© 2015 by Renée van Amerongen

How to…. Gene Nomenclature

Did you know that official rules (or at least some form of consensus) exist(s) for how scientists should refer to genes and proteins in different species? You should familiarize yourself with these rules, as they will help you understand the existing literature (provided that other people adhere to these rules, of course). More importantly, these rules should serve as a guideline when writing your own essays/reports/papers.

Background

Since 1979 the HUGO human gene nomenclature committee has been responsible for *"approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication"* (http://www.genenames.org). However, a lot of research is not done on humans and many genes were first identified in studies using model organisms (mice, flies, worms, yeast, etc.). Until the human genome sequence became available at the turn of the 21st century, scientists have always had great freedom to come up with exotic names and assign them to their favorite genes. You can still see this in many of the gene names that were initially assigned based on mutant phenotypes in *Drosophila*: these genes carry names like "*wingless*" or "*frizzled*".
As science progressed, it became clear that many genes were conserved across species. The mouse homologue of "*wingless*" for instance, turned out to have been identified in the early 1980s as "*int-1*".
All of this changed with the advent of high throughput whole genome sequencing in combination with better gene prediction algorithms: It used to be the case that you would identify a gene and as you were trying to figure out what it's function was, you could name it (*wingless* or *int-1*, for instance). Nowadays, even though we have only studied the function of a small proportion of the total number of genes, they have been assigned standardized (and more boring) names in the genome databases[1]. So how does that work?

Quick nomenclature guidelines

Each field still has their own rules. For instance, for more information on fly gene nomenclature you can go to http://www.flybase.org. and for C. elegans you can visit http://www.wormbase.org. However, the community of vertebrate model organisms has adopted two important general guidelines:

1) Wherever possible, the genes are given the same gene symbol as their human counterpart (and therefore as decided on by the HGNC). This means that genes that the Wnt-signaling community still commonly refers to as "Wnt" or "Frizzled" carry the official gene symbols "WNT" and "FZD", with numbers added to discriminate between the different gene homologues.
2) To discriminate between genes and the encoded proteins, there is a simple rule: genes are *italicized* whereas proteins are not. So if you read a piece of

[1] *As an example: In 2006 I identified a novel binding partner for one of the proteins I was studying at the time. I cloned the full coding sequence of the gene encoding this novel protein and named it "Sorority". But when I blasted the coding sequence, it turned out that the gene already had a name: "Tmem98". To this day we don't fully understand the function of this gene and nobody else has really studied it since. But at the time of its "discovery" it already had a name, because computer algorithms had predicted its existence (and classified it as a transmembrane protein – number 98 to be exact).*

© 2015 by Renée van Amerongen

text where the authors switch between *FZD1* and FZD1, you can deduce that they are talking about the gene and the protein, respectively.

3) Different species get a different combination of capitals and regular letters.

What does it mean in practice?

Lets take *FZD1* as an example. If you go to the HUGO website (http://www.genenames.org) you can find tons of information on your gene of interest:

APPROVED SYMBOL	FZD1
APPROVED NAME	frizzled class receptor 1
HGNC ID	HGNC:4038
PREVIOUS SYMBOLS & NAMES	"frizzled (Drosophila) homolog 1", "frizzled 1, seven transmembrane spanning receptor", "frizzled family receptor 1", "frizzled homolog 1 (Drosophila)"
SYNONYMS	DKFZp564G072, "frizzled, Drosophila, homolog of, 1", "Wnt receptor"
LOCUS TYPE	gene with protein product
CHROMOSOMAL LOCATION	7q21
GENE FAMILY	GPCR / Class F : Frizzled receptors
HCOP	Orthology Predictions for FZD1

When you write an essay/paper you might want to refer to the gene by its official name the first time you mention it (the same way in which you would spell out any other abbreviation the first time). After that, you can use the gene symbol.

If you refer to the same gene/protein in different species, you should stick to the following guidelines:

	Gene	Protein
Humans	*FZD1 (all caps)*	FZD1
Mice/Rats	*Fzd1 (one capital)*	FZD1
Chicken	*FZD1*	FZD1
Xenopus/Zebrafish	fzd1	Fzd1

Please note how you can immediately tell the difference between a mouse and human gene in a piece of text. This comes in handy when you are switching between tumor studies performed in humans and studies performed in mouse models, for instance – a situation that I often encounter myself.

Disclaimer

When reading the literature, you will also quickly find that old habits are hard to change. For instance, when describing the components of the Wnt-pathway, you will rarely see the beta-catenin (or β-catenin) protein being referred to as CTNNB1. Most scientists stick to the habitual name in this case and no one in the field will be confused.

© 2015 by Renée van Amerongen

APPROVED SYMBOL	CTNNB1
APPROVED NAME	catenin (cadherin-associated protein), beta 1, 88kDa
HGNC ID	HGNC:2514
PREVIOUS SYMBOLS & NAMES	"catenin (cadherin-associated protein), beta 1 (88kD)", CTNNB
SYNONYMS	armadillo, beta-catenin
LOCUS TYPE	gene with protein product
CHROMOSOMAL LOCATION	3p21
GENE FAMILY	Armadillo repeat containing
HCOP	Orthology Predictions for CTNNB1

But now at least you know how to do it properly!