Molecular Phylogenetics and Evolution 103 (2016) 133–142

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

An evaluation of logic regression-based biomarker discovery across multiple intergenic regions for predicting host specificity in *Escherichia coli*

Shuai Zhi [a], Qiaozhi Li [a], Yutaka Yasui [b], Graham Banting [a], Thomas A. Edge [c], Edward Topp [d], Tim A. McAllister [e], Norman F. Neumann [a,f,*]

[a] School of Public Health, Room 3-57, South Academic Building, University of Alberta, Edmonton, Alberta T6G 2G7, Canada
[b] Epidemiology & Cancer Control Department, MS 735, Room S6043, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN, United States
[c] Environment Canada, 867 Lakeshore Rd, Burlington, Ontario L7S 1A1, Canada
[d] Agriculture and Agri-Food Canada, 1391 Sandford St, London, Ontario N5V 4T3, Canada
[e] Agriculture and Agri-Food Canada, PO Box 3000, 5403 1st Avenue South, Lethbridge, Alberta T1J 4B1, Canada
[f] Environmental Microbiology Program, Provincial Laboratory for Public Health, Walter McKenzie Sciences Center, 8440-112 Street, Edmonton, Alberta T6G 2J2, Canada

ARTICLE INFO

Article history:
Received 10 May 2016
Revised 23 June 2016
Accepted 14 July 2016
Available online 16 July 2016

Keywords:
Host specificity
Escherichia coli
Intergenic region
Logic regression
DNA sequencing
Bioinformatics

ABSTRACT

Several studies have demonstrated that *E. coli* appears to display some level of host adaptation and specificity. Recent studies in our laboratory support these findings as determined by logic regression modeling of single nucleotide polymorphisms (SNP) in intergenic regions (ITGRs). We sought to determine the degree of host-specific information encoded in various ITGRs across a library of animal *E. coli* isolates using both whole genome analysis and a targeted ITGR sequencing approach. Our findings demonstrated that ITGRs across the genome encode various degrees of host-specific information. Incorporating multiple ITGRs (i.e., concatenation) into logic regression model building resulted in greater host-specificity and sensitivity outcomes in biomarkers, but the overall level of polymorphism in an ITGR did not correlate with the degree of host-specificity encoded in the ITGR. This suggests that distinct SNPs in ITGRs may be more important in defining host-specificity than overall sequence variation, explaining why traditional unsupervised learning phylogenetic approaches may be less informative in terms of revealing host-specific information encoded in DNA sequence. *In silico* analysis of 80 candidate ITGRs from publically available *E. coli* genomes was performed as a tool for discovering highly host-specific ITGRs. In one ITGR (*ydeR-yedS*) we identified a SNP biomarker that was 98% specific for cattle and for which 92% of all *E. coli* isolates originating from cattle carried this unique biomarker. In the case of humans, a host-specific biomarker (98% specificity) was identified in the concatenated ITGR sequences of *rcsD-ompC*, *ydeR-yedS*, and *rclR-ykgE*, and for which 78% of *E. coli* originating from humans carried this biomarker. Interestingly, human-specific biomarkers were dominant in ITGRs regulating antibiotic resistance, whereas in cattle host-specific biomarkers were found in ITGRs involved in stress regulation. These data suggest that evolution towards host specificity may be driven by different natural selection pressures on the regulome of *E. coli* among different animal hosts.

1. Introduction

E. coli is a Gram-negative bacterium that naturally inhabits the intestine of various warm blooded animals. The bacterium has been long been considered a generalist able to colonize a broad range of animal host species and even the non-host environment (Power et al., 2005; Tymensen et al., 2015; Kon et al., 2007; Chandrasekaran et al., 2015). However, recent studies have demonstrated that *E. coli* appears to display some level of host/

Abbreviations: SNP, single nucleotide polymorphisms; ITGRs, intergenic regions; TMAO, trimethylamine N-oxide; TSB, tryptic soy broth; HPP, host predictive power; ML, maximum likelihood; UPGMA, unweighted pair group methods with arithmetic means; NJ, neighbor joining; FM, Fitch-Margoliash; MEA, minimum evolution algorithms; MP, maximum parsimony.

* Corresponding author at: 357-E South Academic Building, University of Alberta, Edmonton, Alberta T6G 2G7, Canada.
E-mail address: nfneuman@ualberta.ca (N.F. Neumann).

http://dx.doi.org/10.1016/j.ympev.2016.07.016
1055-7903/

environmental adaptation and specificity. For example, a human-specific *E. coli* clone was identified by Clermont et al. (2008) and found in human populations from several continents, but not in any other animals. *E. coli* O157:H7 can be grouped into two lineages, amongst which human strains are predominantly present in lineage I, whereas lineage II strains mostly consist of strains from cattle (Kim et al., 1999). By phylogenetic analysis, *E. coli* can be grouped into eight main phylogenetic groups: A, B1, B2, C, D, E, F and E clade I (Clermont et al., 2013) and various studies have demonstrated that the distribution of different *E. coli* phylogroups is host-related. For example, it was observed that *E. coli* strains from group A (40.5%) and B2 (25.5%) are more frequently isolated from humans while group B1 (41%) is most prevalent in animals (Tenaillon et al., 2010). Group B1 was found to be the most predominant phylogenetic group in bird isolates while group B2 are more abundantly distributed in omnivores (Gordon and Cowling, 2003). Furthermore, a study on enteropathogenic *E. coli* (EPEC) demonstrated that 79% of bovine isolates belonged to B1 group, while the majority of swine EPEC isolates (53.5%) were from group A (Wang et al., 2013). In *E. coli* from healthy humans, 54% strains were classified as B2 group (Wang et al., 2013).

More recently, Zhi et al. (2015) used logic regression analysis of DNA sequences in intergenic regions (ITGRs) from human, animal and environmental *E.coli* as a novel bioinformatics approach to identify host-specific single nucleotide polymorphic (SNP) biomarkers in *E. coli*. They found that in the feces of animals *E. coli* populations consist of a mixture of host-specialists and host-generalists, and highly specific SNP biomarkers among host-specialist populations were identified across 15 different animal hosts. For example, 53% of all human isolates displayed a unique ITGR biomarker pattern that was 98% specific to humans, suggesting that a significant proportion of the populations of *E. coli* in humans are specialists. Isolates of *E. coli* collected from deer displayed SNP biomarkers that were 99% specific with 82% of all isolates from deer carrying this biomarker. When this same approach was used for analyzing populations of *E. coli* in wastewater, a unique strain of *E. coli* was identified which was found in various wastewater treatment plants across Alberta, Canada, suggesting the evolution of niche-adapted genotypes in non-host environmental matrices (Zhi et al., 2016). Cumulatively, there is mounting evidence to suggest that *E. coli* may display a significant level of host and non-host adaptation/specificity.

In order to survive under the diverse microenvironmental conditions associated with various animal host gut physiologies or non-host environments (i.e. nutrient availability, pH, temperature, predation, UV, high oxygen), bacteria have evolved a number of adaptive coping strategies. The acquisition of specific genes [e.g., antibiotic resistance genes (Giedraitiene et al., 2011) or virulence genes (Mellata et al., 2010)] is one strategy whereby bacteria can achieve environmental fitness, but bacteria are also able to phenotypically adapt to adverse environments through the regulation of core genes (Prüss et al., 2006; Ziebuhr et al., 1999). Gene regulation can be altered through inactivation and/or activation of gene regulators (Baez and Shiloach, 2013) as well as through mutations in promoter sequences (Ando et al., 2011). For example, mutations in the promoter regions of the *katG* gene can alter expression, resulting in phenotypic changes that affect the susceptibility of *Mycobacterium tuberculosis* to isoniazid (Ando et al., 2011). By acquiring an aerobically expressed promoter for the expression of a previously silent citrate transporter, *E. coli* acquired the capacity to use citrate as an energy source under aerobic conditions (Blount et al., 2012). In *Bordetella pertussis* a new allele in the promoter of the pertussis toxin gene caused a dramatic increase in pertussis in humans in the Netherlands (Mooi et al., 2009).

For *E. coli*, the ability to colonize the gastrointestinal system of a specific animal host will largely be governed by the regulome – i.e., the ability to sense and respond to the unique stimuli in the gut and to compete with other enteric bacterial populations for limited nutrients, resources, and replication sites. We hypothesize that the intense intra-and inter-species microbial competition within the gut, coupled with the diversity in the cellular and molecular physiology of gastrointestinal microenvironments across animal species, will drive the evolution of the regulatory transcriptome of *E. coli* towards host-adaptation and specificity. Although our previous work in this area supports this hypothesis (Zhi et al., 2015), our original findings were limited to identifying host-specific SNP patterns in only three ITGRs across a library of 784 *E. coli* strains isolated from 15 different animals. Evolution towards host-specificity is not likely to follow a single trajectory arising from a single set of defined mutations, rather the evolutionary trajectory is likely to be multidirectional, driven by diverse combinations and permutations of various mutations across the transcriptome. As such, we sought to determine the degree of host-specific information encoded among various ITGRs. In the present manuscript we used two independent approaches to biomarker discovery: (a) a targeted ITGR sequence-based approach, examining 6 different ITGRs across a diverse library of *E. coli* isolates obtained from different animal hosts; and (b) publically-available whole genome data of 160 *E. coli* isolates isolated from different animals. The advantage of using a targeted ITGR sequencing approach was that a large number of isolates from diverse *E. coli* strains collected from an animal-host library could be readily examined. The disadvantage of this approach was that relatively few ITGRs could be examined at once, and for which little *a priori* knowledge on host-specificity existed. Conversely, the advantage of using whole genome data was that a large number of ITGRs could be examined simultaneously (i.e., n = 80 in this study), but with the disadvantage that relatively few *E. coli* whole genomes are available in NCBI from hosts other than humans.

2. Materials and methods

2.1. E. coli isolates

In total, 356 *E. coli* strains isolated from eight different host sources (Table 1) were selected from a previously established *E. coli* library containing 845 isolates (Zhi et al., 2015). All 356 isolates that were PCR positive for the three ITGRs used by Zhi et al. (2015) (*csgBAC-csgDEFG*, *uspC-flhDC*, *asnS-ompF*), as well as for three additional ITGRs (*cutC-torYZ*, *metQ-rcsF*, *araH-otsB*), were chosen for logic-regression-based biomarker analysis. The genes regulated by these ITGRs are as follows: (1) the *csgBAC-csgDEFG* region, regulating synthesis of curli fimbriae; (2) the *uspC-flhDC* region, regulating the expression of the master regulator of flagellum biosynthesis and universal stress response gene C; (3) the

Table 1
E. coli isolates collected from different host animals and PCR results for targeted intergenic regions.[a]

Host	No. of isolates	No. of PCR positive isolates (%)			No. of isolates used for logic regression model building
		cutC-torYZ	*metQ-rcsF*	*araH-otsB*	
Bovine	85	79 (92.9)	83 (97.6)	81 (95.3)	73
Cat	21	19 (90.1)	20 (95.2)	18 (85.7)	17
Dog	61	61 (100)	60 (98.4)	56 (91.8)	55
Goose	20	19 (95)	19 (95)	18 (90)	17
Human	105	101 (96.2)	105 (100)	104 (99.0)	100
Chicken	2	2 (100)	2 (100)	2 (100)	2
Pig	42	40 (95.2)	42 (100)	41 (97.6)	39
Gull	20	18 (90)	20 (100)	17 (85)	15
Total	356	339 (95.2)	351 (98.6)	337 (94.7)	318

[a] All isolates represented in the second column of this table were also PCR positive for *csgBAC-csgDEFG*, *uspC-flhDC*, *asnS-ompF* as determined by Zhi et al. (2015).