nature structural & molecular biology

Name that gene!

Scientists coin new terms, or neologisms, at a tremendous pace, but name choice can have unforeseen results.

Certain biologists (along with club-hoppers, flip-floppers and sabermetricians) felt the glow of literary respectability earlier this year: the words apoptosis, protein kinase and primosome passed the scrutiny of the usage police (along with club-hopper, etc.) and entered that hallowed tome, the Oxford English Dictionary (OED). Thousands of words are considered for the OED every year, and deciding which enter and become 'legit' members of the language involves massive evidence collecting, including proof of established usage by independent sources over a reasonable time span. This is clearly a world far removed from biology, where new terms, particularly gene and protein names, are coined at a dizzying pace that can have unforeseen repercussions.

When disparate fields assign protein names, the results can be surprisingly convergent, especially for abbreviations: is this APC the anaphase-promoting complex, or is it adenomatous polyposis coli? And maybe this PAR is a GPCR activated by proteinases. Wait, no, it's involved in asymmetric cell divisions in *C. elegans* embryos. And substituting NOS (nitric oxide synthase) for Nos (Nanos) would not be good for a fly's abdomen. Confusion can arise verbally as well as on paper: if someone says 'risk' to you, you don't necessarily know what they're into—remodeling (RSC) or repression (RISC)?

Even within fields, nomenclature can be initially daunting. Studying the cell cycle and immune system at college may feel like swimming in a sea of Cs and Ds. Opinions differ on whether cataloguing of *CD* or *unc* one through gazillion is better, or whether out-and-out descriptive names are more accessible. And when it comes to the latter, nobody thinks up monikers like a *Drosophila* geneticist. Contrary to suspicion, these gene names are not designed to confuse, but reflect the mutant phenotype—so *Krüppel* describes the stunted mutant embryos, and *eyeless* is, well, pretty gruesome. However, there is the potential to be jargony and subject to the fleeting zeitgeist ('son of pop-culture icon' is epistatic to 'pop-culture icon').

One problem for editors and fields alike is the situation of homologous genes, different organisms, different names. Without knowing the identity of a gene caught in a genetic screen, you can't know that the homolog has a name. Hence, one man's *NOTCH* is another worm's *lin-12*. Of course, there is also the touchy situation of same gene, same organism, different lab, different name. At one point, *EPHB2* masqueraded under at least eight aliases. A nomenclature meeting eventually resolved that one.

Names for genes associated with human diseases are reminders of who described a disease and crucial for keeping disease associations straight, but there can be a certain longing for a name reflecting function (for example, *pasha*, or *partner of drosha*) as you try to get your mouth around the human version (*DGCR8*, or DiGeorge syndrome chromosomal region-8). However, whimsical names don't necessarily translate well in the clinic. The vertebrate hedgehog genes initially carried tongue-in-cheek names (such as sonic hedgehog). However, mutations are associated with human craniofacial abnormalities, so such names are discomforting for clinicians and patients alike. The Human Genome Organization (HUGO) weighed in last November, approving the abbreviations over the full names (so it's *SHH* for sonic hedgehog). Some would argue though that the abbreviations were already in place for clinical usage and that the field at large can police itself when it comes to appropriateness.

So who coordinates and checks gene names? Sometimes those with the copyright: both Velcro and Pokemon had to be dropped as names after the relevant companies picked up the phone. However, with the advent of the genome projects, the name game was changed by the flood of gene models needing identifiers. Genome databases are thus in an ideal position to both catalog genes of unknown function and examine new submissions, and they all have naming guidelines based on historical precedent for the species. The first test for an *Arabidopsis* gene name, for example, is checking for overlap with previous names (overlap with genes in other species is not checked). At the Zebrafish Information Network (ZFIN), authors are encouraged to name genes according to human and mouse orthology, but there the curators had to initially untangle some nomenclature confusion caused by incomplete sequencing. Although it seems that the plant curators tend not to receive inappropriate names, they have to deal with multiple claims being staked on the same *Arabidopsis* gene. In this case, they check which name has gained established usage in the field, or they find an amicable solution while keeping a record of all published names. At some level, this is not so different from compiling the OED.

Meanwhile, thousands of genes of unknown function are catalogued in the various organism databases, often coded according to which genome project sequenced them and patiently awaiting functional elucidation and naming. It's enough to make the most ardent neologist run out of ideas. The advent of nomenclature checks and even committees is inevitable, though increasing coordination between databases to clarify naming and reduce potential confusion is a welcome development. Along those lines, it is perhaps fitting to end with the words of the first lexicographer, Dr. Samuel Johnson, who back in 1752 sagely noted that

> "Among those who have endeavoured to promote learning and rectify judgment, it has long been customary to complain of the abuse of words, which are often admitted to signify things so different, that, instead of assisting the understanding as vehicles of knowledge, they produce errour, dissension, and perplexity…."

Perhaps a little strongly put, but clearly a discussion point for the ages.

Thanks to Dr. Huala at The Arabidopsis Information Resource and Dr. Van Slyke at the Zebrafish Information Network for information on database naming procedures.