Genomics 102 (2013) 331–337

Contents lists available at ScienceDirect

Genomics

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

Genetic boundaries to delineate the typhoid agent and other *Salmonella* serotypes into distinct natural lineages

Le Tang [a], Chun-Xiao Wang [a], Song-Ling Zhu [a], Yang Li [a], Xia Deng [a], Randal N. Johnston [b], Gui-Rong Liu [a,*], Shu-Lin Liu [a,c,d,e,*]

[a] Genomics Research Center (one of The State-Province Key Laboratories of Biomedicine-Pharmaceutics of China), Harbin Medical University, Harbin, China
[b] Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, Canada
[c] HMU-UCFM Center for Infection and Genomics, Harbin Medical University, Harbin, China
[d] Department of Biopharmaceutics, Harbin Medical University, Harbin, China
[e] Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, Canada

ARTICLE INFO

Article history:
Received 8 June 2013
Accepted 24 July 2013
Available online 6 August 2013

Keywords:
Bacterial species
Genetic boundary
Genomic divergence

ABSTRACT

The deadly human typhoid agent was initially classified as a species called *Salmonella typhi* but later reclassified as a serovar of *Salmonella enterica* together with other pathogenically diverse serovars. The dynamic changes of *Salmonella* taxonomy reflect the need to clarify the phylogenetic status of the *Salmonella* serovars: are they discrete lineages or variants of a genetic lineage? To answer this question, we compared *S. typhi* and other *Salmonella* serotypes. We found that the *S. typhi* and *Salmonella typhimurium* strains had over 90% and ca. 80%, respectively, of their genes identical; however, between *S. typhi* and *S. typhimurium*, this percentage dropped to 6%, suggesting the existence of genetic boundaries between them. We conclude that *S. typhi* and the other compared *Salmonella* serovars have developed into distinct lineages circumscribed by the genetic boundary. This concept and methods may be used to delineate other *Salmonella* serotypes, many of which are polyphyletic, needing differentiation.

1. Introduction

Typhoid is a serious infectious disease with high morbidity and mortality [39]. The causative agent was first isolated in 1881 and was classified as a bacterial species [4,11,12], initially with multiple names and subsequently with the Latinized scientific name *Salmonella typhi*. In the 1980s, *S. typhi* was reclassified as a serological variant (called serovar) of a new species, *Salmonella enterica* [17–19,40], together with other "serological variants" that, though genetically similar, are pathogenically diverse. While the new name of the typhoid agent, *S. enterica* subspecies *enterica* serovar Typhi, has been widely adopted, many authors continue using the previous nomenclature [44], due largely to several uncertainties of the new *Salmonella* taxonomy such as cutoffs for delineating bacteria into phylogenetic groupings. Regarding the typhoid agent, we ask a fundamental question: does this pathogen exist as a natural cluster of bacteria phylogenetically separated from other (and closely related) bacteria? Answering this question requires the identification of some kind of genetic boundary that may circumscribe the typhoid agent unambiguously from its close relatives.

Bacteria, like animals and plants, are categorized into species, based initially on phenotypic similarities and later also, and mainly, on genetic relatedness. However, unlike most animals and plants, which are classified into species based largely on objective criteria such as sexual reproduction potential [33,34], bacteria are essentially asexual organisms, which makes the adopted classification system problematic. The use of DNA–DNA association and 16S rDNA sequence comparison to define bacterial species has revolutionized bacterial classification [8,46,50]. However, evidence is needed to demonstrate that bacteria do exist in discrete natural clusters with unambiguous genetic boundaries separating each of them from others. As it is now increasingly recognized that bacterial species delineated by the modern taxonomy system are essentially complexes with great intra-"species" diversity [47], the current bacterial species definition may categorize multiple natural bacterial clusters into the same "species", hence inevitably causing confusions in many theoretical as well as applied areas. In addition, new advances of bacterial systematics tend to change the taxonomy and nomenclature of bacteria radically, moving some taxa back and forth from one species to another. As a result, a so-named species may contain both mild or non-pathogens and deadly pathogens, as exemplified by the *Salmonella* taxonomy [14,19]. Currently, over 2000 serological types of bacteria are documented under *S. enterica*, most of which are mild or virtually non-pathogenic to humans. Inclusion of the deadly human pathogen *S. typhi* together with them in the same species needs to be re-justified.

All these confusions have stepped from the lack of an answer to a long-asked question: do bacteria exist as discrete phylogenetic clusters? Many lines of evidence, including those from surveys of wild microbial populations, very convincingly indicate the existence of

* Corresponding authors at: Genomics Research Center, Harbin Medical University, 157 Baojian Road, Harbin 150081, China. Fax: +86 451866614075.
E-mail addresses: grliu.natsumi@gmail.com (G.-R. Liu), slliu@ucalgary.ca (S.-L. Liu).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.07.014

genetically discernible bacterial clusters that possess the attributes applicable to species (for a couple of comprehensive reviews, see [1,3]). As such, *S. enterica* may consist of discrete phylogenetic clusters with unambiguous genetic distinctions among them — the serologically differentiated types (previously called serotypes but treated as separate species and later called serovars) actually may either be monophyletic (such as 9,12:d:– for *S. typhi* and 1,2,12:a:[1,5] for *Salmonella paratyphi* A [2,13,27,35,42]) or polyphyletic (such as 6,7:c:1,5 for *S. paratyphi* C, *Salmonella choleraesuis* and *Salmonella typhisuis* or 1,9,12:a:1,5 for *Salmonella miami* and *Salmonella sendai* [17]). In this study, we compared genomic sequences of *S. typhi* and other representative monophyletic *Salmonella* serotypes, primarily *Salmonella typhimurium*, to identify possible genomic features that may unambiguously discriminate them into discrete clusters without any genetic overlaps among them. We found that strains of the same *Salmonella* monophyletic serotype shared very high sequence identity as expected; however the sequence identity dropped abruptly between different *Salmonella* serotypes as closely related as between *S. typhi* and *S. typhimurium*, forming a clear-cut genetic boundary between them. We propose that the sharp genetic sequence divergence between highly related *Salmonella* serotypes can be used to distinguish *Salmonella* into phylogenetically definite lineages, each being biologically unique such as in host range or in the nature of diseases to elicit in the host. Polyphyletic *Salmonella* serotypes, many of which have been successfully differentiated by biochemical assays such as 1,4,[5],12:b:1,2 into *S. paratyphi* B (adapted to humans causing paratyphoid) and *Salmonella java* (infecting a broad range of hosts) by the *d*-tartrate test, may be now resolved into distinct lineages based on their genomic differences reported in this article.

2. Results

2.1. Genomic comparison to identify characteristics for distinguishing the typhoid agent from other Salmonella lineages

We hypothesize that, as a unique human pathogen dwelling in a specific niche [15,39], the typhoid agent might have diverged enough from other *Salmonella* lineages to become a nascent cluster of bacteria circumscribed by clear-cut genetic boundaries between itself and its close relatives. To test this hypothesis, we needed first to find possible genomic features that may unambiguously recognize the typhoid agent as a cohesive genetic cluster of bacteria. For this, we compared strains of *S. typhi* and strains between *S. typhi* and some very closely related *Salmonella* lineages, focusing particularly on *S. typhi* and *S. typhimurium*, which are the two most representative *Salmonella* lineages [25,28,36,38].

We began with comparisons of the complete genomes of three *S. typhi* and six *S. typhimurium* strains. As these genomes were analyzed by different groups of investigators, the different criteria and methods used for defining genes would inevitably lead to different annotation results. Whereas no annotation tools or methods could be said to be better than others, the same parameters or criteria were necessary for gene-by-gene comparisons between pairs of the genomes in our study. We thus re-annotated these genomes using tools provided in http://rast.nmpdr.org/ with manual checking where necessary. As shown in Table 1, we obtained different numbers of genes after the re-annotation. For example, *S. typhimurium* LT2 and 14028S had 4598 and 5419 genes, respectively, in the literature; so 14028S seemed to have 821 more genes than LT2. After the re-annotation, LT2 and 14028S had 5078 and 5081 genes, respectively, and, most importantly in this study, homologous genes were strictly matched. We do not think that the gene numbers of LT2 and 14028S were "truer" after the re-annotation, but re-annotation by the same set of parameters made the gene-by-gene comparison in this study meaningful.

We found that, within *S. typhi* or *S. typhimurium*, independent strains had well over 90% of their gene contents in common (93–95% within *S. typhi* and 92–94% within *S. typhimurium*; Table 2), whereas between strains of *S. typhi* and *S. typhimurium*, the bacteria had less than 80% of the gene contents in common (76–79%; Table 2). The ~20% difference in gene contents is consistent with previous findings on a single strain each of *S. typhi* and *S. typhimurium*, CT18 and LT2, respectively [36,38]. However, it is hardly appropriate to take this kind of broad spectrum of genomic differences to set up objective criteria for distinguishing *S. typhi* and *S. typhimurium* or other closely related *Salmonella* lineages as separate bacterial clusters.

Table 1
Numbers of genes of the *Salmonella* strains after re-annotation in comparison with the previous numbers.

Strain	Number before re-annotation	Number after re-annotation
S. typhimurium		
LT2	4598	5078
14028S	5419	5081
SL1344	4623	5117
D23580	4628	5112
ST4/74	4775	5108
UK/1	4562	5014
S. typhi		
Ty2	4641	5168
CT18	4702	5223
P-stx-12	4884	5130

2.2. Abrupt genomic sequence divergence between S. typhi and S. typhimurium

We then focused our analysis on the nucleotide sequences common to *S. typhi* and *S. typhimurium* strains. Of great significance, we found abrupt genomic sequence divergence between *S. typhi* and *S. typhimurium*: whereas within *S. typhi* or *S. typhimurium*, most genes had 100% sequence identity among the wild type strains compared (over 90% of all shared genes in *S. typhi* and close to 80% of all shared genes in *S. typhimurium*; Supplementary Table 1), only 6% of the genes common to the compared *S. typhi* and *S. typhimurium* strains had 100% sequence identity (Table 3). These results demonstrate the existence of clear-cut boundaries between *S. typhi* and *S. typhimurium* that may limit their genetic exchanges and suggest the isolation of gene pools of the two *Salmonella* lineages for considerable evolutionary times.

To reveal the genomic divergence between *S. typhi* and *S. typhimurium* more intuitively, we carried out pair-wise and progressive whole genomic comparisons among six strains, including *S. typhi* Ty2, CT18 and P-stx-12, and *S. typhimurium* LT2, 14028S and SL1344. We first picked up Ty2 and then added CT18 to obtain the number of genes that have 100% sequence identity between the two strains; in

Table 2
Numbers of genes in common between pairs of *Salmonella* strains.

Strain	Ty2	CT18[a]	P-stx-12[a]	LT2[a]	14028S[a]	SL1344[a]
Ty2	5168	4875 94%; 93%	4873 94%; 95%	3997 77%;79%	3951 76%; 78%	3996 77%; 78%
CT18		5223	4847 93%; 94%	4032 77%; 79%	3957 76%; 78%	4028 77%; 79%
P-stx-12			5130	3991 78%;79%	3941 77%; 78%	4010 78%; 78%
LT2				5078	4696 92%; 92%	4711 93%; 92%
14028S					5081	4774 94%; 93%
SL1344						5117

[a] Below the figure for the number of genes in common for the two genomes in comparison, the first percentage indicates the ratio of shared genes in the strain in the left-most column and the second percentage indicates the ratio of shared genes in the strain in the first row of the table.