A New Function Evolved from Gene Fusion

Manyuan Long

References	This article cites 25 articles, 13 of which can be accessed free at: **http://genome.cshlp.org/content/10/11/1655.full.html#ref-list-1** Article cited in: **http://genome.cshlp.org/content/10/11/1655.full.html#related-urls**
Creative Commons License	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**
Correction	A correction has been published for this article. The contents of the correction have been appended to the original article in this reprint. The correction is also available online at: http://genome.cshlp.org/content/11/2/308.1.full.html

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

A New Function Evolved from Gene Fusion

Manyuan Long[1]

Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois 60637, USA

What constitutes genetic difference among organisms? How do new gene functions originate in nature? Since the early days of molecular biology, we have known that homologous genes between species differ in DNA and protein sequence. Noncoding regions have also been evolving with repetitive sequences, transposable elements, and other elements continuously reshaping genomes of organisms. As more genomes of humans and other organisms are examined, it also becomes clear that species differ not only in these two genomic parameters but also in the number and kinds of genes.

Genes are subject to a life and death process: New genes have originated continuously throughout evolution. For example, *Drosophila melanogaster* contains 87 cuticle protein genes, while *Caenorhabditis elegans* contains no such genes in its genome (Rubin et al. 2000). If this is thought to be comparing too-divergent organisms, take a look at recently divergent sibling species. *Drosophila teisseiri* and *Drosophila yakuba* contain a gene called *jingwei* (Long and Langley 1993; Wang et al. 2000), which originated only 2.5 million years ago. *D. melanogaster* itself has a unique gene *Sdic,* which expresses particularly in the sperm tail and does not exist in even its closest relative species (Nurminsky et al. 1998).

New genes often give rise to new biological functions driven by adaptive Darwinian selection (Long and Langley 1993; Chen et al. 1997; Begun 1997; Nurminsky et al. 1998). New genes may even have controlled the origination of new species, for example, *Odysseus,* a homeobox duplicate gene in *Drosophila* (Ting et al. 1998). Such new genes are associated with two conspicuous

changes consistent with origin of new functions: High protein substitution rates and drastic changes in gene structure. *Drosophila* is not the only organism whose genome has been found to originate new protein-coding genes differentiating one species from another. Other organisms, including plants and mammals, also have newly originated genes. For example, the *Mus musculus* genome contains multiple copies of the new gene *SP100-rs,* which is absent in its sibling species *Mus caroli* (Weichenhan et al. 1998), though little detail of its evolution and function is known. In potato, a new cytochrome c1 originated a mitochondrial targeting function (Long et al. 1996). Retrosequences may have contributed to the origin of new vertebrate regulatory elements or new parts of vertebrate coding regions (Brosius 1999). In these cases, recombination of protein modules and gene duplication played essential roles in creating the initial gene structures, and natural selection participated in the subsequent evolution.

Although insights from young chimerical genes in *Drosophila* have enormously changed our views of new gene evolution, good data from humans or mammals have been lacking. This is a significant hurdle for understanding new gene evolution in the genetic systems of the human and its primate relatives. In this issue, Thompson et al. (2000) present a clear example of how new genes with novel functions can originate in humans and other mammals, including the molecular process and derived biological function. A closer look at the origination of this new gene, *Kua-UEV*, offers insights into the general problem of human gene origination.

UEV is a conserved gene, distributed across all major eukaryotic lineages ranging from animals to fungi, plants, and protozoa. The UEV proteins in these organisms share multiple functions, for example, cell protection, c-FOS tran-

scription, and cell-cycle progression (Sancho et al. 1998; Thomson et al. 1998; Xiao et al. 1998). In *Saccharomyces cerevisiae,* the UEV protein controls elongation of polyubiquitin chains when associated with ubiquitin-conjugating enzymes (E2; Hoffman and Pickart 1999). The *UEV* genes in divergent organisms have maintained a very conserved structure in its common domain (C domain). However, there exists an additional domain (B domain) in one isoform of the human gene that does not exist in other organisms and, thus, creates a new, chimerical gene structure. How did this new structure originate, and where does the B domain come from?

From the first glimpse, this human gene is reminiscent of the chimerical structure of two *Drosophila* young genes. The first example is *jingwei,* which is composed of a major domain and an additional N-terminal domain (Long and Langley 1993). Recent work implies that the mosaic structure of *jingwei* was created by insertion of the retrosequence of the alcohol dehydrogenase gene into a previously existing gene, recruiting a portion of the N-terminal domain (Long et al. 1999; Wang et al. 2000). The second example is *Sdic,* which was created by a deletion in two adjacent genes at the DNA level (Nurminsky et al. 1998). However, the human *UEV* gene seems to have taken a different evolutionary route to acquire its additional B domain (Fig. 1).

In the genomic databases of *D. melananogster* and *C. elegans,* two small DNA fragments unrelated to the *UEV* gene in these species were found to be significantly similar to the B domain of the human *UEV* gene. Further analysis showed that these are seven exons encoding a 319–amino acid protein in *C. elegans* and five exons encoding a 326–amino acid protein in *D. melanogaster.* This newly discovered gene, named *Kua* (derived from the word "Cua" in Catalan, which means "tail" or "queue") en-

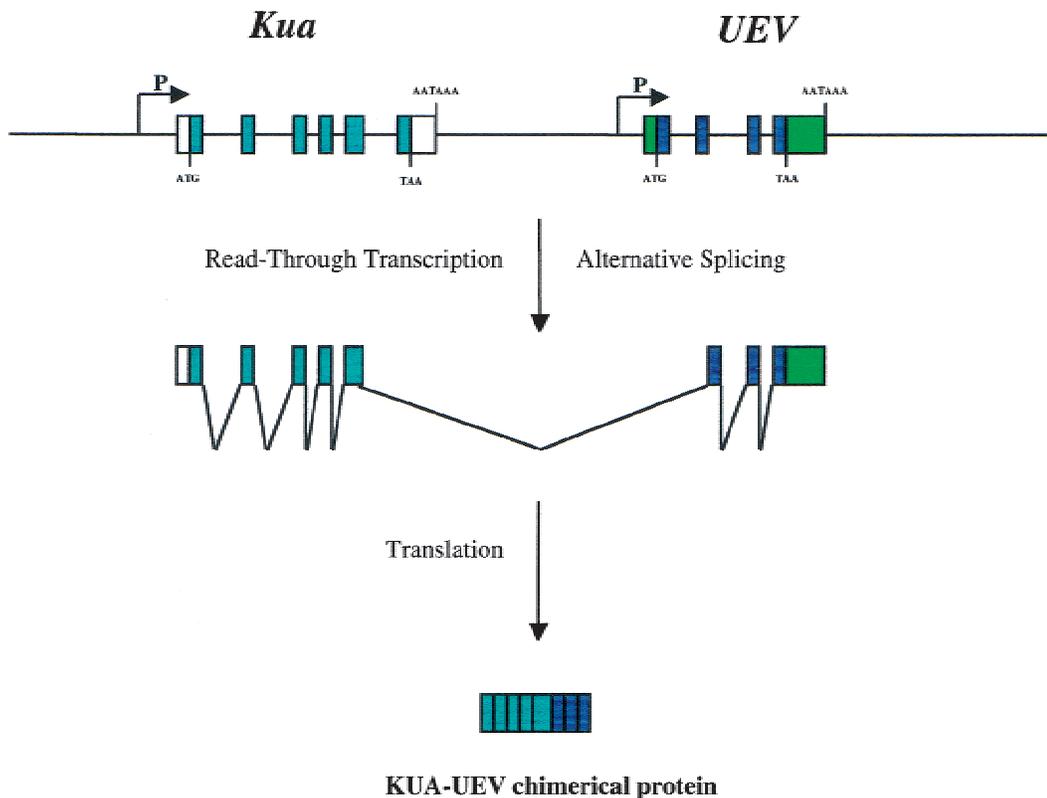[1]E-MAIL mlong@midway.uchicago.edu; FAX (773)702-9740.

Figure 1 The molecular process for *Kua-UEV* gene fusion.

codes a protein having features reminiscent of fatty acid hydroxylase. *Kua* was also detected in other species (*M. musculus, Trypanosoma cruzi,* and *Arabidopsis thaliana*) but was not found in *S. cerevisiae* genome sequences.

What is the linkage relationship between *Kua* and *UEV*? In *D. melanogaster, Kua* and the *UEV* gene are separated by 2.5 million bases in chromosome 1, while in *C. elegans* these genes are located on two different chromosomes. Thus, the genes *Kua* and *UEV* are simply different loci. However, in the human genome these two loci are adjacent by several kilobases, and a portion of RNA transcripts from the two genes is fused into a single RNA. This fused transcript structure may result from a relatively weak terminating signal for *Kua* gene transcription. A similar mechanism is responsible for generating read-through transcripts of the L1 element and its downstream cellular gene sequences (Boeke and Pickeral 1999; Moran et al. 1999).

Gene fusion has been observed be-

fore in various organisms. The classic examples are the fatty acid synthase gene (McCarthy and Hardie 1984) and tryptophan synthethase gene in fungi (Burns et al. 1990). Other noted cases include *HisA* and *HisF* in the histidine pathway (Lang et al. 2000), glutamyl- and prolyl-tRNA synthetase genes (Berthonneau and Mirande 2000), the young fusion gene Sp100-rs in *M. musculus* (Weichenhan et al. 1998), and the old fused genes of ubiquitin and ribosomal proteins in diverged organisms like yeast and human (Kirschner and Stratakis 2000). In bacteria and archaea, gene fusion was genomically surveyed in a number of species whose genomes have been sequenced (Snel et al. 2000). However, the human *Kua-UEV* gene fusion provides a revealing case regarding several important aspects of new protein origin.

First, a fused transcript is not a synonym for a fused protein. Distinct proteins in prokaryotic organisms are organized in operons, long transcripts encoding many proteins; many *C. elegans*

genes are also encoded in an operon-like structure (Blumenthal and Spieth 1996). Thus, an authentic gene fusion should possess a particular mechanism to override the nonsense codon used to stop translation of the N-terminal protein. For example, a mutation-like insertion in the stop codon would continue translation for a fused protein (Burns et al. 1990). However, the *Kua-UEV* human gene uses another, more sophisticated mechanism to solve the problem. Taking advantage of the more efficient splicing system in eukaryotes, *Kua-UEV* employs alternative splicing to skip the exon k6 of *Kua* that contains the *Kua* stop codon and exon A of *UEV* that contains a translation initiation codon. Given that many vertebrates genes often contain long UTR regions and an intergenic region, alternative splicing may be an efficient mechanism to avoid the stop codon in upstream gene(s), as represented by the *Kua-UEV* gene. These long stretches of noncoding DNA may contain many stop codons, and the random peptides translated from such DNAs may not be able to provide useful folds. Thus, one can predict that, in the future, it would not be unusual to find gene fusion products using this existing cellular mechanism, rather than waiting for a mutation in the stop codon.

What is the evolutionary advantage of gene fusion? Conspicuously, covalently connected proteins would ensure coregulation of gene expression of related functions. The covalently linked proteins can ensure stoichiometric production of the component peptides (McCarthy and Hardie 1984). Gene fusion also confers other advantages for particular proteins. For example, the multifunctionality of fatty acid synthase prevents dissociation at low protein concentration (McCarthy and Hardie 1984). In these ideas or experiments, the fused proteins are viewed as linked independent functional units. In the case of *Kua-*

UEV, however, a new advantage arises: The fusion creates a new function for *UEV* enzymatic activity. The nonfused form of UEV proteins, *UEV1A,* is intracellularly located in the nucleus, while KUA proteins are distributed in endomembranes. Consistent with the location of KUA proteins, KUA-UEV proteins were shown to be associated with cytoplasmic structures. Thus, the fused UEV enzymes work in new intracellular locations, suggesting the origin of a new gene function. This fused gene mimics some chimerical genes created by exon shuffling, for example, the coxII gene (Nugent and Palmer 1991) and the potato cytochrome c1 (Long et al. 1996), where the N-terminal-recruited portions also ensure a particular intracellular position for the enzymatic activity encoded in the C-terminal peptide.

Keeping original functions may be a premise in the creation of new genes. Gene duplication is often involved in exon shuffling, suggesting selection pressure for maintaining the function of donor genes. Even as *UEV* and *Kua* are fused together, they have kept their original separate functions. *UEV* has a duplicate copy (*UEV2*), but no duplicate copy of *Kua* has been found yet. Is this a reason for the fused gene to generate independent transcripts? While there may be other possible reasons for this pattern of transcription, speculation like this cannot be discounted.

It is often thought that the function of fused genes and chimerical genes is simply an addition of functions in preexisting component genes. If so, one would predict that the component sequences in such genes would evolve at a neutral substitution rate. However, this prediction is inconsistent with an observed phenomenon of accelerated evolution in chimerical genes (Long and Langley 1993; Long et al. 1996). A recent structural analysis of the histidine biosynthesis components *HisA* and *HisF* indicated that the protein structure after gene fusion was also subject to structural and functional adaptation (Lang et al. 2000). In this sense, gene fusion may be a critical step toward creating a new gene with novel function.

Is the function of the fused *Kua-UEV*

created recently in humans and its close relatives? In mouse, *Kua* and *UEV* may be in close proximity, because a hybrid transcript of these two genes was also observed (T.M. Thomson, pers.comm.). However, these two mouse genes generate different hybrid transcripts, suggesting that the fused protein and its functions may have evolved recently in humans or their primate ancestors. It should be possible to demonstrate or falsify this hypothesis by characterizing *Kua* and *UEV* genes in our primate relatives.

Finally, *UEV* genes also possess an interesting exon-intron structure that is telling about intron evolution. *UEV* has unusually conserved positions of intron 2 and 3, identical among plants, fungi, animals, and protozoa. These two introns thus should date back to 1–2 billion years ago. However, the authors' interpretation of intron 4 in *Schizosacchromyces pombe* requires some caution. This intron is interpreted as a new arrival by recent intron insertion. This explanation seems plausible because it is the only intron among the analyzed organisms in phase 2 (i.e., the intron breaks a codon after the second nucleotide) and because it breaks the secondary structure of the UEV protein. However, an alternative hypothesis generates a biologically sensible prediction and, hence, may be considered for a test: If this intron is an ancient intron, like intron 2 and 3, the missing corresponding intron in all species except *S. pombe* would be the result of intron loss. The position of all these missing introns would be in the 3′ end exon. This would extend the interesting model of Fink (1987) from yeast to the organisms under investigation. In this model, introns loss, by reverse transcription and homologous recombination, should show a gradient from 3′ to 5′ in loss frequency. This alternative hypothesis predicts that some eukaryotic organisms, in addition to *S. pombe*, may still retain this intron.

REFERENCES

Begun, D.J. 1997. Genetics **145:** 375–382.

Berthonneau, E. and Mirande, M. 2000. *FEBS Lett.* **470:** 300–304.

Blumenthal, T. and Spieth, J. 1996. Curr. Opin. Genet. Dev. **6:** 692–698.

Boeke, J.D. and Pickeral, O.K. 1999. *Nature* **398:** 108–109.

Burns, D.M., Horn, V., Paluh, J., and Yanofsky, C. 1990. *J. Biol. Chem.* **265:** 2060–2069.

Chen, L., DeVries, A.L., and Cheng, C.H. 1997. *Proc. Natl. Acad. Sci.* **94:** 3817–3822.

Fink, G.R. 1987. *Cell* **49:** 5–6.

Hofmann, R.M. and Pickart, C.M. 1999. *Cell* **96:** 645–653.

Kirschner, L.S. and Stratakis, C.A. 2000. *Biochem. Biophys. Res. Commun.* **270:** 1106–1110.

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. 2000. *Science* **289:** 1546–1550.

Long, M. and C.H. Langley. 1993. *Science* **260:** 91–95.

Long, M., de Souza, S.J., Rosenberg, C., and Gilbert, W. 1996. *Proc. Natl. Acad. Sci.* **93:** 7727–7731.

Long, M., Wang, W., and Zhang, J. 1999. Gene **238:** 135–142.

McCarthy, A.D. and Hardie, D.G. 1984. *Trends Biochem. Sci.* **9:** 60–63.

Moran, J.V., DeBerardinis, R.J., and Kazazian Jr., H.H. 1999. *Science* **283:** 1530–1534.

Nugent, J.M. and Palmer, J.D. 1991. *Cell* **66:** 473–481.

Nurminsky, D.I., Nurminskaya, M.V., De Aguiar, D., and Hartl, D.L. 1998. Nature **396:** 572–575.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. *Science* **287:** 2204–2215.

Sancho, E., Vila, M.R., Sanchez-Pulido, L., Lozano, J.J., Paciucci, R., Nadal, M., Fox, M., Harvey, C., Bercovich, B., Loukili, N., Ciechanover, A., et al. 1998. *Mol. Cell Biol.* **18:** 576–589.

Snel, B., Bork, P., and Huynen, M. 2000. *Trends Genet.* **16:** 9–11.

Thomson, T.M., Khalid, H., Lozano, J.J., Sancho, E., and Arino, J. 1998. *FEBS Lett.* **423:** 49–52.

Thomson, T.M., Lozano, J.J., Loukili, N., Carrió, R., Serras, F., Cormand, B., Valeri, M., Díaz, V.M., Abril, J., Burset, M., et al. 2000. *Genome Res.* **10:** 1743–1756.

Ting, C.T., Tsaur, S.C., Wu, M.L., Wu, C.I. 1998. *Science* **282:** 1501–1504.

Wang, W., Zhang, J., Alvarez, C., Llopart, A., and Long, M. 2000. Mol. Biol. Evol. **17:** 1294–1301.

Weichenhan, D., Kunze, B., Traut, W., and Winking, H. 1998. *Cytogenet. Cell Genet.* **80:** 226–231.

Xiao, W., Lin, S.L., Broomfield, S., Chow, B.L., and Wei, Y.F. 1998. *Nucleic Acids Res.* **26:** 3908–3914.

Errata

Genome Research 10: 1655–1567 (2000)

A New Function Evolved from Gene Fusion
Manyuan Long

The following reference was omitted:

Brosius, J. 1999. *Gene* **238:** 116–134.

Genome Research 10: 1697–1710 (2000)

Sequence and Comparative Analysis of the Mouse 1-Megabase Region Orthologous to the Human 11p15 Imprinted Domain
Patrick Onyango, Webb Miller, Jessica Lehoczky, Cheuk T. Leung, Bruce Birren, Sarah Wheelan, Ken Dewar, and Andrew P. Feinberg

The URL http://www.jhmi.edu/feinberg_lab found in the Methods section under the heading 'Sequencing of the Mouse Contig' should instead be http://www.hopkinsmedicine.org/imprinting.