Gene names: the approaching end of a century-long dilemma

The trouble started, of course, with Thomas Hunt Morgan. Yet, it would be unfair to blame him. Who amongst us, faced with the same problem, would not have done precisely the same? Nevertheless, he began a tradition whose legacy has been nearly a century's worth of subtle confusion and misunderstanding.

Morgan's problem seemed simple enough. He had to choose an appropriate name for the first discovered mutant gene. The mutation caused the production of white-eyed fruit flies, either when in single copy (hemizygotes) in males or, when present in two copies (homozygotes) in females. (The gene is sex-linked, that is carried on the X chromosome.) What should he call the mutant gene? Well, the choice was obvious: *white* or, as abbreviated, *w*. The corresponding normal form or allele, the "wild-type", Morgan designated w^+.

Thus began the long tradition of naming genes with respect to their mutant phenotype. In effect, the role of the wild-type gene, the form fashioned by evolution, comes to be defined by what happens when its activity is *not* available to the orgenism. This, as anyone can see, is a highly indirect, and therefore dubious, way to characterize a gene. Perhaps the back-to-front character of this procedure is not too serious when the mutant phenotype is a direct consequence of the absence of the wild-type gene product, as in the case of *w*. It becomes highly problematical, however, when the mutant phenotype is the outcome of a long and complex series of events. An analogy, from a tale attributed to Benjamin Franklin, illustrates why. The story recounts a simple chain of events: for want of a nail, a horseshoe was lost; for want of a horseshoe, a horse was lost; for want of a horse, a rider was lost; for want of a rider, a message was lost; for want of a message, a battle was lost; for want of a victory, a kingdom was lost. The idea that big consequences can follow from small events is absolutely valid. But to conclude from the story that the function of horseshoe nails is to prevent the loss of kingdoms would be silly. Many gene names, however, have equivalent implications.

There is another kind of confusion that creeps in. Many genes are named for their *first discovered* mutant phenotype. Long use of that gene name subtly reinforces the notion that *the* function of the wild-type gene is to participate in the process designated by the name. Apart from the fact of indirect consequences described above, two difficulties arise. First, many gene products have multiple uses. Second, where there are different roles—either from different usage or from multiple biochemical roles—these can be differentially affected by different mutations. When one considers these additional factors, it is easy to see just how large the potential for confusion associated with a particular gene name is. This situation is illustrated in the "Problems & Paradigms" article by Justin Courcelle, Philip Hanawalt and Ann Ganesan, in this issue. They discuss how the naming of the *recA* gene of *Escherichia coli* has shaped perceptions of the function of this gene product and why some rather different interpretations of its roles may now be in order.

The fundamental problem is that newly-identifed genes do require names, to permit communication about them, but that these names have tended to do double-duty. They have been picked to act as both *signifiers* and as *descriptors*. As a signifier, a gene name has to identify unambiguously a particular gene, such that it is distinguished from all previously identified genes. As a descriptor, the gene name is intended to tell you something about the properties of the gene. When mutant phenotypes were all that there was to go on, these were, inevitably, drawn upon to provide quick, shorthand descriptors. Yet, for all the reasons enumerated above, such descriptive elements are both incomplete and misleading.

In a sense, this situation has been bad enough for biologists, in creating subtle distortions in the ways that they think about the functions of particular genes. But it has had an additional spill-over effect in popular perceptions about genes and gene actions, which may be even more serious. In the early '90s, in particular, there was a host of claims, based on statistical evidence, that there were specific genes "for" particular traits, ranging from shyness and thrill-seeking to gender-preference and even criminality. The implication was always that they were special single genes "for" these traits. In virtually none of the press reports was it made clear that the scientists were talking about allelic variants of particular normal (wild-type) genes that might, indirectly and subtly, cause some alteration. The result is that many members of the public apparently think that there are specific aberrant genes that exist only in particular individuals rather than variant alleles that might affect certain behaviors in certain circumstances. In most of these cases, the neuro-developmental subtleties involved in these situations were completely swept under the carpet, at least in the journalistic accounts. This stuff is not only pernicious in its support of rigid genetic deterministic thinking but it has considerably increased public confusion about genes and the basis of mutant effects.

Hence, the tradition of gene-naming that Morgan unintentionally founded has a lot to answer for. The good news, however, is that this era of confusion is probably coming to a close. The agent of this change is genomics. Ideally, in future,

the descriptive element attached to each gene will flow naturally from the molecular-biochemical properties of the gene family to which it belongs, as is already frequently the case. Furthermore, since new genes are increasingly being identified through genomics rather than from their mutant effects, gene names will be predominantly given on this basis. Even when the name of the gene family was derived from either a mutant or a biochemical/cell biological test, the repeated use of the acronym instead of the full name should dull any potentially misleading associations. For instance, when a new FGF is found and named (usually by the family name plus a number, e.g. FGF-23), one is unlikely to assume automatically that its primary function involves either fibroblasts or promoting growth. As more and more genes are identified through genomics and placed in gene families, the tendency to assume that the new gene has anything to do with the first-identified function in that gene family is bound to diminish.

Yet, many new genes will, at least initially be identified through their first-discovered mutant effects or cellular roles while others, bearing no resemblance to known genes, will beidentified through gene sequencing. How should one name such new genes? At least two possibilities suggest themselves. The first is that, when the gene has been identified through a mutant effect, a humorous or whimsical name, only remotely connected to the phenotype, be used. The *hedgehog* gene, named for the appearance of fruit fly embryos mutant for that gene, is a good example—no one will be tricked by the name into thinking that the gene acts to specify hedgehogs— while *sonic hedgehog*, a member of the same family but named after a cartoon character, extends the tradition.

The second possibility is that words from languages other than English be used. While many of the mutant phenotype-related names from French, German and Japanese that have flooded into the genetics literature over the past two decades have associations for people who speak those languages, respectively, those names will serve purely as signifiers, lacking confusing descriptor-aspects, for most English-speaking scientists. Another advantage of using words from other languages is that the supply is virtually limitless. No matter how many novel genes are picked up in genomics projects, the world's languages are equal to the task. Think of the 700 or so languages believed to exist in Papua-New Guinea . . .

That last possibility, though suggested tongue-in-cheek, is linked to a more general idea. Perhaps, if geneticists were to start consciously mining foreign languages for new gene names, one consequence would be an increasing exposure to languages and linguistics. Linguists, comparably, with their interest in information transmission are increasingly intrigued by genetic phenomena. Such cross-disciplinary interest has been developing, in both directions, for at least a decade. (For a recent example, see Howe, C. et al. (2001). Manuscript evolution. *TIG* **17**, 147–152.) Yet, there is bound to be more scope for fruitful interactions between geneticists and linguists. It would be a particularly satisfying conclusion to a near-century's worth of semantic muddle in genetics, involving gene names, if an improved and enriched relationship between genetics and linguistics came into being, as we approach the centenary of the discovery of Morgan's white-eyed fly.

ASW