OPEN ACCESS Freely available online

PLoS GENETICS

Perspective

Order and Disorder during *Escherichia coli* Divergence

Heather Hendrickson*

Microbiology Unit, Department of Biochemistry, University of Oxford, Oxford, United Kingdom

> "… I was much struck how entirely vague and arbitrary is the distinction between species and varieties" — Charles Darwin, On the Origin of Species (1859)

Escherichia coli is a single species with numerous recognized roles, from lab workhorse to beneficial intestinal commensal or deadly pathogen. The extant strains have disparate lifestyles as a result of differential niche expansion since their divergence 25–40 million years ago, ten times longer than the estimated divergence between chimpanzees and humans [1,2]. Not only do these roles vary by strain (variant) of the species, but the recognition of a strain's role in one context does not exclude radically different behaviour in another, due to differential gene expression [3]. These are organisms adapting on evolutionary and lifetime scales to myriad environments and pressures. How do these strains differ from one another and what sustains their identification as a single species?

To address these questions, Touchon et al. have completely sequenced and annotated six strains of *E. coli* while re-annotating previously sequenced strains, as discussed in this issue of *PLoS Genetics* [4]. Comparative genomics analyses of 20 *E. coli* strains and one out-group provided insights into the contributions of horizontal gene transfer (HGT) and mutation on evolution in this species. In addition, the strains were tested in a mouse model to compare their virulence.

How Many Genes Could an *E. coli* Possibly Have?

It was known as early as 2001 that over 30% of the genes in *E. coli* O157:H7 Sakai, a dangerous pathogen, are unique to that organism, compared with the K12 laboratory strain [5]. With the expansion of the data (from two genomes to 20) carried out by Touchon et al., the stark nature of the potential for similarities and differences between strains is revealed. The regions that are similar, the 4.1-Megabase "backbone" of the genomes, are 98.3% identical at the sequence level. This is remarkable considering the time they have had to diverge. Outside of this backbone, genes are in flux as a result of HGT and deletion. If expressed, genes gained through HGT can provide entirely new capabilities for a bacterium, ranging from carbon utilization to toxicity [6].

The collection of all genes found in the *E. coli* strains sampled is called the "pan genome" (Figure 1). Touchon and colleagues have found the *E. coli* repertoire to be a staggering 17,838 genes. Individual strains have an average of 4,721 genes, and it is estimated that only 1,976 of these will be the "core genes" that (nearly) all *E. coli* strains have.

Consequences of a Large Pan Genome

Touchon et al. observe "… although some fundamental functions can be well studied by using a model strain, no single strain can be regarded as highly representative of the species" [4]. At first glance, this may seem a small point, but it calls into question a basic tenet of the International Code of Nomenclature of Bacteria, which still relies on the establishment of a "type species" that should not be "exceptional, including species which possess characters stated in the generic description as rare or unusual" (recommendation 20d.4). According to the authors, every *E. coli* strain is endowed with unusual characters, at least in terms of its gene content.

Large-scale genomic comparisons within a single species, particularly one with the range of lifestyles present in *E. coli*, have not been undertaken. Is *E. coli* atypical in terms of its catalogue of potential genes, or is it entirely normal? If such diversity continues to be observed at the single-species level, then we need to think carefully about what is meant by bacterial taxonomy [7].

Signs of Selection

In addition to showing how these *E. coli* strains are different, the authors elucidate what makes them similar. The aligned core genome (those genes shared by a majority of the strains studied) was analyzed for linkage disequilibrium. Touchon et al. noted evidence for a high level of gene conversion: any single nucleotide was 100 times more likely to be involved in a gene conversion than a mutation. Mutation causes gradual change in DNA sequence, whereas homologous recombination restores similarity.

Even though a huge flux of HGT was observed, entrance of new DNA across strains was not random. In the 21 genomes analyzed, 133 locations were found to accumulate 71% of all the non-core pan-genome genes. For the majority of these, the participation of phage or integrase was ruled out. The formation of such hotspots for genomic flux cannot be explained by any known mechanism. Touchon et al. suggest that, once a rare, large integration event disrupts chromosome order, perhaps this less perfectly adapted region opens the way to future events through a "founder effect" for additional HGT. Phylogenetic incongruence tests revealed two chromosomal regions that were recombination hotspots with large selective footprints, indicating that this variation was being maintained by selection: the *rfb* and *leuX-fimH* loci. This is in agreement with other studies in both *E. coli* and *Salmonella enterica* [8,9].

Citation: Hendrickson H (2009) Order and Disorder during *Escherichia coli* Divergence. PLoS Genet 5(1): e1000335. doi:10.1371/journal.pgen.1000335

Editor: Josep Casadesús, Universidad de Sevilla, Spain

Published January 23, 2009

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: heather.hendrickson@bioch.ox.ac.uk

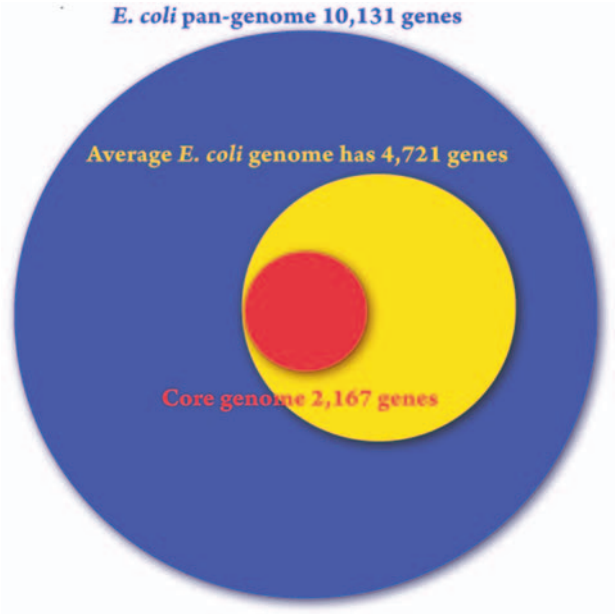

Figure 1. The breadth of genomic potential for *E. coli*. A Venn diagram of the pan-genome (blue), average genome (yellow), and core genome (red) of the sequenced *E. coli* strains according to Touchon et al. [4].
doi:10.1371/journal.pgen.1000335.g001

The terminus regions were found to have lower G+C percent contents than the rest of the genome, as well as a reduced ration of non-synonymous-to-synonymous polymorphisms and lower recombination rate. We may have much to learn about this region of the chromosome, potentially another example of the conflict between genomic flux and genomic organization.

A Species by Any Other Name

Is the nature of melange-like *E. coli* truly captured by referring to its variants as individual strains? Far from a simple semantic argument, we need new concepts in evolutionary microbiology to refer to and understand organisms possessed of truly chimeric chromosomes. Even as we grapple to understand the breadth of present-day *E. coli*, they continue to evolve at a breathtaking rate. Since the first detection of *E. coli* 0157:H7 in 1982, new sub-populations have emerged that have the capacity to cause even more serious illnesses [10]. Comparative genomics of the sort done by Touchon and co-authors unveils the complex evolutionary events taking place within these dynamic bacterial populations.

References

1. Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet 3(2): e7. doi:10.1371/journal.pgen.0030007.
2. Lawrence JG, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A 95: 9413–9417.
3. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, et al. (2008) Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. Proc Natl Acad Sci U S A 105: 4868–4873.
4. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5(1): e1000344. doi:10.1371/journal.pgen.1000344.
5. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, et al. (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8: 11–22.
6. Lawrence JG, Hendrickson H (2003) Lateral gene transfer: when will adolescence end? Mol Microbiol 50: 739–749.
7. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, et al. (2005) Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol 3: 733–739.
8. Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG (2004) Protozoan predation, diversifying selection, and the evolution of antigenic diversity in Salmonella. Proc Natl Acad Sci U S A 101: 10644–10649.
9. Weissman SJ, Chattopadhyay S, Aprikian P, Obata-Yasuoka M, Yarova-Yarovaya Y, et al. (2006) Clonal analysis reveals high rate of structural mutations in fimbrial adhesins of extraintestinal pathogenic Escherichia coli. Mol Microbiol 59: 975–988.
10. Kaper JB, Karmali MA (2008) The continuing evolution of a bacterial pathogen. Proc Natl Acad Sci U S A 105: 4535–4536.