A rose by any other name . . .

It appears to me that at the moment the responsibility for naming genes is separately vested in the traditional published literature, and in the databases and genome sites. It needs to be rationalised.

In 1998, I proposed renaming the gerCA, gerCB and gerCC ‘spore germination’ genes of Bacillus subtilis as hepA, menG and hepB to reflect their actual function in menaquinone biosynthesis (1). But as some of these names have been used for different functions in E. coli, and in deference to the annotators in SWISS-PROT and Subtilist, their names should now be – or have been – changed to hepS, menH and hepT.

Members of the bacterial genetics community are used to having the responsibility of defining the name of a new gene in the primary publication, with the agreement of the journal’s referees. The name reflects, to a greater or lesser degree, the function of the product or at least the mutant phenotype. Following the rules of Demerec et al. (2), the capital letter has often served only to distinguish cistrons, or open reading frames, reflecting the degree of ignorance of detailed function. Criteria were (and are) as trivial as which letters had been used already in that organism, or what the order of genes in a cluster might be, as the genetics usually predated biochemical studies.

It is not uncommon for genes with the same name from different bacteria to encode products with different functions, and the name–function relationship might even be reversed, as in the case of the leuB and leuC genes of Escherichia coli and B. subtilis. Some genes would have two alternative names, as two labs converged on different aspects of the mutant phenotype, and genes could be renamed in later publications by other authors, if the referees thought it acceptable.

When studying a single metabolic pathway or organism, such anomalies could be learned and lived with, both in literature searches and in the use of mutant strains. No longer – we are in the genome age now. We are fortunate that sequence annotators, for example, in my direct experience the Ecogene, Subtilist, and SWISS-PROT teams, have created and continue to improve sites that bring much useful information together in a rational format. These are brave men and women that attempt to bring some logic and order to gene and protein sequence collections, in order to avoid amplifying ambiguities as new open reading frames are named on the basis of their similarity to already characterized gene products. Classifying protein products requires even more ingenuity, as gene nomenclature is very different across groups of organisms; the short protein ‘entry names’ in SWISS-PROT may increasingly serve to guide the user to possible function, and/or to protein class – for example, the leuB gene product of E. coli and its orthologue in B. subtilis have the ‘entry names’ LEU3_ECOLI and LEU3_ BACSU respectively.

Gene and protein names are being revised in genome databases, with consultation between their curators, and generally for very good reason. The names of the leuB and leuC genes of B. subtilis are now being reversed in Subtilist, for example, to make sure that, whenever possible, orthologues between E. coli and B. subtilis share the same gene name. But are such changes obvious enough? How should the literature reflect and record them? Is the journal-based literature on gene and allele names, with its wealth of associated phenotypes, searchable by title and keywords, at risk, for a few genes at least, of becoming out of step with the sequence databases and genome sites as names change?

There are references to the alternative names, and links to the literature in these websites, but it will need some care in the coming years to ensure that a mutant allele remains easily traceable and can also be correctly correlated with the current genetic (or do I mean genomic?) nomenclature. I can see several examples close to my research area – you will know whether it applies more widely...

If there is a significant risk of confusion, I have a few suggestions. The current accepted gene name should be defined internationally – is it the latest one published in the refereed literature, or a more recent one adopted in a prescribed genome database? The databases take care to list alternative gene names, but perhaps a date could be associated with changes in name, along with the reason for the change, so that future researchers can mine the literature effectively. The compilers of these database sites welcome comments and improvements, and clearly more dialogue is required. Proposed changes might be posted by annotators, to encourage consultation. We probably need to be more proactive in checking and commenting on the entries for genes...
we identified or study, discussing nomenclature as necessary. Finally, should journals develop instructions for authors that reflect the need for accord with the nomenclature in databases? Clarity is all, and it is at risk.

An annotator’s view

Moir makes some excellent points in her commentary, none of which I would disagree with. I would like to add to, and hopefully strengthen, some of these from the point of view of a sequence annotator.

Naming of genes used to be performed by researchers who had often spent many years studying the pathway, or mutation, defining the gene, and would thus be in an excellent position to choose a fitting and descriptive name. As Moir describes, a wealth of subsequent literature often hangs on these names which can often be extremely evocative. This is not, however, always the case. I recently came across a family of genes from different organisms, all called gufA, and each confidently described as encoding ‘GufA protein’. When I tried to trace the source of this designation, the function was unclear until a careful reading of the original manuscript revealed that the name stood for ‘gene of unknown function A’ (1). This kind of whimsical naming extends even to such well-studied organisms as *E. coli*, where the final coding sequence of the genome has been named *lasT* (2). The protein encoded by this gene is a member of a widely-conserved family of methyltransferases, few of which have, so far, acquired the same gene name.

Many of the problems Moir describes are a consequence of the pace of modern genomic sequencing, which means that a sequence annotator will often look at, and name, several tens of genes per day. Having the complete sequence of a genome does confer at least one advantage – all the genes can be systematically named (or numbered), giving an unequivocal label to each gene within any particular organism. Annotators will often attach functional gene names, where these can be confidently assigned, but these should always be in addition to the systematic name. Many genes now have multiple names, and this should not be a problem where all of these names are cross-referenced within databases. Of course the annotator’s source for these additional gene names is usually the primary databases (EMBL/GenBank/DDBJ), but in the cases where gene names have been subsequently changed in the literature, these changes are often not fed back to these databases, leading to the propagation of old names in the genomic annotation.

The solution to these problems lies, as Moir suggests, in greater communication between the microbial geneticists in the field, and the genome annotators and database curators. Genome annotations should not be set in stone, but should be seen as a ‘first pass’ and subject to continual updates and modifications. Genome databases are, in the long run, likely to become the *de facto* primary source of information on the genetics of the organism. Where annotations are at odds with the literature, they should be corrected, ideally through communication between the genetics community and the database curators. Where the gene designations have been changed for good reason, this needs to be recorded, and previous names retained as synonyms where necessary.

All of this underlines the requirement for continuous curation of genome databases, with the active participation of the communities working on the organism in question. If, as is increasingly the case, the genetics community is using these databases as a source of information, it is in their best interests to ensure that that information is as up-to-date and correct as possible.

Julian Parkhill

The Sanger Centre, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.

Tel: +44 1223 494 975. Fax: +44 1223 494 919.

e-mail: parkhill@sanger.ac.uk
