Escherichia coli: From genome sequences to consequences

(or “Ceci n’est pas un éléphant...”)

Mark Pallen MA MBBS MD PhD FRCPATH

M Pallen. *Escherichia coli*: From genome sequences to consequences. Can J Infect Dis Med Microbiol 2006;17(2):114-116.

The present article summarizes a presentation given by Professor Mark Pallen of the School of Medicine at the University of Birmingham (Birmingham, United Kingdom) for the Fourth Stanier Lecture held in Regina, Saskatchewan, on November 9, 2004. Professor Pallen’s lecture, entitled ‘*Escherichia coli*: From genome sequences to consequences’, provides a summary of the important discoveries of his team of research scientists in the area of genetic sequencing and variations in phenotypic expression.

Key Words: *Comparative genomics; Darwin; Genetic sequencing; Genomics; Phenotypic expression; Proteomics; Stanier*

THE GENOMIC CHALLENGE: A DARWINIAN RESPONSE

The advent of the genomic era brings many new opportunities to the field of bacteriology. It is clear that ‘discovered biology’ before the genomic era encompassed but a small fraction of a much larger landscape. The majority of bacterial genes remain uncharacterized, and genome sequencing has revealed an unexpected abundance of horizontal gene transfers between and within bacterial species. However, comparative genomics has also shown that similar sequences have been reused by nature in numerous combinations in humans, flies, worms, yeast and bacteria, recalling Darwin’s dictum, “Nature is prodigal in variety, though niggard in innovation” (1).

Genomics also brings challenges, such as the risk of data overload and the dangers of oversimplistic interpretation of sequence information. One response to these challenges is a return to basics, that is to say, to the founding principles of biology as outlined by Charles Darwin 150 years ago in *The Origin of Species* (1), in which one of Darwin’s key points was that variation mattered in biology and that one should distrust the typological approach that sets one example of the species above all others as an archetype. Even now, how often do we hear that the *Escherichia coli* genome was sequenced in 1997,

Escherichia coli : Des séquences génomiques aux conséquences

Le présent article résume une présentation donnée par le professeur Mark Pallen, de la faculté de médecine de l’université de Birmingham (Birmingham, Royaume-Uni), dans le cadre de la quatrième conférence Stanier qui avait lieu à Regina, en Saskatchewan, le 9 novembre 2004. La conférence du docteur Pallen, intitulée « *Escherichia coli* : Des séquences génomiques aux conséquences » fournit un résumé des importantes découvertes réalisées par son équipe de chercheurs scientifiques dans le domaine du séquençage génétique et des variations de l’expression phénotypique.

when in fact just one atypical laboratory strain of a highly diverse species was genome sequenced!

Similarly, while Darwin stressed that one should expect imperfection in nature, the default assumption of many bacteriologists is that every gene in a genome must have a function. To quote Darwin (1):

On the view of descent with modification, we may conclude that the existence of organs in a rudimentary, imperfect and useless condition, or quite aborted, far from presenting a strange difficulty...might even have been anticipated and can be accounted for by the laws of inheritance.

Finally, Darwin almost appears to have anticipated the power of sequence homology when he describes patterns of morphological variation in nature (1):

Let two forms have not a single character in common, yet if these extreme forms are connected together by a chain of intermediate groups, we may at once infer their community of descent, and we put them all into the same class.

School of Medicine, University of Birmingham, Birmingham, United Kingdom

Correspondence and reprints: Dr Mark Pallen, School of Medicine, University of Birmingham, Room ELG41, Division of Immunity and Infection, Birmingham B15 2TT, United Kingdom. Telephone 0121-414-7163, fax 0121-414-3454, e-mail m.pallen@bham.ac.uk

Received and accepted for publication February 28, 2006

E COLI K-12 – A CALIFORNIAN INVENTION

Armed with these Darwinian principles, we have been exploiting genomics, bioinformatics and laboratory-based research to study *E. coli*. This bacterium is a common human gut commensal, yet also a model laboratory organism. The most commonly used laboratory strain, *E. coli* K-12, was isolated in 1922 from the stool of a convalescent diphtheria patient in Palo Alto – like PCR, it is a Californian invention! *E. coli* is also a pathogen, causing urinary tract infections, blood stream infections, meningitis and diarrhea. This phenotypic diversity is matched by an astonishing genomic variability, with more than one-third of the genome varying among strains.

Genomic diversity

Recently, we have addressed the problem of genomic diversity in *E. coli* by establishing *coli*BASE <<http://colibase.bham.ac.uk>> (2). Unlike other *E. coli* databases, the focus of *coli*BASE is on genomic diversity and on *E. coli* as a pathogen. The database now includes dozens of completed and unfinished genomes from *E. coli* and its relatives. The site allows rapid and user-friendly comparisons between genomes using MUMmer (a system for rapidly aligning entire genomes) and PROmer (a program used to generate alignments based on input sequences when species are too divergent for DNA sequence alignment to detect similarity). Thus, the site acts as a one-stop shop for *E. coli* sequence analysis. Most recently, we have generalized the *coli*BASE schema so that we can build similar facilities for other groups of bacteria <<http://xBASE.bham.ac.uk>>. A related focus of comparative genomics efforts has been the exploitation and development of long polymerase chain reaction-based methods to reveal similarities and differences between bacterial genomes (3,4).

Type III secretion: Questioning assumptions

The several pathotypes of *E. coli* that cause diarrhea include enteropathogenic *E. coli* and enterohemorrhagic *E. coli* (EHEC) (especially *E. coli* O157). Both of these pathotypes possess a special virulence factor – the type-III secretion system encoded by the locus for enterocyte effacement (LEE). Type-III secretion systems consist of five components – regulators, chaperones, the secretion apparatus, the translocators and the effectors – that enable bacterial effector proteins to be translocated through a molecular syringe into host cells, where they can then subvert host cell biology to the bacterium's advantage (5).

In the decade since the LEE was discovered, most scientists working on it have made two assumptions, albeit often unwittingly. First, it has been assumed that the LEE acts as a self-contained unit, encoding all of the effector proteins that are translocated through the associated type-III secretion system. Second, it has been assumed that many of the genes within the LEE have no homologues in other organisms or in other type-III secretion systems.

Recently, using careful application of bioinformatics approaches, we have challenged both assumptions. We have shown that many hitherto uncharacterized or poorly characterized genes within the LEE do indeed have homologues in other systems; most notably, we have established sequence homology between an LEE-encoded protein, SepL, and two well-known proteins, YopN and TyeA, from a type-III secretion system in *Yersinia* (6).

In a similar vein, using bioinformatics analyses, we have uncovered several dozen new candidate effector genes in EHEC, more than doubling the number of potential or proven effectors in this organism. These genes are scattered throughout the EHEC chromosome, but often cluster in prophage genomes. We are currently verifying our predictions by showing that these putative effectors are indeed injected into host cells. A major challenge for the coming years will be to assign functions to them all.

E COLI TYPE-III SECRETION SYSTEM 2: A DEGENERATE SYSTEM PRESENT EVEN IN THE ANCESTOR OF K-12

In addition to the LEE, the EHEC genome sequence revealed a gene cluster thought to encode a second unsuspected and cryptic type-III secretion system, termed '*E. coli* type-III secretion system 2' (ETT2). Initial assumptions were that the ETT2 genes represented an insertion into the EHEC genome relative to the laboratory strain K-12 and that they encoded a functional type-III secretion system. Through careful genomic comparisons and analyses, we overturned both of these assumptions, showing that the model strain encoded remnants of the ETT2 gene cluster and that the EHEC ETT2 gene cluster was riddled with mutations that abrogate function (3). Furthermore, we established that ETT2 genes are present in the majority of *E. coli* strains but have usually suffered mutational attrition. In addition, they show no link to virulence, but instead reflect the phylogenetic origin of the strain. Only in one strain, the genome-sequenced 042 strain, did we find evidence of an ETT2 cluster that may be functional. This same strain also contained a second related type-III secretion locus containing translocation genes absent from the main ETT2 gene cluster.

THE GRIN OF THE CHESHIRE CAT

Having overturned the assumption that the ETT2 cluster encodes a functioning type-III secretion system, we fell into believing that the whole cluster represented a nonfunctional 'baggage of history'. However, this fresh assumption was soon overturned with the discovery that two regulators encoded within the ETT2 gene cluster regulate the expression of genes within the LEE (7)! We have termed this effect, where regulatory or other influences outlive the decay of most of a functionally cohesive gene cluster, the 'Cheshire Cat Effect', after a cat in Alice in Wonderland, whose grin outlived the rest of the character.

FLAG-2: A SECOND EVOLUTIONARY SURPRISE

All but one of the genome-sequenced *E. coli* strains contain a curious pair of apparently orphan flagellar genes, *fliA* and *mbhA*, which sit adjacent to each other but diverge from a promoter-less start point. As with ETT2, through careful comparative genomics, we have established that they are, in fact, remnants of a large gene cluster potentially capable of encoding a second, previously unsuspected flagellar system in *E. coli*, which we have termed 'Flag-2' (4). Surprisingly, a seemingly intact Flag-2 cluster occurs in approximately 20% of *E. coli* strains. Unfortunately, a functioning version of the system has so far proved elusive. However, several distinctive characteristics of the Flag-2 system make it of interest even in the absence of an intact example: it appears to be dependent on the alternate sigma factor RpoN; it encodes a flagellum

more akin to lateral flagella than to the conventional flagellum; and it appears not to be subject to the same extreme antigenic variation that characterizes the conventional flagellar system.

CONCLUSION

Darwin versus Monod in the postgenomic era

One of Roger Stanier's final publications, in 1977, was the obituary of the Nobel laureate Jacques Monod (8). A few years earlier, in December 1972, long before the genomic era, Monod made a memorable Delphic utterance, "Tout ce qui est vrai pour le Colibacille est vrai pour l'éléphant" (9), which translates roughly into English as "All that is true for *E coli*, is true for the elephant". By contrast, in the postgenomic era, our bioinformatics and laboratory-based studies lead us to conclude that what is true of one strain of *E coli* is not even true of another strain from within the same species! This calls to mind Darwin's admonitions from *The Origin of Species* (1):

No one supposes that all the individuals in the same species are cast in the very same mould ... There are not many men who will laboriously examine internal and important organs and compare them in many specimens of the same species.

This adds fresh impetus to efforts to sequence ever more strains from this highly varied species, *E coli*.

ACKNOWLEDGEMENTS: The author thanks Dr Morris Goldner, Professeur associé de biologie médicale, Faculté de médecine, Université Laval and President of the Stanier Institute/Institut Stanier, for his assistance in arranging and supporting the regular Stanier Symposia. The author also thanks Shawna Bourne, Resident Writer and Editor for the Stanier Institute/Institut Stanier, for reviewing and formatting this manuscript for publication.

REFERENCES

1. Darwin C. *The Origin of Species*. Boston: Harvard University Press, 2005.
2. Chaudhuri RR, Khan AM, Pallen MJ. *coliBASE*: An online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res* 2004;32:D296-9.
3. Ren CP, Chaudhuri RR, Fivian A, et al. The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J Bacteriol* 2004;186:3547-60.
4. Ren CP, Beatson SA, Parkhill J, Pallen MJ. The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. *J Bacteriol* 2005;187:1430-40.
5. Pallen MJ, Beatson SA, Bailey CM. Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: A Darwinian perspective. *FEMS Microbiol Rev* 2005;29:201-29.
6. Pallen MJ, Beatson SA, Bailey CM. Bioinformatics analysis of the locus for enterocyte effacement provides novel insights into type-III secretion. *BMC Microbiol* 2005;5:9.
7. Zhang L, Chaudhuri RR, Constantinidou C, et al. Regulators encoded in the *Escherichia coli* type III secretion system 2 gene cluster influence expression of genes within the locus for enterocyte effacement in enterohemorrhagic *E. coli* O157:H7. *Infect Immun* 2004;72:7282-93.
8. Stanier RY. Jacques Monod, 1910-1976. *J Gen Microbiol* 1977;101:1-12.
9. Jacob F. *The statue within: An autobiography*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1998.