



◀ Lawrence Lawry / Science Photo Library

What is a bacterial species? **W. Ford Doolittle** discusses how genome data are giving microbiologists cause to think carefully about how to define the 'species'.

Biologists have long struggled with the problem of defining and recognizing species. Even animal and plant species are sometimes difficult to delimit, and philosophers disagree about the ontological referents of the word. But microbes – especially prokaryotes – seem to pose special problems. Their small size and general uncultivability (only a very small percentage can grow in the lab) confound efforts to describe and archive type specimens. Worse, prokaryotes reproduce asexually and are thus in principle unable to conform to Ernst Mayr's popular Biological Species Concept – 'groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups'.

Thus microbiologists have in general been willing to admit that practical needs for identification and naming might best be met by some widely agreed-upon but provisional (and at least in some sense arbitrary) species definition, while the more theoretically driven search for a unifying species concept that would explain patterns of microbial diversity in ecological and population genetic terms could wait for the accumulation of more data, especially gene and genome sequence data. Operationally, molecular definitions have won the day, and species are usually expected to share at least 70% binding in standardized DNA–DNA hybridization and/or over 97% gene-sequence identity for 16S ribosomal RNA (rRNA) (see the article by Stackebrandt & Ebers on p. 152).

#### Diversity within diversity

As a bonus, molecular methods allow us to identify and enumerate species in the environment without isolating and cultivating any organisms – for instance, using specific PCR primers to amplify and sequence 16S genes from unfractionated environmental DNA preparations. What we often find is an astonishing number and diversity of apparent species, with few 16S sequences assignable at the 97% level to any cultivated isolates – including isolates from the same sampling site. Moreover, many species, as defined by the 97% 16S identity cut-off, will themselves be represented by multiple different but similar individual sequences in any sample (Fig. 1).

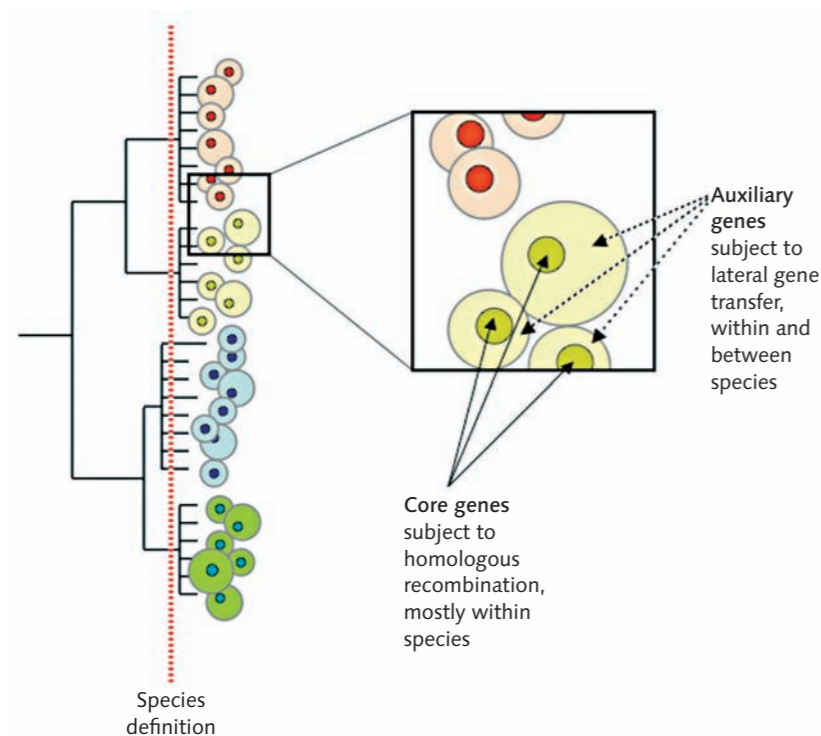
Such gene sequence 'microdiversity' is the rule rather than the exception

with environmental sampling of many genes (in addition to 16S), and may be matched by another kind of variation at the level of genome composition (gene content). Martin Polz and collaborators have used pulsed-field gel electrophoresis to show that *Vibrio splendidus* isolates (with >99% 16S sequence identity) from a sample site on the Massachusetts coast can differ by as much as a megabase in size, comprising 'at least a thousand distinct genotypes, each occurring at extremely low environmental concentrations (on average less than one per millilitre)'.

Gene content diversity has also emerged as a principal message of more 'traditional' complete genome sequence studies, based on cultivated isolates. When such activities began, a decade ago, the thought was that one

sequence (that of strain K12) would surely be enough to define *Escherichia coli*, one would do for *Bacillus subtilis*, and so forth. Completion of the second *E. coli* (O157:H7) gave us a shock. This sometimes lethal food contaminant proved to have 1,387 genes not present in K12, scattered in hundreds of small or large clusters around its genome. (Reciprocally, K12 had 528 genes not in O157:H7.) The two genomes were otherwise (aside from one inversion) colinear, exhibiting 98% average nucleotide identity (ANI) between shared genes.

Now that there are more than three dozen species with more than one strain sequenced, results like this seem to be almost the norm. Strains which are very close on the basis of the sequences of the genes they do share



◀ Fig. 1. Diversity and microdiversity. Trees based on 16S rRNA or other phylogenetic marker genes sequences often show 'microdiversity', exhibiting clusters of sequences more closely related than the value accepted as defining species (vertical hatched red line). Even in species so defined, genomes can show substantial (up to 30 %) variation in size and gene content. A 'species genome' or 'pangenome' can be imagined to comprise core genes shared by all its strains, and a set of auxiliary genes found only in some strains. These two classes of genes may have different evolutionary modes and tempi. *W.F. Doolittle*

may nevertheless differ by up to 30 % in gene content, the differences being attributable to scores or hundreds of events of gene gain (mostly by lateral gene transfer) and gene loss after divergence from a common species ancestor. Although many variable sequences are phages or transposable elements, others are genes vitally important in defining a strain's specific niche. Sometimes such genes are transferred together: 'pathogenicity islands' exemplify this, but the phenomenon is not limited to pathogens. Endosymbiotic nitrogen-fixing strains of *Mesorhizobium*, for instance, possess an approximately 500 kb 'symbiosis island' which encodes not only the dozen and more genes needed to form root nodules, but most or all genes needed for nitrogen fixation and the island's own strain-to-strain transfer.

It has recently become popular to think in terms of 'species genomes' or 'pangenomes' (Fig. 1), consisting of a core or backbone of genes shared by all strains and an auxiliary or flexible gene pool found only in one or some strains. For some groups, the number of such auxiliary genes already exceeds the number of genes in the core, and just keeps on climbing with each new genome sequenced. Core genes in contrast get fewer with each new genome. But still there will be hundreds to thousands of core genes for any group we might want to call a species, and they will usually comprise a colinear backbone, within which laterally transferred genes and islands can be seen to be embedded. We might use concatenated (strung-together) core gene sequences to define species with greater precision and nuance than either

DNA-DNA hybridization or 16S allow. Konstantinidis & Tiedje noted that an ANI value of greater than 94 % for core genes characterizes most species defined by other means.

We might also use phylogenetic trees of concatenated core gene sequences to establish lineage relationships of strains within a species, although for this, homologous recombination poses a complication. Increasingly, many bacteria (and some archaea) turn out to avidly indulge in between-strain homologous recombination. Recombination means that different genes will have different evolutionary histories, and there will be no unique phylogeny relating the genomes of different strains of a species. The rate of recombination in bacteria will of course never approach that of animals, who must recombine every time they reproduce, but still it can exceed mutation as a generator of evolutionary novelty. This raises the possibility that the Biological Species Concept might appropriately be applied to some bacteria after all, at least in so far as it entails sharing of core genes by recombination in a common gene pool.

### Getting the concept

Homologous recombination is one process that confers genomic coherence (within-species similarity and between-species divergence) in prokaryotes, just as it does in animals. Members of species A will more closely resemble each other than they do members of a sister species B, because at most core gene loci they share alleles that have arisen within the A gene pool. Periodic selection is another coherence-generating

*We will be more likely to realize a full understanding of microbial diversification if we accept that the word 'species', for all its utility, may have no precise referent.*

process, which will also homogenize gene content. In this mode, favourable mutations sweep to fixation within a physically or ecologically bounded finite asexual population, carrying the rest of the genome in which they first occurred (and whatever auxiliary genes it might bear) along for the ride. Mutations at other loci than that under selection may of course occur along the way, and if homologous recombination is frequent, in the end only the favoured mutation – not the genome in which it first appeared – will achieve fixation. In different groups and at different times these two forces will vary in strength to promote coherence in gene sequence and/or gene content, but there is no guarantee that either will maintain or create sharp species boundaries. Divergence in sequence suppresses recombination, but recombination with divergent sequences might often offer more significant selective advantage. Agents of genetic exchange (phages and conjugation systems) have recognition specificity, but as selfish elements are under pressure to broaden, not restrict, host range. Sometimes genetic processes and ecological selective forces may create groups as genomically coherent as animal species, but there is no compelling theoretical argument that they must always or usually do so, and so far there are insufficient data to say that they generally have.

Lateral gene transfer is no respecter of species boundaries and disrupts genomic coherence, however defined or achieved. More to the point, transfer is often the source of those genes whose

expression in phenotype most strongly differentiates closely related taxa, and concerns us most as practising microbiologists. So molecular definitions of species based on whole genomes may be only poorly coupled to the ecological drivers of the processes of diversification and adaptation we would like to think of as speciation. For this reason Konstantinidis & Tiedje suggest that microbiologists might better bring their ideas about species in line with those of zoologists by 'including only strains that show a >99 % ANI or are less identical at the nucleotide level, but share an overlapping ecological niche...' (emphasis mine). They are willing to be flexible about using overall genomic coherence in recognizing species, when phenotypic differences (which will often be due to laterally transferred genes) seem to warrant it. Our current species definitions and the more genomically based criteria likely to be adopted soon serve their practical purposes as well as we can expect, but there can be no very precise mapping to any unitary model of diversification and adaptation. That is, it seems unlikely that we will soon, if ever, have a uniform species concept that will allow us to give unqualified (definition-independent) answers to such questions as how many prokaryotic species are there at some particular site, or in the whole world. It's not that there is no hope for a full understanding of microbial diversification, adaptation and dispersal – it's just that we will be more likely to realize that hope if we accept that the word 'species', for all its utility, may have no precise referent.

**W. Ford Doolittle**

Canada Research Chair, Dept of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 1X5 (e ford@dal.ca)

### Further reading

- Hanage, W.P., Fraser, C. & Spratt, B.G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol* 3, 6.
- Konstantinidis, T.K. & Tiedje, J.M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102, 2567–2572.
- Perna, N.T., Plunkett, G., Burland, V. & others (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 410, 529–533.
- Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M. & Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci U S A* 103, 12115–12120.
- Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J., Sarma-Rupavtarm, R., Distel, D.L. & Polz, M.F. (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307, 1311–1313.