobacter pylori are bacteria with substantial inter-strain variability and phylogenetic reconstructions of sequence data from the organism have common homoplastic sites. Although frequent recombination events have been proposed to contribute to the variation, the effects of nucleotide substitution rate heterogeneities on the reconstruction of *H. pylori* genealogies have not been studied. We analyzed the substitution pattern of a housekeeping gene, a homologue of the ribonuclease reductase gene (*rnr*), to characterize rate heterogeneities between 11 *H. pylori* isolates. Evidence of limited recombination was demonstrated by the Sawyer's runs test, but the homoplasy test and site-by-site compatibility tests indicated frequent recombination events. Within the 1935 nucleotide gene, 292 sites were polymorphic with an average pair-wise difference of 5.01%. Xia's distances for amino acids at non-synonymous codon substitution sites were smaller at homoplastic sites than at sites that were not homoplastic. Transitions were significantly more common among homoplastic than among non-homoplastic nucleotide substitutions. Simulations of evolution with or without recombination indicated the transition/transversion ratio is expected to be higher in homoplastic sites with no recombination. Despite evidence of recombination, analyses of the *rnr* genealogy does not show a random tree but rather base substitution behaviors characteristic of both recombination and substitution saturation at some sites. Analyses of sequences in the *H. pylori* multilocus sequence-typing database provided similar evidence for substitution saturation in multiple housekeeping genes.

Published by Elsevier B.V.

1. Introduction

Homoplasies are defined as when loci (sites in a sequence) have a number of steps on a reconstructed phylogenetic tree equal to or greater than the number of different alleles (characters) found at that locus (Maddison and Maddison, 1992). A homoplasy is an allele identified more than once in a phylogenic tree that is not derived from a common ancestor. Thus, a given character at a specific site will appear to arise more than once in the tree. Assuming an accurate reconstruction of the phylogeny, homoplasies can arise as a result of evolutionary convergence, parallelism or reversals. Analysis of multiple DNA sequences is the most informative method of population genetics, but accuracy is dependent on use of correct models of evolution (Hillis et al., 1994; Li, 1997). Rate heterogeneity, which is variation in substitution rate among the different nucleotides or base positions, when extreme, can lead to site saturation, i.e., multiple events at sites, some of which will show apparent reversion to an ancestral state (Xia, 2000b). However, horizontal gene transfer can also introduce homoplasies because genetic materials with different phylogenetic histories are brought together such that apparent reversals are introduced. Some researchers believe that horizontal gene transfer is more likely than reversals due to point mutations to introduce homoplasies in bacteria (Falush et al., 2001), but at this time there are no diagnostic methods to distinguish between these two mechanisms.

Helicobacter pylori plays an important role in gastroduodenal diseases in humans and has remarkable adaptations that allow persistent survival in the gastric niche (Doig et al., 1999; Montecucco and Rappuoli, 2001). *H. pylori* also are extremely variable, with a low probability that independent isolates will be the same (Go et al., 1996; Han et al., 2000) and appear panmictic

* Corresponding author at: USDA Agricultural Research Service, Russell Research Center, P.O. Box 5677, Athens, GA 30604, United States. Tel.: +1 706 546 3236; fax: +1 706 546 3633.
E-mail address: Rick.Meinersmann@ars.usda.gov (R.J. Meinersmann).

1567-1348/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.meegid.2008.04.003

(Salaun et al., 1998; Suerbaum et al., 1998), meaning that genetic exchange occurs too rapidly for linkage of alleles at different loci to be observed. Studies of the population genetics of *H. pylori* may be helpful in identifying selective pressures on the organism. Reconstructing ancestral relationships between strains is confounded by recombination events, as well as by rate heterogeneity. Although most of the limitations in *H. pylori* phylogenetics have been attributed to recombinant events (Maynard-Smith and Smith, 1998; Suerbaum et al., 1998), rate heterogeneity has not been fully considered.

This study began as a survey of *H. pylori* diversity within individual patients (Romero-Gallo and Blaser, unpublished). In one patient separate isolates displayed a gene with heterogeneity and sequence analysis showed high similarity to the gene encoding ribonucleotide reductase (*rnr*) (Tobe et al., 1992; Cheng et al., 1998), a housekeeping gene expected to be well conserved. Since sampling of partial *rnr* sequences from several independent isolates of *H. pylori* showed substantial diversity, the present study sought to determine the mechanisms underlying the observed patterns of *rnr* diversity, and whether units of recombination could be identified. Our findings indicated that there was a bias in the type of changes seen among homoplastic sites that was more consistent with rate heterogeneity than by horizontal gene transfer. We further tested to see if the bias was seen in controlled computer simulations of evolution and in other examples of *Helicobacter* gene evolution.

2. Materials and methods

2.1. Gene sequences

The *H. pylori* isolates studied are listed in Table 1. Isolates 97-793 and 97-645 were isolated from the fundus and antrum, respectively, from biopsies obtained simultaneously from the same patient, as described above. This pair of isolates appeared to be clonal variants, based on RAPD analysis (data not shown). DNA sequences for strains 26695 and J99 were obtained from the genomic sequences deposited in Genbank (accession numbers AE000630 and AE001544). For the remaining isolates, templates for DNA sequencing were generated by polymerase chain reaction using primers listed in Table 2. Sequencing was performed by dye terminator reactions with the same and other primers (Table 2) followed by analysis on an ABI Prism 373 automated sequencer. The quality of the sequences was evaluated and contiguous sequences were constructed with Sequencher (Gene Codes Corporation, Ann Arbor, MI).

Table 1
H. pylori isolates used in study

Strain designations		Source
97-679	A-4	India (Ladakh)
97-645	A-30[a]	India (Ladakh)
97-793	F-30[a]	India (Ladakh)
98-884	ATCC51407	Monkey
98-927	1308-3	Monkey
98-924	1309-3	Monkey
97-12	DB 011	Hong Kong
26695		UK (genomic sequence strain)
Hpk5		Japan
HPJ166		USA
J99		USA (genomic sequence strain)

[a] A-30 and F-30 were the index isolates from the stomach antrum and fundus, respectively, of the same person and appear to be clonal variants by RAPD (data not shown); both were included to bias toward finding recombination junctions.

2.2. Analyses

DNA sequences were aligned using ClustalX (Thompson et al., 1994). No gaps were introduced into the alignment and no additional editing was necessary. Phylogenetic reconstructions were performed using PAUP* ver. 4.0b4a (Swofford, 1998) and with DAMBE ver. 4.0.39 (Xia, 2000a). Window analyses of nucleotide diversity were produced using MULTICOMP ver. 1.01 (kindly provided by Ruiting Lan, University of Sydney, Sydney, New South Wales, Australia) using values for synonymous (K_s) and non-synonymous (K_a) substitution rates calculated by the method of Li (1993). Substitution patterns based on phylogenies were analyzed using DAMBE (Xia, 2000a). Homoplasies were identified on the same phylogenies using PAUP*. Homoplasies were observed at identical sites when the rooting was changed to other isolates or to midpoint rooting. Recombination analyses using site-by-site compatibility methods were done with RETICULATE (Jakobsen and Easteal, 1996) and with SITES (Hey and Wakeley, 1997). Maynard–Smith's homoplasy test was performed with HOMOPLASY (Maynard-Smith and Smith, 1998), and Sawyer's runs test was performed using GENECONV (Sawyer, 1999). The informative sites test was performed with PIST (Worobey, 2001) on datasets containing only the third codon position base. RECOMBINE (Kuhner et al., 2000) was run with trees generated in PAUP* as the initial phylogeny, the transition/transversion ratio set to 4.0, the base frequencies and initial value of theta determined by the program, and the remainder of the parameters set at the default values. The shape parameter (α) of the gamma distribution was determined with the data generated by PAUP* using the HKY85 model and four rate categories in a maximum likelihood test (Swofford, 1998). MODELTEST ver. 3.7 (Posada and Crandall, 1998) was used to compare different models of evolution.

2.3. Computer simulations

Computer simulations were run with "Seq-Gen" (Rambaut and Grassly, 1997) in which either recombination or rate heterogeneity without recombination was responsible for the bulk of the homoplasies in the data. The HKY85 model of evolution (Hasegawa et al., 1985) was used with the base composition and the transition/transversion ratio set to approximate those calculated from the *rnr* data. Seq-Gen uses guide trees to generate simulated evolutions. For non-recombinant data, a single guide tree was used (the tree reconstructed from the *rnr* sequence) and the rate heterogeneity shape parameter (α) was set to the value calculated for the *rnr* data. To simulate recombination the rate heterogeneity was left at the default (equal rates for all sites) and the sequence was divided into nine partitions. The nine partitions were defined by bootscan analysis (Salminen et al., 1995) implemented in RDP (Martin and Rybicki, 2000) of the *rnr* data and trees that were used as guide trees by Seq-Gen were deduced for each partition with PAUP*. One thousand simulations were run with either high rate heterogeneity (recombination essentially nil) or with high recombination (rate heterogeneity essentially nil). The output for each simulation was put through PAUP* to diagnose changes at each site to determine if they were homoplastic and if they were transitions or transversions. Scripts were developed in SAS and Excel to tabulate the number of changes in each simulation for each category: homoplastic or non-homoplastic, and transition or transversion. The informative sites among the non-homoplastic sites were also considered separately since all homoplastic sites are automatically informative. Excel was used to tabulate the frequency of each observation and to produce the graphics.