

# Genomics Confounds Gene Classification

*Large-scale genomic studies challenge traditional definitions of genes and require new approaches to classifying life at the molecular level.*

Michael Seringhaus and Mark Gerstein

Scientists strive to make sense of the natural world by defining its vital parts. As physicists anointed the atom, molecular biologists selected the gene as their basic unit. It was a smart choice: virtually every observable property of any organism on Earth is derived from the action of one or more genes. Early on they were conceived as the physical embodiment of Gregor Mendel's basic rations of heredity. By the mid-20th century, molecular science sharpened the picture. Genes became distinct spans of nucleotide sequence, each producing a RNA transcript translated into a protein with a tangible biological function.

Today, high-throughput genomics is generating data on thousands of gene products every month, improving our view once more. Biology's basic unit, it is now clear, is not nearly so uniform nor as discrete as once was thought. Scientists must adapt to these complexities and improve the systems we use to

name and classify genes and their vital products. As Confucius once warned: defective language produces flawed meaning. But what is the best route toward improved precision? To try to answer that, we must understand how we reached where we stand today.

## An Idea Blossoms

The word "gene" originally arose as a derivative of *pangene*, a term used to describe entities involved in pangenesis, Darwin's hypothetical mechanism of heredity. The etymology of that term derives from the Greek genesis ("birth") or genos ("origin"). The term gene itself was first used by Wilhelm Johannsen in 1909, based on a concept Mendel had developed in 1866. In his famous breeding experiments with pea plants, Mendel showed that certain traits (such as height or flower color) do not appear blended in offspring. Instead, these traits are passed on as distinct, discrete entities. Furthermore, he demonstrated that variations in such traits are caused by variations in heritable factors. (In modern terminology, he showed that genotype dictates phenotype.) In the 1920s, Thomas Hunt Morgan demonstrated that genetic linkage, the tendency of certain traits to appear together, corresponds to their physical proximity on chromosomes. The one-gene, one-protein view soon followed, as George Beadle and Edward Tatum demonstrated that mutations in genes could cause defects in specific steps of metabolic pathways. A series of experiments then established that DNA is the molecular vehicle for heredity, culminating in James Watson and Francis Crick's famous 1953 solution of the three-dimensional structure of DNA.

The double-stranded double helix, with its complementary base-pairing, neatly explained how genetic material is copied in successive generations and how mutations can be introduced into daughter chromosomes by occasional replication errors. Crick's continued work decrypting the genetic code laid the groundwork for the so-called central dogma of molecular biology: namely, that information travels from DNA through RNA to protein. In this scheme, a gene is a DNA region (or "locus") that is expressed as messenger RNA (mRNA) and then translated into a polypeptide (usually a protein needed to build or operate a portion of a cell). This version of the general blueprint of life, with exceptions such as the RNA-based genomes in some viruses, is the overarching view that brought scientists to the doorstep of the genomic era.

But this view has ramifications far beyond the nucleotide-sequence level. The central dogma also seeded what we'll call the "extended dogma" of molecular biology. Within this conceptual framework, a transcribed mRNA (corresponding to a gene) gives rise to a single polypeptide chain that in turn folds to form a functional protein. This molecule is thought to perform a discrete and discernible cellular function such as catalyzing a specific chemical reaction. The gene itself is regulated by a promoter and transcription-factor binding sites assumed to be located on nearby DNA.

Genetic nomenclature developed to reflect the view that every gene has a discrete function. Each gene was given a name, and these names and their as-

---

*Michael Seringhaus received his PhD in genomics and bioinformatics from the Department of Molecular Biophysics and Biochemistry at Yale University, where he studied with Mark Gerstein. He is currently a student at Yale Law School. Mark Gerstein is the A.L. Williams Professor of Biomedical Informatics at Yale University, where he co-directs the Yale Computational Biology and Bioinformatics Program. His laboratory uses computation to annotate genome sequences, mine data on gene expression and molecular networks, analyze protein families and simulate macromolecular structures. A former W.M. Keck Foundation Distinguished Young Scholar, he received his PhD at Cambridge University. Gerstein address: PO Box 208114 MBB, Yale University New Haven, CT 06520. Internet: Mark.Gerstein@yale.edu, <http://gersteinlab.org>*





Hillel Steinberg / Alamy

**Figure 1.** Scientists, especially biologists, are avid classifiers. Few are as famous for their intensive tallying and categorizing ways. Like eager stamp collectors, they sort related but disparate objects by age, place of origin, shape or other significant traits. How else could one try to make sense of life? Consider where a careful inventory of Galapagos finches carried Charles Darwin or a close analysis of fruit flies brought Thomas Hunt Morgan. High-throughput genomic experiments — sequencing studies, transcriptional research and the rest — are generating huge waves of new data that will require smart sorting on a massive scale. To fully take advantage, it's time for biologists to consider creating new methods for doing what they've always done so well.

sociated functions were arranged in a simple classification system. Such classification begins with broad functional categories (for instance, genes whose products catalyze a hydrolysis reaction or bind to other molecules) and moves to more specific functions (for example, the designation “amylase” describing the specific hydrolysis reaction involved in breaking down starch.) Early attempts at functional classification of this sort, starting in the 1950s, include the International Commission on Enzymes Classification and the Munich Information Center for Protein Sequences. This unitary approach toward function still influences many well-known protein databases such as UniProt (the Universal Protein Resource) where genes are arranged by name and research articles are indexed to these

names. To accomplish this, curators peruse manuscripts and synthesize from them a simple summary statement of each gene's function as described in the literature. This functional annotation is used to situate a given gene within the larger functional landscape.

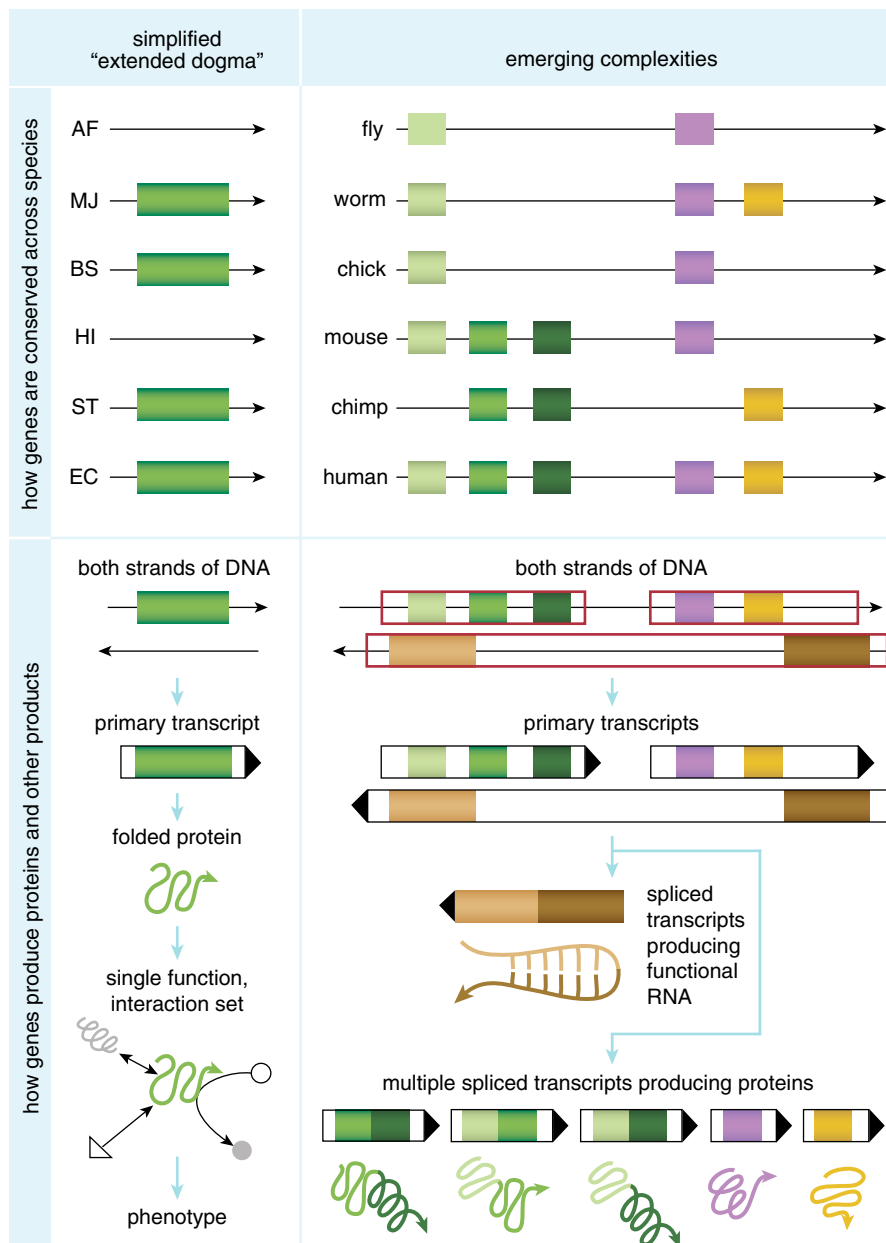
This iterative one-gene, one-protein, one-function relationship paints a relatively straightforward picture of subcellular life. When describing the function of a given gene in a cell, an individual protein can be imagined as a single indivisible unit or node within the larger cellular network. In turn, when mapping genes across species using sequence similarity, a protein is assumed to be either fully preserved in various organisms or entirely absent. Thus, related proteins in different organisms can easily be grouped together into

consistent families, which can be given simple, unitary descriptions of their function. Thus, the extended dogma expands the central dogma to include regulation and function. (See Figure 2)

### Complex Reality

To the modern genomics scientist, the classical image of a gene and the extended dogma associated with it are quaint. High-throughput experiments that simultaneously probe the activity of millions of bases in the genome deliver a far less tidy view. First, the process of creating a RNA transcript from a DNA region is more complex than once was imagined. Genes make up only a small fraction of the human genome. But RNA expression studies on human DNA suggest that a substantial amount of the genome outside the boundaries of known or pre-





**Figure 2.** At left is the traditional picture of a gene, somewhat simplified and extended. A region of DNA gives rise to a primary transcript and to a folding protein with a single biochemical function. The protein participates in a single set of interactions with other molecules, which give rise to a single discernible phenotype. At right is the more complicated, contemporary view. DNA segments coding for functional products are split up into different regions, called exons, on a chromosome. Transcribed regions form long primary transcripts, which are spliced to give shorter transcripts that give rise to functioning products. Splicing can occur in multiple ways, mixing and matching different regions of a genome. Transcripts may produce folding functional RNAs as well as folding proteins. In the simplified view on the left, a polypeptide associated with a gene is conserved entirely across species, shown here in representative prokaryotes (EC, ST, etc.). In contrast, in higher eukaryotes at right, bits and pieces of the gene — often corresponding to particular exons — may or may not be conserved across organisms. Sometimes conservation patterns do not even conform to exonic boundaries.

dicted genes is transcribed. Among the evidence are results published last year from the pilot phase of the Encyclopedia of DNA Elements (ENCODE) project. This massive, international collaboration intends to functionally annotate each base pair of the human genome.

The pilot studies on a representative 1 percent of genome (roughly 30 million base pairs) suggested that non-genic transcription is very widespread. Precisely how wide is not yet known.

Moreover, the function of this non-gene, transcribed material is unclear.

So is how best to classify and name it. Since genetic nomenclature is keyed to link to discrete genes, short transcribed regions located outside of identified genes are troublesome. They sometimes end up listed in sequence data-banks sporting similar identifiers to genes, which can be confusing. To further complicate things, experiments on non-gene transcription show that some of this activity occurs in pseudogenes, regions of genome long considered fossils of past genes. In a transcriptional sense, dead genes appear to come to life, with some clues even suggesting they may help regulate other genes.

Finally, the phenomenon of alternative splicing complicates matters further. In eukaryotes, genes typically are composed of short exons, or coding regions of DNA, that are separated by long DNA stretches called introns. Scientists have long understood that introns are transcribed to mRNA that is discarded (or “spliced out”) before proteins are produced. However, it now appears that for a given gene-containing locus this splicing can be done in multiple ways. For instance, individual exons can be left out of the final product. Sometimes, only portions of sequence in exons are preserved. (See Figure 2) When a sequence from outside the conventional bounds of a gene is spliced in as well, the number of variants climbs further. What once was thought to be a system to reliably remove introns can itself yield many variants of a single gene. This variation too appears to be considerably more prevalent than once was thought.

Our understanding of gene regulation is also changing. The traditional view of the gene assumed that the protein coding portion of a gene and its regulatory sequences existed in tight proximity on a chromosome — in some definitions they were part of the gene. In particular, the classical picture of gene regulation has long been taught via the *lac* operon, a simple bacterial example of repressors, operators and promoters clustered near each other. This model describes a direct, proximal relationship between transcription factors and genes, with regulatory sequences of a particular gene directly upstream. But this simple rubric does not apply very well to mammalian systems and other higher eukaryotes. In that setting, genes can be regulated very far upstream by enhancers over 50,000 base pairs away, even beyond adjacent genes. The loop-

ing and folding of DNA can bring distant spans into close spatial proximity. (See Figure 3) Moreover, gene activity can be influenced by chemical alterations called epigenetic modifications. These can come in the form of modifications to the DNA itself (such as the attachment of methyl groups) or modifications to histones, support structures in chromosomal DNA. Depending upon such modifications, a gene may be functionally active or silent in different circumstances with no change to its sequence. This further complicates the notion that a DNA sequence in a single region is sufficient to describe a gene.

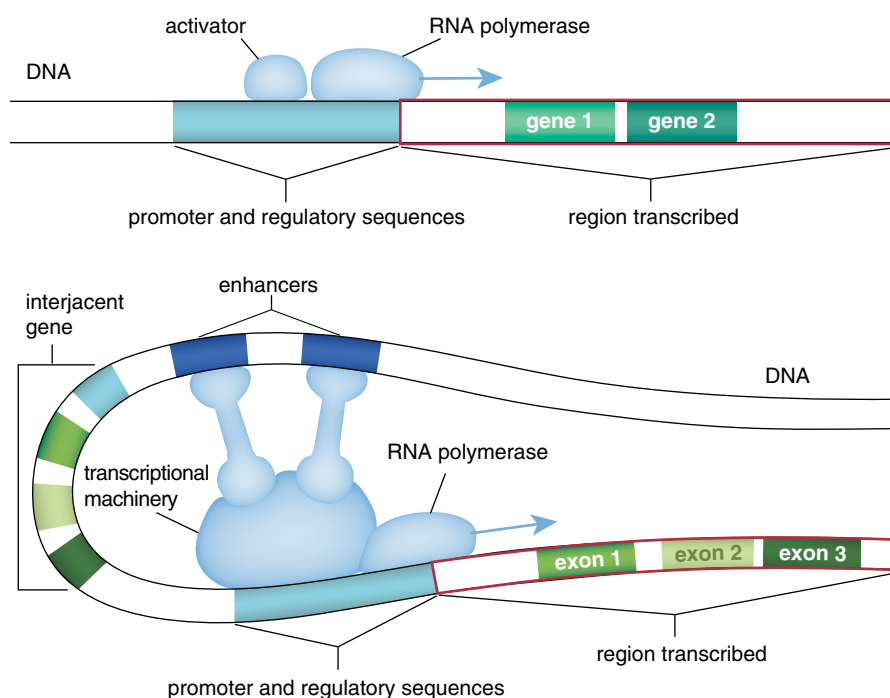
The transcriptional and regulatory peculiarities described above never meshed well with the traditional notion of the gene, but they were thought to be fairly rare. Again, the recent ENCODE results suggest that deviations from the traditional model could be the norm.

### Capturing Function

In the quest to accurately describe biological systems, defining basic units is only part of the job. Scientists ultimately want to understand biological function. Function in the genetic sense initially was inferred from the phenotypic effects of genes. A person might have green or blue eyes and a gene related to this characteristic could then be ascribed the “eye color” function. Phenotypic function of this sort is most directly seen through deleting or disrupting, or “knocking out”, a particular gene. Disrupting a gene in this way might cause an organism to develop cancer, to change color or to die early. Disabling the yeast mitochondrial gene *FZO1*, for instance, causes mutant strains to display slow growth and a petite phenotype.

But a phenotypic effect doesn’t capture function on the molecular level. To really elucidate the importance of a gene, it’s vital to understand the detailed biochemistry of its products. For instance, the yeast gene *FZO1* mentioned above displays GTPase enzyme activity, a molecular-level action not immediately apparent from its ultimate phenotypic effect. *FZO1* protein, it’s now clear, helps fuse mitochondrial membranes in yeast, protecting the cellular power plants involved in energy production. The biochemical effect explains the phenotypic effect.

Also key to understanding function are the processes or pathways a gene product engages with in a given cell.



**Figure 3.** In bacteria, the genetic regulatory machinery involves repressors or activators operating in close proximity to genes. The gene is actually transcribed by RNA polymerase. In eukaryotes polymerase still transcribes gene but important control regions can occur tens of thousands of nucleotide pairs away from the targeted coding region – with uninvolved genes sometimes positioned in between. (In the figure this is indicated by the “interjacent” gene.) The physical qualities of DNA, its ability to loop and bend, brings distant regulatory components into close proximity.

A gene, for instance, may be involved in secretion or amino-acid biosynthesis and thus could be classified functionally in this manner. Identifying where a protein is found within various cell compartments offers additional functional insight. A protein may be found only in the nucleus or in a cell membrane. *FZO1*, as would be expected, localizes to the mitochondrial membrane in yeast.

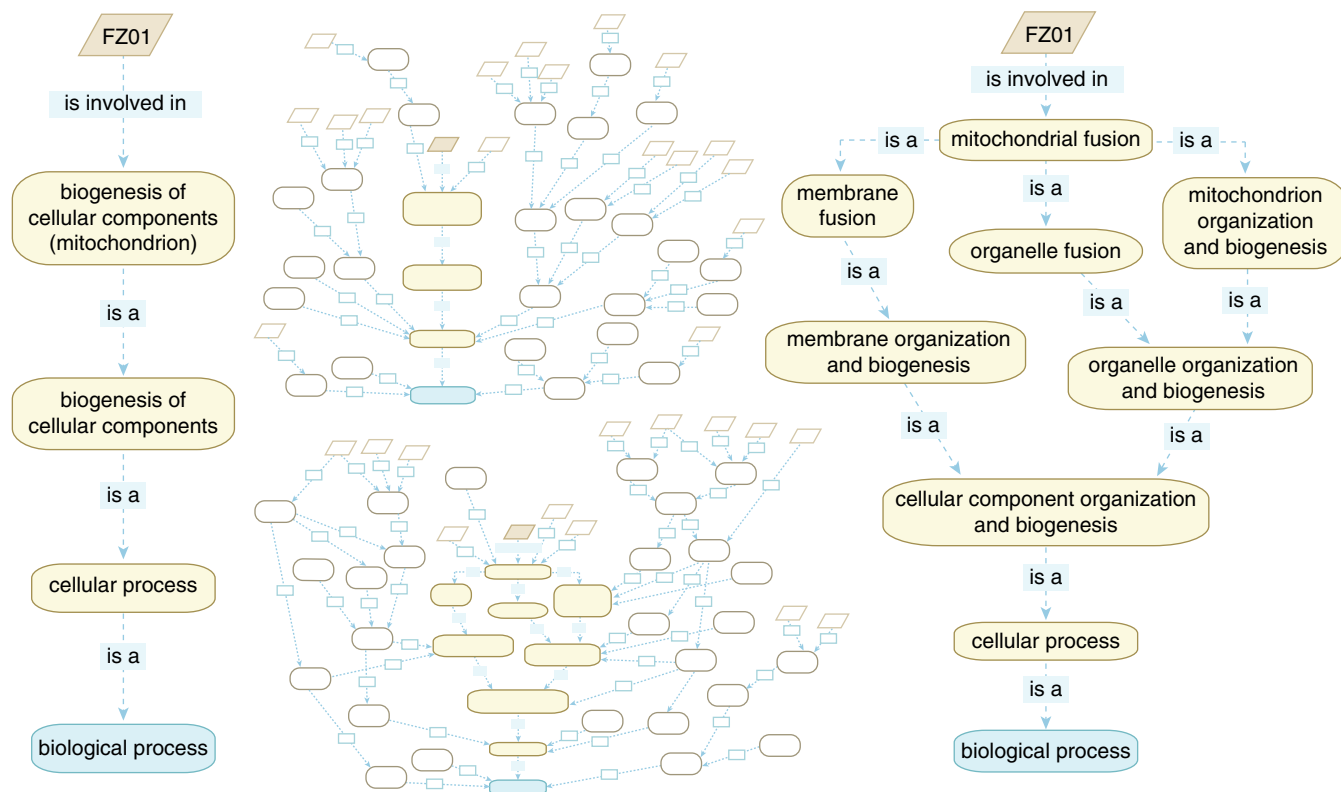
Deciding which qualities of a gene and its products to record, report and classify is not trivial. All of the aforementioned approaches to functional classification assume a simple hierarchical classification scheme (for example, gene X is a member of group Y, which is part of superclass Z). Because of the complexity of functional classification, scientists now are pursuing several additions and enhancements to hierarchies to integrate information from multiple levels of function.

One popular approach that arose with widespread genome sequencing and the avalanche of data it produces is the Gene Ontology, best known as GO. This system is more complicated than a simple hierarchy since it employs a directed acyclic graph (DAG) structure. (See Figure 4) In a simple hierarchy a gene can have only one functional

“parent” or classification whereas in the DAG approach, it can have any number of parents. For instance, a gene product might belong to a subset of proteins involved in cell-cycle control while at the same time belonging to a group of transcription factors. Individual entities thus are more fully described.

Both DAGs and simple hierarchies classify from the general to the specific, but because a DAG can have multiple parents for any given node, the latter is more flexible. The DAG approach does have some shortcomings. Expansion of the classification depends on the degree of knowledge about a particular cellular process. All aspects of subcellular life are not equally well studied, a fact that can reflect the interests of researchers or funding agencies more than actual biological complexity. So some areas of a DAG can be much richer structurally than others for the wrong reasons.

Another approach to making the most of floods of new genomic data, particularly from large-scale experiments that sample many genes at once, is to assign uniform attributes to each gene. For instance, biologists can measure the expression level of the same gene in variety of cellular conditions



**Figure 4.** In a simple hierarchy, a gene is described in terms of single relationships where one unit always descends from just one “parent”. Directed acyclic graphs (DAGs) can capture more complexity. In this example, the hierarchy captures that FZO1 plays a role in the biogenesis of cellular parts but the DAG gives a wider view of the scope of those roles.

using DNA microarrays. Or they can compare how stringently its protein products bind to a battery of metabolites with a protein array.

Finally, scientists can attempt to describe gene function completely in terms

of molecular networks. This approach focuses less on what a particular gene does and more on which other genes it is connected to, in much the same way social network research does with people. As the old saying goes, what you do

might matter less than who you know.

Further adding to this complexity, a single gene frequently does not yield a protein with a single function, despite the one-to-one-to-one implications of the extended dogma. Individual proteins are often comprised of domains, each a different segment of polypeptide sequence with a distinct folded structure that might serve a discrete cellular need. A protein that catalyzes a certain reaction through one domain may have an additional domain responsible for DNA binding. Conversely, an assemblage of proteins produced by two or three genes can be necessary to carry out a single identifiable function in a cell.

Moreover, any given domain might belong to a distinct protein family occurring in many different species. But only certain domains, or even subdomains, may be preserved across species. Trying to describe such “sub-genic” conservation with traditional gene names or database identifier tags can quickly become confusing. The names don’t always clarify whether a whole protein is conserved, or just some part of it. (See Figure 2)

Given this, some people have argued that protein domains offer a more

#### quirky gene names

literature	pop culture	cars and driving
amontillado	sonic hedgehog	Sunday driver
luciferase	tribbles	limo
thor	pokemon	18wheeler
tigger	kojak	long island expressway
ariadne	groucho	people
malvolio	maggie	
tinman	glass-bottom boat	
food	swiss cheese	
	moonshine	

**Figure 5.** The story behind gene names gets complicated. For instance, consider the gene temporarily called evander, which was detected in zebra fish. Fish with mutant forms of the gene had deformed ears and jaws. Early on, researchers named it after Evander Holyfield, the boxer whose ear Mike Tyson bit in a match. Due to reservations by a collaborating researcher, the mutated form of the gene was changed to hearsay before its discovery was published. For more information on gene names visit: <http://tinman.nikunnakki.info/?q=browser>.

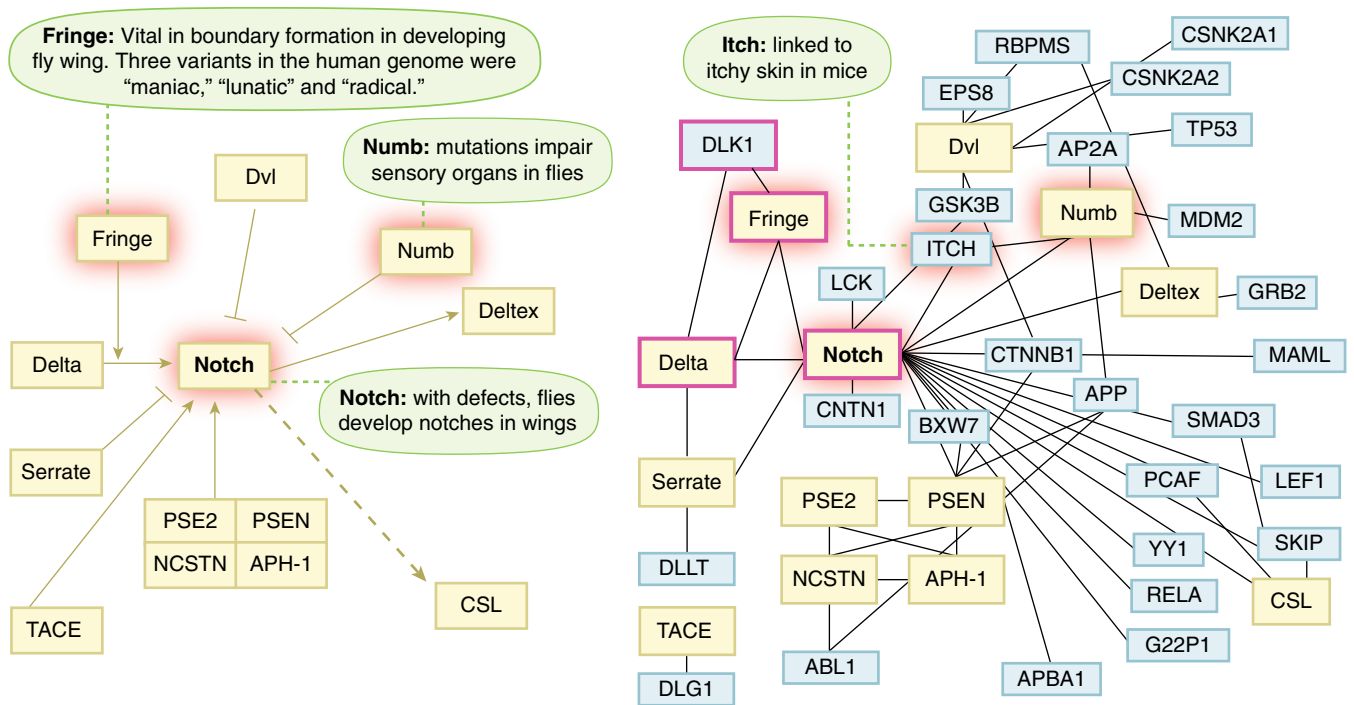


Figure 6. The Notch pathway is highly conserved among many species. Notch encodes a receptor protein first identified in fruit flies that produces a notch-like wing shape when the gene is defective. The protein spans the cell membrane and acts like a trigger. When certain signaling molecules bind to its extracellular domain, the intracellular domain detaches and influences gene expression. At left, is a traditional view of part of the Notch pathway (adapted from the KEGG PATHWAY Database) with Notch and its interaction partners depicted in yellow. At right, results from high throughput, interaction experiments in humans identified many more proteins involved with the pathway (many of which are shown in blue). Several stylized gene names (Numb, Itch, etc.) are present here while others (PSE2, BXW7) are codes, an example of clashing nomenclature in contemporary genomics. At left, Fringe is denoted by a traditional “funny” name related to one of its roles. At right, Fringe is portrayed at the biochemical level as interacting with three partners: DVL, Delta, and Notch. Likewise, the Notch gene can also be described variously: as a single-pass, transmembrane receptor protein, as a gene whose deletion leads to a phenotypic effect, as a member of a particular signaling pathway or as related to interaction partners pictured here.

workable basic unit for molecular biology than entire genes. Focusing on domains does facilitate certain functional descriptions, but this approach is far from perfect. For instance, two mRNA splice variants may share the same exon and presumably the same domain structure – but they may still produce different products, depending on the presence or absence of a short, several-nucleotide leader sequence.

### What’s in a Name?

The lack of reliable central nomenclature standards is becoming a more urgent concern in biology. Prior to the genomics age, gene nomenclature was a relatively small-scale endeavor that produced, on the whole, carefully chosen and sometimes whimsical, even sassy, names. Without gobs of raw sequence and super-powerful computers, homology mapping between species was modest. Research communities working on a given model organism, say a fruit fly, a yeast or a cress weed, developed informal naming standards among themselves.

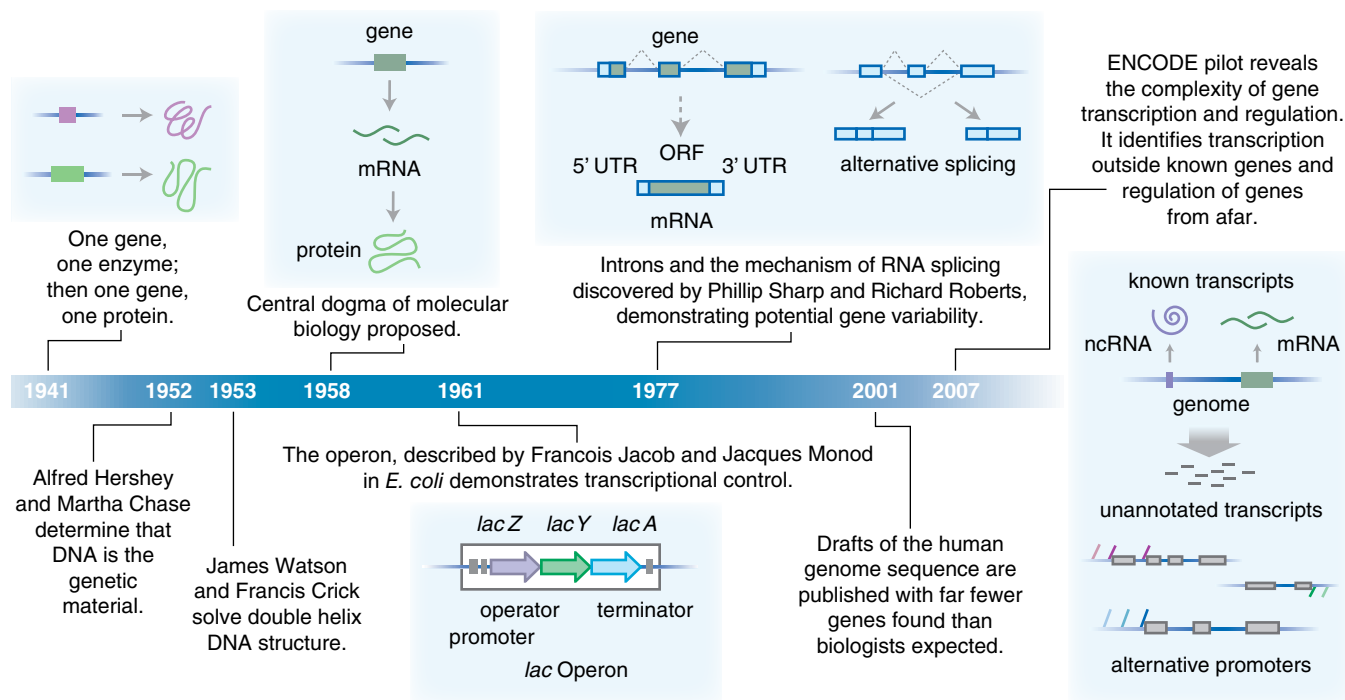
Scientists lucky enough to identify a novel gene were free to name it without consulting any central body or rules beyond those community standards. These standards generally came in the form of species-specific naming conventions. Researchers working with the yeast *Saccharomyces cerevisiae* frequently combined three letters and a number, for instance *FZO1*, to name genes. Some specialties were more flamboyant. Consider the gene *sonic hedgehog*, named for a spiky computer-game character, and the gene *yippee*, capturing a scientist’s apparent delight at discovery. Both genes were detected in the fruit fly *Drosophila melanogaster*.

Many cases exist where, with so many genomes now sequenced, conserved genes are named one thing in one organism and something else in another. One example is *lov-1* in round worms and *PKD1* in people, genes of interest in part because the latter is implicated in human polycystic kidney disease. Names that seemed meaningful when used among a handful of genes can be confusing in the context

of thousands. For instance, the pair of gene names *superman* and *kryptonite* is significant for a research community concerned with one model organism, the cress *Arabidopsis thaliana*, where a suppressing action of *kryptonite* upon *superman* is observed. But such monikers would make no sense in another organism where only one of the pair is present.

To tackle this problem, a number of organizations are attempting to standardize gene naming. The Life Science Identifiers project aims to make biological identifiers consistent and usable across different databases. A parallel goal motivates the Human Proteome Organisation, a global collaborative group dedicated to furthering proteomics (the genome-level study of proteins). As a small part of its mission, the Gene Nomenclature Committee of the Human Genome Organization is weeding out some “jokey” names that in the wrong context could be disturbing. Last year they renamed three human homologs of the “fringe” genes found in flies (which previously were





**Figure 7.** Huge insights into the structure and function of DNA, proteins and genes came in the 20th century through intensive efforts by molecular biologists in the U.S. and abroad. In the 21st century, genomics promises to greatly complicate what is already known. One certainty is that high-throughput experiments, including coming ENCODE studies, will produce huge amounts of new data. A fundamental challenge will be figuring out how to make the most of it.

named lunatic fringe, manic fringe and radical fringe) hoping to avoid worrying medical patients with such labels. These genes participate in the widely conserved notch gene signalling pathway. (See Figure 6)

### An ancient problem

The problem of devising standard nomenclature and classification categories is not unique to molecular biology. This problem lurks in many walks of life, whether in organizing stars in distant galaxies, books in a library or inventories on store shelves. The tremendous success of the World Wide Web has highlighted the importance of classifying information well on a large scale.

One of the most exciting proposals to improve standardization of online information is called the Semantic Web, a hallmark technology of the next generation of the World Wide Web. There, hyperlinks take on a meaning beyond simple connection and represent standardized relationships between a pair of entities. For instance, a given link from one high school student's homepage to another's might represent her relationship to that person (eg. "friend of"). Links from a page about automotive parts might convey the car model each was designed for (representing a

"part of" relationship). One can easily see how the code of a web page marked up this way could be mined to extract meaningful information.

But how should all these relationships be standardized and organized? Computer scientists address this formally by developing precise specifications for a knowledge-classification system -- an ontology. One of the challenges in creating an ontology is the tremendous amount of domain-specific knowledge that is required. Often no single person (or group) holds enough, so complete coverage of a domain can require a collective effort. Informally, collecting distributed intelligence from "many eyes" into a rough classification has proven successful on a number of prominent Web 2.0 sites. Rough classifications are established on Flickr for photographs, on Delicious for links and, perhaps most visibly, on Wikipedia, for knowledge in general. The latter is an ever-evolving encyclopedia that is far larger than any one person, or reasonably sized group, could produce. And it can be updated quicker than its traditional print counterparts. A parallel movement has begun in biology to develop a community tagging system for genes and proteins. The Wiki-Proteins project, for instance, encourages distributed annotation of proteins.

So what does all this have to do with molecular biology? Today effective gene classification can be thought of as a four-step process. First, a gene is identified and named as precisely as possible given the limited information available at the time of discovery. Second, based on functional experiments or sequence comparison, brief descriptions of that gene get compiled. Third, from that data, standardized keywords can be created and used to categorize genes. Finally, those categories can be arranged into a hierarchy or related organizing template. The fourth and final step represents the state of the field at the moment. For instance, it is this sort of arrangement of functional terms that GO provides.

The main limitation to this approach will be its dependence on human curators. To understand where gene nomenclature could move next, consider another example from the Internet and the very different organizational approaches initially taken by the search engines Google and Yahoo. Yahoo originally was a manually curated directory, a DAG-like structure devised and revised by human hands. Web users submitted sites and curators slotted them into various categories. This approach made sense initially, when the number of new Web

sites to add daily was small. But it was quickly challenged by the scalable, automated strategy of Google, which harvested keywords from each Web page and computed relationships between them based on their pattern of links. The very early Yahoo approach is similar in a sense to classifying genes with GO, dependent on meticulous study of each new specimen by careful human curators. But a large volume of new objects can completely paralyze even the most dedicated team. Precisely this problem is occurring in biology today.

It may be that an information handling strategy that blossomed on the World Wide Web can help scientists solve this problem. What if molecular biologists abandoned manual classification of the massive amounts of genomics data pouring their way and gave up asking where a novel or confusing entity should fit on a chart? Instead, as results arrive, measures of similarities could be automatically computed, generating a Google-style collection of interlinked, interwoven information. With this approach, we would lose the comforting anchor of quaintly named genes and reassuring images of hierarchical classification. But we would gain, perhaps, a more useful and robust understanding

of the myriad ways biological entities can be similar and can interact.

Of course, significant scientific labors await anyone pursuing this approach. Researchers will have to choose which standardized attributes are important enough to explore in any given high-throughput experiment. They must decide how to best cluster genes and their products to illuminate the truly important connections. Such challenges -- opportunities really -- are among the fruits genomics brings to biology.

## References

- The ENCODE Project Consortium 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799-816.
- Gerstein, M. B. et al. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Research* 17:669-81.
- Seringhaus, M. R., Cayting, P. D., and Gerstein, M. B. 2008. Uncovering trends in gene naming. *Genome Biology* 9:01.
- Kapranov, P. et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316:1484-8.
- The Gene Ontology Consortium. 2008. The Gene Ontology project in 2008. *Nucleic Acids Research* 36:440-4.
- Lu, L. J. et al. 2007. Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends in Biochemical Science* 32: 320-31.
- Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* 284:29-37.
- Finn, R.D. et al. 2008. The Pfam protein families database. *Nucleic Acids Research* 36:281-8.
- Mewes, et al. MIPS: analysis and annotation of genome information in 2007. 2008. *Nucleic Acids Research*. 36:196-201.
- UniProt Consortium 2007. The universal protein resource (UniProt). *Nucleic Acids Research* 36:190-5.
- Zheng, D., Gerstein, M. B. 2007. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends in Genetics* 23:219-24.
- Kapranov, P., Willingham, A. T., Gingeras, T.R. 2007. Genome-wide transcription and the implications for genomic organization. *Nature Review Genetics*. 8:413-23.
- Wu, J. Q. et al. 2008. Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biology* 9:R3.

For relevant Web links, consult this issue of *American Scientist Online*:

[http://www.americanscientist.org/Issue TOC/issue/1001](http://www.americanscientist.org/Issue%20TOC/issue/1001)