Correspondence

Uncovering trends in gene naming

Michael R Seringhaus[†], Philip D Cayting[†] and Mark B Gerstein[*† ‡]

Addresses: [*]Program in Computational Biology and Bioinformatics, [†]Department of Molecular Biophysics and Biochemistry, and [‡]Department of Computer Science, Yale University, Whitney Avenue, New Haven, CT 06520, USA.

Correspondence: Mark B Gerstein. Email: mark.gerstein@yale.edu

Published: 31 January 2008

*Genome **Biology*** 2008, **9:**401 (doi:10.1186/gb-2008-9-1-401)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/1/401

Abstract

We take stock of current genetic nomenclature and attempt to organize strange and notable gene names. We categorize, for instance, those that involve a naming system transferred from another context (for example, Pavlov's dogs). We hope this analysis provides clues to better steer gene naming in the future.

Certain scientific discoveries confer the privilege of coining a lasting name. In biology, curious identifiers abound, ranging from playful acronyms for biosoftware tools to Latinized species names and tongue-twisting descriptive labels. The most problematic of these are gene names. The prevalence of silly and awkward gene names has been discussed before [1-3]. Critics lament the lasting implications of whimsical naming, particularly when the gene in question turns out to be involved in human illness. How would you react when an oncologist explained that your *Pokemon* mutation, say, gives you mere months to live?

Unusual or inconsistent gene names raise other concerns as well. In the genomics era, the genetic harvest of entire species is now processed *en masse*, and genetic information is accessed via massive databases where the context of clever or conflicting names is easily lost. As a consequence, initial cuteness has bred current confusion. Here, we take stock of the current genetic nomenclature and attempt to systematize these strange and notable names.

A survey of nomenclatures: genes, streets and stars

For human observers, the genome is merely the latest example of a vast, partially understood landscape of objects to label. Pioneering explorers, cartographers, astronomers and city planners all faced a similar task, generating nomenclatures with a variety of coherence. Street names within cities provide a good analogy to genes within species. North American cities, for instance, often share a corpus of conserved street names, some of which convey useful information (consider Church Street or College Street). This is reflected in the commonsense naming of certain 'landmark' genes and proteins (for example, those for ribosomes) in most species. Of course, it is possible for names to lose their meaning (though the church is demolished, Church Street remains). Similarly, some genes whose names describe their function later turn out to perform an entirely different activity.

A central-planning approach to naming, in cities and genes alike, is sensible but often bland. Genes in many newly sequenced organisms are named according to a rigorous system (for example, sequential open reading frame numbering), just as certain newer cities adhere to a rational system (consider Pierre l'Enfant's plan for Washington DC or the numbered grid of Manhattan).

In astronomy, the names of the primal heavenly bodies - the Sun and the Moon - come down to us from prehistoric times. The constellations were named several thousand years ago on the basis of their semblance to animals and mythical beings. Roman astronomers offered up names of deities for the planets in accordance with their observed characteristics. This nomenclature sufficed until increasingly powerful telescopes revealed unending swathes of astral objects to name. Accordingly, celestial nomenclature evolved into a pseudo-consistent system of numbered galaxies, stars and other objects. The resulting bricolage of astronomical names parallels that found in gene nomenclature: a manageable set of initial core objects gives way to waves of thematic naming, until the avalanche of new genes brought on by large-scale sequencing forces us to bland, systematic identifiers.

What's in a (gene) name?

Many gene names are straightforward: ordered sets of letters and numbers conforming to a specific pattern. Some carry no meaning beyond pure record-keeping, reflecting the need to quickly assign a unique identifier to every genomic entity. Other systems confer information in a structured manner. Consider the *Saccharomyces cerevisiae* gene name YAL042W as used by the *Saccharomyces* Genome Database. Each part has a specific meaning: Y denotes the species (yeast), A indicates the chromosome (I), L denotes the chromosome arm, W is the coding strand (Watson) and 042 is a sequential identifier.

Early gene names were often generated in a loose 'namespace': several letters, sometimes followed by numbers. Such names are often abbreviations for scientific terms describing initial findings about the gene; in some cases, these have dual meaning - for instance, *LOV1* refers not only to the noblest human emotion but also to light-, oxygen- and voltage-sensitive domains. In the fruit fly *Drosophila melanogaster*, gene names are frequently full words or phrases, drawn from a variety of languages.

Classification overview

For this survey, we defined gene names of interest as those with extraneous or unrelated ('skewed') meaning. Following a survey of biological databases and with the help of several websites dedicated to 'interesting' gene names [4,5], we gathered over 100 notable names from several species. Whereas these websites have attempted mainly to catalog the names according to their source (such as history and literature), we explored the underlying patterns.

We established four main classes (T, P, M and ~M) and 11 subclasses. This classification is shown in Figure 1, and an expanded version is available in Additional data file 1 and online at [6]. These categories, admittedly arbitrary, reflect several observable implementations of nonstandard biological naming.

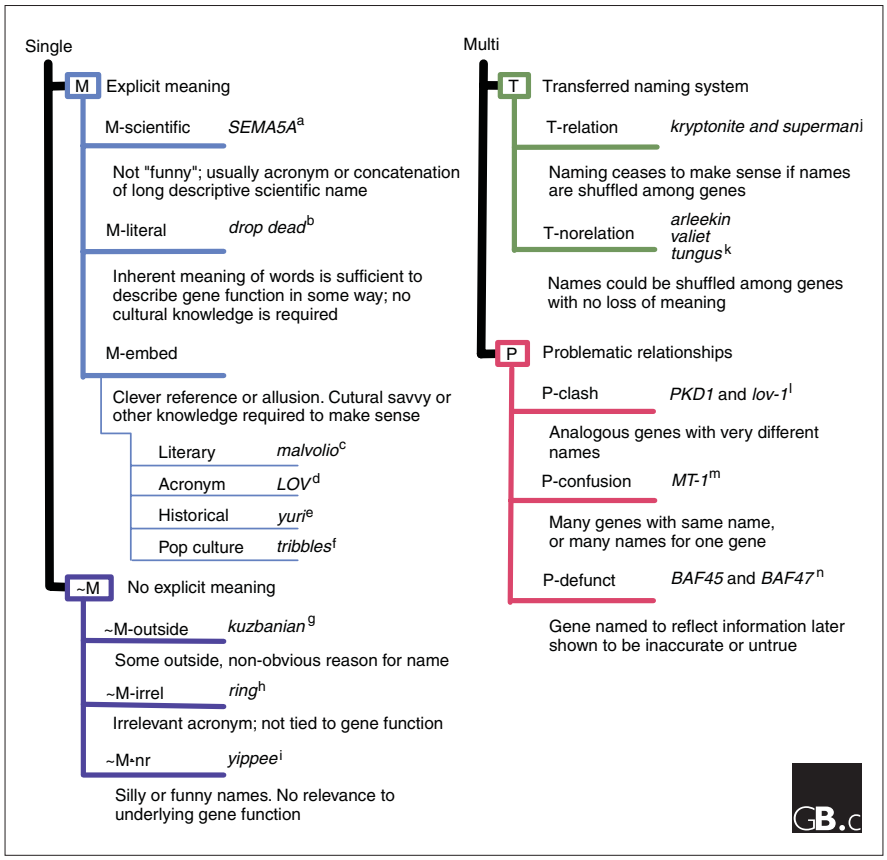

Figure 1

Criteria and examples of gene name classification. [a]*SEMA5A* (human): sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A; [b]*drop dead* (*Drosophila*): flies with mutations in *drop dead* die rapidly after their brain rapidly deteriorates. [c]*malvolio* (*Drosophila*): gene needed for normal taste behaviour. Malvolio in Shakespeare's *Twelfth Night* tasted "with distempered appetite". [d]*LOV* (many different organisms): light, oxygen, or voltage (*LOV*) family of blue-light photoreceptor domains. [e]*yuri* (*Drosophila*): this gene was discovered on the anniversary of Yuri Gagarin's space flight. Mutants have problems with gravitaxis and cannot stay aloft. [f]*tribbles* (*Drosophila*): cells divide uncontrollably, like the eponymous *Star Trek* characters. [g]*kuzbanian* (*Drosophila*): mutants have uncontrollable bristle growth. Koozbanians are alien Muppets with uncontrollable hair growth; spelling was changed to avoid copyright infringement. [h]*ring* (*Drosophila*): really interesting new gene. [i]*yippee* (*Drosophila*): a graduate student's reaction on cloning the gene. [j]*kryptonite* and *superman* (*Arabidopsis*): the *kryptonite* mutation suppresses the function of the *SUPERMAN* gene. [k]*arleekin*, *valient*, *tungus* (*Drosophila*): mutations in *arleekin*, *valient*, *tungus* and 29 other genes affect long-term memory. Named after Pavlov's dogs. [l]*PKD1* (human) and *lov-1* (worm): these are homologs, although their names do not suggest it. [m]*MT-1* (human): this label can refer to at least 11 different human genes. [n]*BAF45* and *BAF47* (mouse): names for the same gene, reflecting a revision of the molecular weight of product.

Explicit meaning (M)

The first class (M) contains individual genes whose names have meaning; that is, they reflect in some intelligible way an underlying characteristic of the gene. This is accomplished in three ways, reflected in three subclasses.

Scientific meaning (M-scientific) covers genes with standard, descriptive scientific names, sometimes shortened to yield quaint abbreviations. To a scientist, these names are the most descriptive, conveying meaningful information about the gene (for example, *SEMA5A*; see Figure 1 legend for a description of this and other gene names). Description can also be achieved through literal meaning (M-literal). Such labels as *drop dead, brokenheart* and *stuck* refer

to effects noticed in mutants. Although descriptive, these names are often too vague to be instructive on their own. Last, a large subclass of gene names covers those with embedded meaning (M-embed). These are similar to M-literal in that they aim to be descriptive and memorable, but here the names are drawn from pop culture, history or literature. These names are opaque unless the audience grasps their cultural meaning (for example, *tribbles* refers to *Star Trek* creatures that reproduce uncontrollably).

No explicit meaning (~M)

The second class of genes (~M) constitutes individual genes named in such a way as to convey virtually no information about the gene itself. We categorized names with no apparent reason (~M-nr) as those whose significance, if any, is wholly irrelevant to the underlying function of the gene (for example, *yippee*). A similar subclass is the irrelevant acronym (~M-irrel), in which the official gene name reflects an abbreviation of equally random terms (for example, *ring*). Other names reflect not only the inclinations of the researchers who coined them, but also outside pressures (~M-outside). For instance, the Muppets reference *kuzbanian* was purposely misspelled to avoid copyright infringement, and the *Drosophila* gene *fruity* - defects in which cause males to lose interest in females - was later renamed *fruitless*, revealing an intrusion of political correctness into gene naming.

Transferred naming system (T)

The third class of genes (T) contains those for which entire naming systems have been transferred from other domains. Such transfers can occur in two forms. In the first (T-relation), internal relationships among names are preserved. Thus, meaning is conveyed not only by the system as a whole, but also by the assignment of individual names within the transposed system (for example, *superman* and *kryptonite*). Early astronomical naming was similarly nonrandom: the names of Mars, Jupiter, Venus and Mercury, among others, clearly relate notable features of various gods to characteristics of the planets as seen from Earth. In the second subclass (T-norelation), any internal relationships that existed among the names are lost when applied to genes; new meaning is tied only to the transposed system as a whole (for example, the names of Pavlov's dogs). This is common in computer networks; we have encountered arrays of printers named after characters from *The Simpsons*, *Star Trek* and *Lord of the Rings*.

Problematic relationships (P)

The fourth and final class (P) constitutes multi-gene naming problems. Such gene names become troublesome when situated in the wider landscape of biological nomenclature. When multiple names clash (P-clash) - for instance, orthologs with the same function exist in different species but their names are completely different - the result is divergent names for similar genes (consider *PKD1* and *lov1*). Conversely, confusion (P-confusion) frequently results when different genes share the same name, or many names are attached to one gene (for example, *MT-1* refers to at least 11 genes in humans, while *asp* refers to at least 14 genes across eight species, many with entirely disparate functions). Finally, defunct names (P-defunct) occur when the function or characteristics implied by the name prove incorrect or misleading (for example, *Baf45* and *Baf47* were both names for the same mouse gene, now called *Ini1*, which reflected conflicting estimates of the molecular weight of the product).

Naming on a genomic scale

The prevalence of unusual gene names can be very different among species. While *Drosophila* brims with creative names, many recently sequenced organisms use strict numbering systems and other species impose limitations on length and format.

We assessed naming profiles on a genomic scale by gathering all gene names in a given species and plotting the occurrence of each name, both on the web (cultural impact) and in PubMed (scientific impact) (see schematic, Figure 2a). The left-hand region represents well-studied genes whose names are distinctive and not found in common parlance (that is, 'normal' gene names); the right-hand region represents those with many hits in the web search (for example, genes with common English names or abbreviations with additional meaning).

By comparing such graphs for several species, we can discern which species contain a high fraction of gene names with 'skewed' meaning. We compared four species, the baker's yeast *S. cerevisiae* (classic gene naming convention: three letters plus one number), the bacterium *Escherichia coli* (four letters plus one number) and the fruit fly *D. melanogaster* (no limits on gene names). We also examined the first free-living organism to have its full genome sequenced, the bacterium *Haemophilus influenzae*; gene names in this species conform to strict standards and are best described as identifiers.

A strict identifier-only model of naming yields names unlikely to have meaning in any other context. Restrictive namespaces must generate names using combinations of a small set of letters and numbers, and the resulting names are likely to be jumbles of characters with no secondary meaning. As the namespace grows and scientists have greater freedom to choose names as they please, there is a higher prevalence of names with dual meaning. On a species level, we therefore expect gene names in *H. influenzae* to return virtually no Web hits, and those in *Drosophila* to return many. By juxtaposing scatterplots we can gauge the prevalence of names with distorted meaning in a species as a whole (Figure 2b).

Given this wide range of naming conventions, the main question is how to move forward. Gene nomenclature might move to a two-tiered model, like the famous Linnaean binomial naming

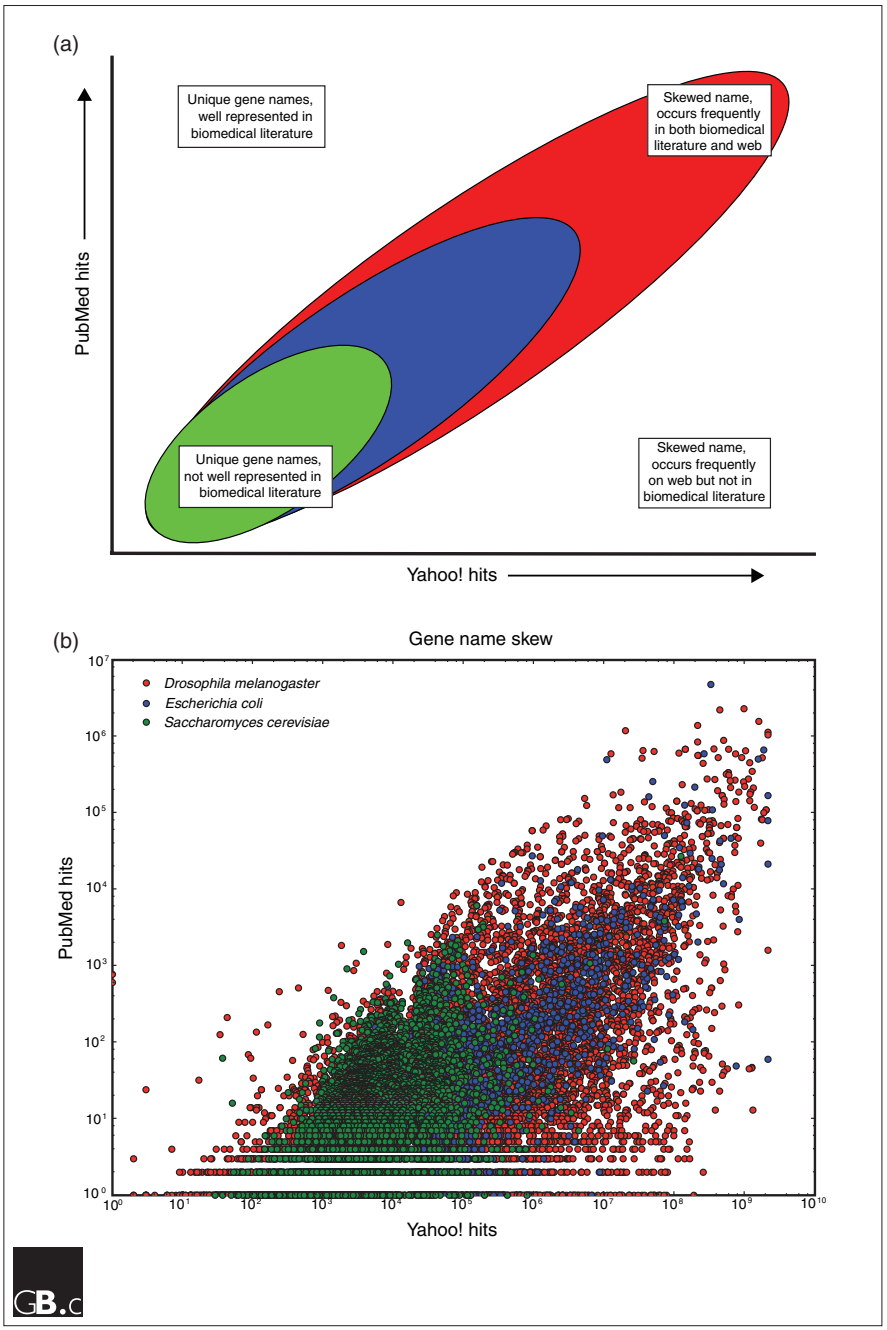

Figure 2
Comparisons of gene names. **(a)** Schematic of 'skewed' gene names inter-species comparison. Horizontal axis, Web search results for gene name using Yahoo! search engine. Vertical axis, search results using gene name as PubMed query. Overlapping ovals, predicted name distribution for *S. cerevisiae* (green), *E. coli* (blue) and *D. melanogaster* (red) based on naming systems employed in these species. **(b)** 'Skewed' gene names inter-species comparison. Actual name distribution for *S. cerevisiae* (green), *E. coli* (blue) and *D. melanogaster* (red). *H. influenzae* is not shown; the strict, identifier-style names in this species generated virtually no Web hits, so these names appeared entirely along the base of the horizontal axis and were omitted.

scheme. This approach could pair existing names with descriptive prefixes or suffixes conveying meaningful information, such as Dr/Mrs/Sir and suffixes like degree titles or top-level Internet domains such as .com, .edu and .gov. Adding this systematized binomial structure is just one possible way to update gene nomenclature without completely uprooting its existing structure. And perhaps it is worth saving. Biological nomenclature is undeniably idiosyncratic and perhaps dysfunctional, but even the silliest gene names are meaningful in a sense - from cultural influences to wordplay, allegory, and clever puns, gene names reflect our essential humanity, the minds behind the science. The work of those early pioneers remains enshrined in the whimsical gene names dotting the species they studied.

Additional data files

Additional data are available with this paper online. Additional data file 1 contains the full list of genes from which the examples presented were chosen.

References

1. Vacek M: **A gene by any other name: whimsy and inspiration in the naming of genes.** *Am Sci* November/December 2001 [http://www.americanscientist.org/template/AssetDetail/assetid/14672]
2. Schwartz, J: **'Sonic Hedgehog' sounded funny, at first.** *The New York Times*, November 12, 2006.
3. Petsko, G: **What's in a name?** *Genome Biol* 2002, **3:**comment1005.1-1005.2
4. **Clever gene names** [http://tinman.nikunnakki.info]
5. **My favorite gene names** [http://jpetrie.myweb.uga.edu/genes.html]
6. **Funny gene** [http://www.gersteinlab.org/proj/funnygene]