Evolution of novel genes

Manyuan Long

Much progress in understanding the evolution of new genes has been accomplished in the past few years. Molecular mechanisms such as illegitimate recombination and LINE element mediated 3′ transduction underlying exon shuffling, a major process for generating new genes, are better understood. The identification of young genes in invertebrates and vertebrates has revealed a significant role of adaptive evolution acting on initially rudimentary gene structures created as if by evolutionary tinkers. New genes in humans and our primate relatives add a new component to the understanding of genetic divergence between humans and non-humans.

Addresses
Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA;
e-mail: mlong@midway.uchicago.edu

Current Opinion in Genetics & Development 2001, **11**:673–680

0959-437X/01/$ – see front matter

Abbreviations

AFGPs	antifreeze glycoproteins
L1	LINE-1
LDL	low-density lipoprotein
LINE	long-interspersed nuclear element
MCH	melanin-concentrating hormone
TEs	transposable elements

Introduction

Notwithstanding the fact that speculations about evolutionary novelties can be traced back to Darwin [1], the mechanisms and processes involved in the origin of evolutionarily new genes were not observable until the late 1970s and early 1980s when appropriate molecular and biochemical techniques were developed. Productive approaches emerged in the early 1990s via two favorable conditions: first, enormous databases of DNA and protein sequences and structures; second, the introduction of the concept of young genes.

The complete process for the birth of a novel gene comprises initial mutation events, yielding a particular gene structure, and the subsequent evolutionary process, in which the new gene structure is fixed in the whole species and improved for some novel function(s). Present-day sequence databases have archived sequence and structural information of an astronomical scale, enabling the comparison of various gene structures (e.g. intron–exon structure or protein modules) to postulate initial molecular processes [2–4].

Much information for early changes in gene structure and sequence for the further improvement of acquired novel functions may be missing from those genes that are retrieved from the database, however, because these genes are often so old that this early evolution has been obscured by later changes. An efficient approach is the direct examination of a gene that originated recently (i.e. several million years ago) whose sequences and structures retain initial features of evolution. A technical challenge is the identification of young genes limited to related species. Fortunately, genome sequences and innovative molecular techniques remove the impasse to finding young genes; a number of informative examples have been identified.

Results from these new approaches have changed our view about the makeup of genes and genomes. In this review, we summarize the major progress that has been achieved recently and consider its implications to the understanding of evolution of genetic systems.

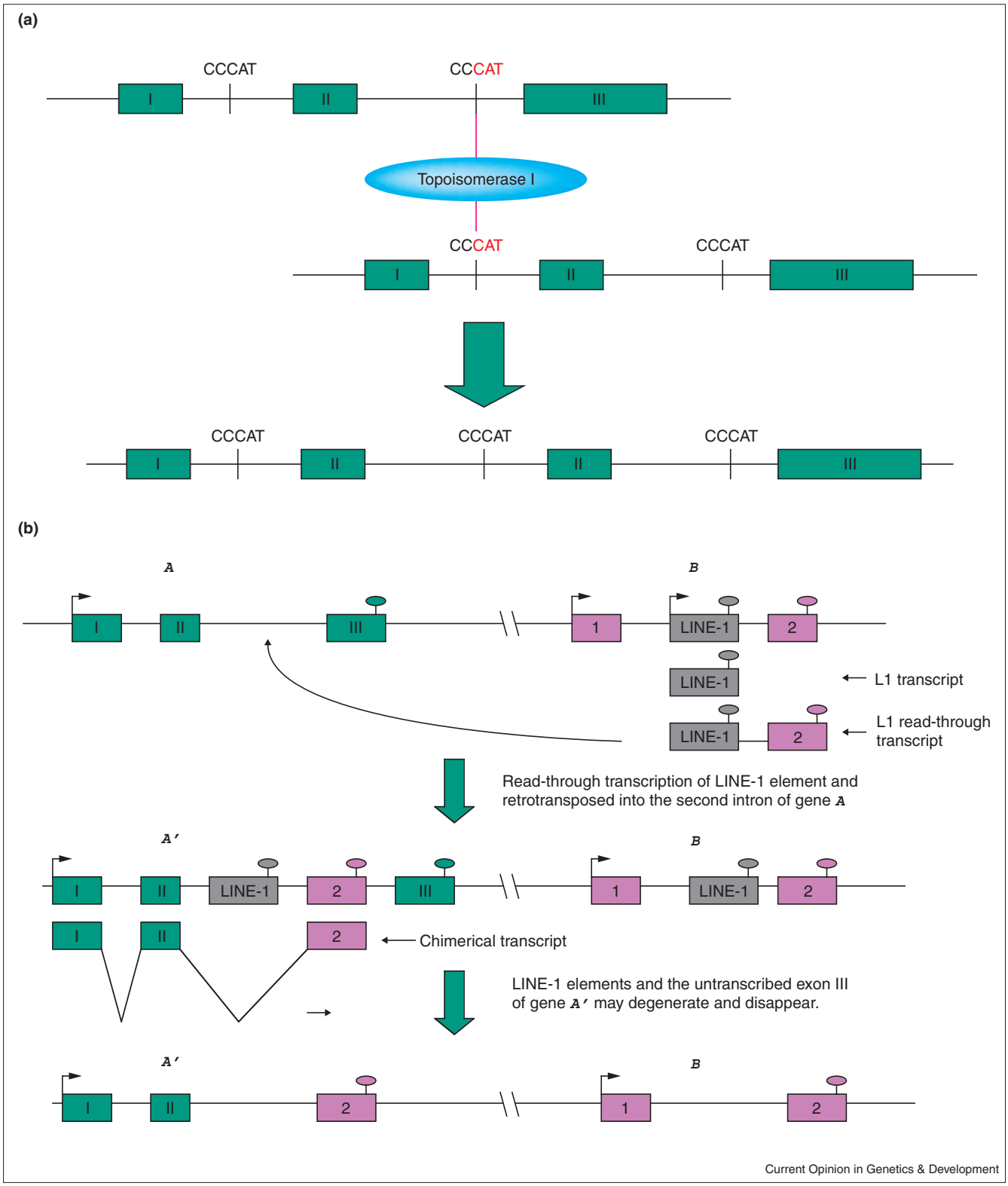
Exon shuffling: mechanisms and rates

Since its proposal in the late 1970s, exon shuffling has been demonstrated to be an efficient process for generating new genes [5]. This proposal assumed that new combinations of exons from different genes, representing subunits of functional structure, may arise by non-homologous recombination and evolve into new genes with novel functions and that introns will speed up the formation of new combinations and, thus, the rate of new protein evolution. Among the first-detected examples of exon shuffling was the low-density lipoprotein (LDL) receptor [6]. Patthy [7,8] surveyed protein databases systematically and showed that exon shuffling occurred in many genes of vertebrate and invertebrate organisms. New genes created by exon shuffling in plants were also observed (e.g. in potatoes [9] and sunflowers [10]). Unfixed spontaneous events of exon shuffling were also observed in mouse [11] and tomato [12]. These observations suggest that exon shuffling is a general mechanism for the origin of new genes.

The mechanisms that drive exon recombination, however, remained a mystery for two decades until recent work in two areas: illegitimate recombination and LINE-1 (L1) element mediated recombination. Illegitimate recombination, the DNA recombination between sequences sharing little or no homology, was initially proposed as a molecular mechanism for exon shuffling [5,13]. Brosius [14] emphasized the importance of retroposition in gene evolution and Gilbert *et al.* [15] indicated that retroposition was a dominant form for exon shuffling.

Bloemendal and co-workers [16,17] revealed the first clear molecular mechanism resulting in illegitimate recombination leading to exon shuffling in the small heat-shock protein αA-crystallin gene of the hamster. By transfecting a construct of the hamster αA-crystallin gene into a mouse muscle cell line, these authors identified a mutant αA-crystallin gene with a large intragenic duplication (Figure 1a). The sequence of this mutant suggested it had been created

Figure 1



Molecular mechanisms for exon shuffling. **(a)** Illegitimate recombination leads to exon duplication for super αA crystallin protein. The recombination took place between two non-homologous sites with the identical sequence CCCAT in intron I and II of the super αA crystallin protein. Topoisomerase I preferentially nicks the sequence CAT, suggesting that the two new CCCATs function in illegitimate recombination. **(b)** Exon-shuffling by LINE-1 mediated 3′ transduction. *A* and *B* are two hypothetical genes to explain the transduction process. *A′* is a new gene created by the process. The transcription start site and terminating signals are marked by arrows and ovals, respectively. LINE-1 has weaker transcription terminating signals (ovals) that yield a read-through transcript including exon 2 of gene *B*.

by illegitimate recombination between two CCCAT sites in two αA-crystallin genes, one at intron 3 and the other in exon 2 (which is an intron in another isoform of αA cystallin), resulting in an internally-duplicated exon structure. It has been demonstrated that the triplet sequence CAT is preferentially nicked by topoisomerase I involved in illegitimate recombination, implicating these two CCCATs in the act of illegitimate recombination.

Such a specific sequence requirement, however, would imply a low rate of recombination, which may not be able to account for the observed higher frequency of exon shuffling. Other mechanisms can also lead to exon shuffling. A mechanism reported to be potentially efficient [18–20] is called L1 element mediated 3′ transduction (Figure 1b). L1 is a retrotransposon that can reverse transcribe and move in the mammalian genome. Moran *et al.* [18] demonstrated in a transfection experiment that L1 can, as a consequence of its associated weak transcription terminating signal, yield a read-through transcript at an appreciable frequency. A 3′ flanking genomic region-derived fragment in such a transcript would also move with the LINE element on the same transcript. Depending upon the position of LI, a whole nuclear gene or exons downstream of L1 can be carried with L1 and recombine with exons of a recipient locus. Given the high abundance of L1 elements in mammalian genomes [21–23] (~15% of the human genome), L1 element mediated 3′ tranduction likely represents a frequent mechanism to shuffle genome sequences. Genomic analyses revealed that sequences transduced by this mechanism account for ~1% of the human genome [24•,25•], although the new functional chimerical genes created by this mechanism have not been documented from these L1-derived human sequences [19].

How often does exon shuffling generate new genes? Patthy [7,8] lists hundreds of protein families that have been made by this mechanism; most are mammalian genes. Statistical analyses of intron phase, defined as the relative position of introns either within or between codons, have revealed dominant phase zero introns and significant intron phase correlation within genes (i.e. exons tend to be flanked by introns of the same phase). This non-random distribution of intron phase suggests that a large portion of eukaryotic genes originated from exon shuffling [26,27]. An alternative explanation, however, is that non-random distribution of hypothetical proto-splice sites for intron insertion [28,29] (e.g. AG↓G exon sequence ['↓' indicating the intron insertion site]), would also generate a non-random distribution of intron phases [30,31]. Assuming that the sites where introns reside are just a random sample of those candidate proto-splice sites, two studies investigated the distribution of these sites [32,33••]. The currently observed intron phase distributions differ significantly from the distributions of the hypothetical proto-splice sites including the AGG base triplet. These recent analyses continue to support the conclusion that exon shuffling plays an important role in new gene evolution.

Retroposition and gene translocation

Retroposition has been viewed as sowing the 'seeds' for evolution of novel gene functions [14] rather than representing merely the dead ends (pseudogenes) in evolution, because of two properties. The first is that retroposition often occurs into a new genomic environment where a fortuitously evolved regulatory system from the nearby genomic sequence might resurrect the otherwise doomed insertion with a new expression pattern. The second property is that the retrosequence may become part of a new gene by recruiting nearby pre-existing coding and noncoding sequences to form a new chimerical gene. Many cases have been uncovered for these two routes to generate new gene function. For example, the *BC200* RNA gene that is expressed specifically in the nervous system of primates was created after the insertion of a 7SL RNA-derived monomeric *Alu* element recruited a unique genomic region [34]. Brosius [35] reviewed the role of retroposition in evolution of new genes and new regulatory systems.

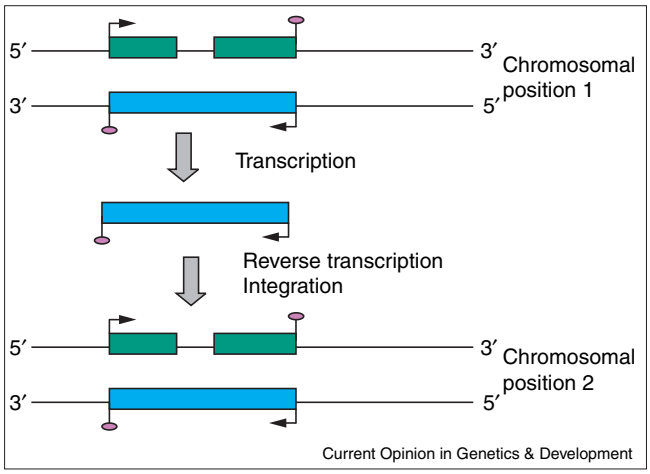
Several molecular features are often viewed as hallmarks of retropositions: the loss of introns, poly A tracts, and flanking short direct repeats [36]. These are indeed useful characters for identifying the genes created by retroposition, leading to the identification of numerous examples in the human genome [37,38]. These three features, however, might only represent a portion of actual retroposition events. Retroposition of intronless genes is not associated with intron loss. Moreover, these features can be maintained in nature only for a limited evolutionary time — introns can be inserted into intronless genes and the poly A tracts and flanking repeats may be eroded in a short time.

It had not been anticipated that a retroposition event could retain an intact intron–exon structure by reverse-transcribing the RNA encoded by the antisense strand of a gene. Nahon and co-workers [39••,40,41] identified two chimerical genes originating recently in Hominidae. PMCHL1 and PMCHL2 (proMCH-like 1 and 2 genes) have derived from the melanin-concentrating hormone (MCH) gene by reverse transcription and integration of a long exon in the *AROM* (antisense-RNA-overlapping-MCH) gene encoded in the complementary strand of the MCH gene. Because the RNA sequence for *AROM* does not itself contain the splicing site sequence of the MCH gene, the intron–exon structure of the MCH gene overlapping *AROM* has been maintained after the integration of retroposed sequences. A more general process to summarize the mechanism to preserve intron–exon structure for retroposition is given in Figure 2. This mechanism casts doubt on a prevailing view that chromosome translocation is responsible for dispersed gene family members with introns in many situations. In such cases, identification of poly A tracts and flanking repeats may provide clues to origin by retroposition, if the genes are young enough to retain these signatures.

New functions evolved from gene duplications

The classic model for the origin of new gene functions is based on gene duplication [42], proposing that while one

Figure 2



Schematic of an intron-containing gene duplicating and moving to another chromosomal position. (Arrow, transcription start site; oval, transcription terminating site.) Two strands of the same DNA regions encode two unrelated genes, with one (upper) having an intron and the other having no intron. Retroposition from the transcript of the intronless gene creates a new locus in chromosomal position 2, in which the structure of the intron-containing gene is also retained.

copy of a pair of duplicate genes maintains the original function, the other copy can accumulate mutations for further evolution of new functions. Although many theoretical models have been proposed to describe the gene duplication process, it should be emphasized that gene duplication is not synonymous with the gaining of new functions. A duplicate gene can have several evolutionary fates: first, it becomes a pseudogene; second, it maintains redundant functions; and third, it gains new functions. Walsh [43] examined the probability that a gene duplication evolves new functions under a simplified assumption incorporating only the first and third of these scenarios. He showed that a duplicate copy is in general much more likely to become a pseudogene instead of a new functional gene. He also showed that if a duplicated gene acquires an even slightly advantageous function, then it is unlikely to become nonfunctional in subsequent evolution. Ohta [44] proposed a probability model to examine the role of neutral mutation in the evolution of new gene functions. Gu [45] proposed a statistical method to detect potential functional divergence between duplicate genes.

There are many new functional genes evolved from gene duplications — for instance, hemoglobin genes in humans, in which duplicate copies are expressed at different developmental stages. A compelling case was reported recently in centromeric H3-like proteins in *Drosophila* [46••]. A member of this family, the *Cid* gene, in *D. melanogaster* functions in the centromere to determine the specificity of centromeric DNA binding. An evolutionary feature of centromeric DNA is the fast change of its component satellite repeats as a consequence of insertion of new mobile elements or loss of old repeats. For example, in a closely related *Drosophila* species that diverged for only 2–3 million years, the components of centromeric DNA diverged significantly. A biological question is how could the function of *Cid* respond to such rapid change in centromeric DNA? Malik and Henikoff [46••] find that *Cid* in *D. melanogaster* and *D. simulans* adaptively evolved new binding functions by equally rapid changes in its protein sequence. A recent laboratory experiment by Rosenzweig and co-workers [47] also showed that new duplicate copies of hexose transport genes in *Saccharomyces cerevisiae* originated only following 450 generations of selection in a glucose-limited medium. This interesting experiment, with early selection experiments in micro-organisms for the origin of new gene functions [48,49] demonstrated the significant role of selection in the origin of new gene functions.

Mobile elements and lateral gene transfer: more sources for new protein function

Transposable elements (TEs) have been shown to contribute to the creation of protein diversity. Makalowski and Boguski [50] and Makalowski [51•] have surveyed vertebrate genomes, identifying >200 cases in which various mobile elements encoded portions of cellular proteins and changed the functions of genes recipient for the TE insertions (also see 'Update'). Because of the abundance of *Alu* elements in primate genomes, it serves more often than others as a donor of new peptides within the new proteins. An example is the human decay accelerating factor (DAF) that contains an *Alu* fragment. Because this new protein has a novel hydrophilic carboxy-terminal region as a consequence of the *Alu* insertion, it would have intracellular locations different from the original gene [51•,52]. These cases implicate TEs in the origin of new proteins, offering the opportunity for further investigation. For instance, it would be interesting to know the population and/or species distributions of these newly created genes, because it is unclear whether these new forms of proteins have been fixed in whole populations of humans with other primate species. It would also be interesting to decipher which evolutionary forces have acted on these genes.

Lateral gene transfer, the pass of genetic information from one genome to an unrelated genome, has been shown to be important in genome evolution of prokaryotes ([53,54]; see also reviews by Ochman [pp 616–619] and Ragan [pp 620–626], this issue). It is likely that novel genes with new functions can evolve from such a process in recipient species of the gene transfer. De Koning *et al.* [55•] provide one such case. It is found that the gene encoding N-acytylneuraminate lyase in the protozoan *Trichomonas vaginalis* shares a high similarity (80% identity) with the neuraminate lyase bacteria *Haemophilus influenzae* in protein sequence, suggesting that this is a recent transfer event. Remarkably, the newly transferred gene recruited a new leader sequence 24 amino acids long, which seems to be a signal peptide. This structural evolution is reminiscent of origin of presequences or transit peptides in many nuclearly encoded organellar proteins [9,56]. The *T. vaginalis* N-acytylneuraminate

lysase becomes a secreted protein whereas the bacteria neuraminate lyase is cytosolic, suggesting that the new gene in *T. vaginalis* may have evolved a new function.

The role of 'tinkers'

Various mechanisms for new gene evolution have been investigated, which often understandably point to deterministic optimization of the structure and functions of new genes. It may give insight to revisit the concept of 'tinkerism', the important but often overlooked idea of 'evolution as tinkering', proposed by Jacob [57] to describe how evolution works to generate novelties. In this view, evolution does not behave like a good engineer who always wants to do the best job with a well-prepared plan and specifically provided materials. Instead, evolution works like a tinker who uses whatever material comes to hand in making a device that crudely serves some new functions but does so in a far-from-perfect manner at first. Thus, a tinker can make a roulette table from an old bicycle wheel or a TV stand from a broken chair.

Nurminsky *et al.* [58] investigated the origin of *Sdic*, an evolutionarily new gene in *Drosophila*. It originated in the single lineage of *D. melanogaster*, after its split only three million years ago from its sibling species. *Sdic* is one of the two youngest genes known (the other gene, *Jingwei*, also in *Drosophila*, is under 2.5 million years old) [59]. Remarkable in the early life of this gene are the unusual origins of various of its essential parts: deletion created the chimera; a new exon evolved from an intron of the *Cdic* parent; and the new testes-specific promoter formed from an exon in parent *AnnX*. As the resources for its various parts are so dissimilar to their eventual uses, both functionally and structurally, one could scarcely have predicted that they would be connected together. These changes provide evidence of tinkerism.

Furthermore, the high rate of protein evolution in the above examples indicates imperfection of the original parental genes or gene fragments for the novel functions they eventually serve. Otherwise, these proteins would not have been so rapidly changed by the force of natural selection. The high substitution rates of the shuffled exons, as shown in *Sdic* and other new genes [9,59], however, suggests that the original exons were not adept in their new roles and needed further modifications by the diligent tinker. Thus, the concrete case of the *Sdic* gene and the rapid evolution of new genes known previously reveal tinkering evolution. This route of evolution, added to the powerful mechanisms of exon shuffling and duplication that provide novel but often imperfect genetic materials, would create a vast diversity of genes.

New gene functions follow Darwin's scenario

Whereas the mutation process that creates initial structure paints a dramatic picture for the first step in the evolution of new genes, the various evolutionary forces involved in the next step — the fixation of the new genes, at the various stages of their improvement in whole species — can next be considered.

Most new genes that originated from exon shuffling and gene duplication have undergone significantly elevated rates of evolution when compared to their parental genes. *Jingwei* has a significantly elevated rate of evolution in its protein sequence and gene structure, signifying strong protein adaptive evolution throughout its evolutionary history [59,60,61••]. *Sdic* evidences rapid sequence change and low within-species variation [58,62,63,64••]. Similarly, another new gene in *Drosophila*, *Fannegan*, which was generated by gene duplication 20 million years ago, also showed fast protein sequence evolution [65]. A plant cytochrome C1 precursor gene recruited a novel mitochondrial-targeting domain 100–120 million years ago, which evolved 30–50 times faster than its ancestral counterpart [9]. Ohta [66] noticed higher rates of evolution associated with functional divergence in some anciently duplicated genes. These examples are consistent with a role for Darwinian selection in shaping the structures of new genes.

Most attention has been focused on new genes that adopt different functions, a process shown to be governed by Darwinian selection. New progress has also been achieved in understanding convergent evolution. Novel genes in different lineages can evolve the same new function under similar selection pressures. The crystallins — eye lens proteins that contribute to the high refractive index needed for the lens to focus light — provide a good example for the same function being evolved from different proteins. Ancestral genes code for proteins as diverse as small heat shock proteins, lactate dehydrogenase B, and ornithine cyclodeaminase (e.g. [67,68]).

Antifreeze proteins provide an explicit example of convergence from disparate origins [69–72]. The antifreeze glycoproteins (AFGPs) in polar fishes form a family of proteins that bind to ice crystals in the cells and block further ice crystal growth. Whether in Antarctic or Arctic fishes, these proteins have a similar molecular phenotype — many glycotripeptides $(\text{Thr-Ala/Pro-Ala})_n$ interspersed with short peptide spacers — but these genes are very different in exon–intron structures, codon usage, and spacer sequences [69]. Furthermore, the AFGP in the Antarctic fish has very high similarity to trypsinogen genes, suggesting its recent origin (5–14 million years) from the latter gene [70]. Identification of a hybrid gene of AFGP and trypsinogen genes in an Antarctic fish, *Dissostichus mawsoni*, further confirmed the origin of the Antarctic AFGP [71]. Although it is unclear how Arctic AFGP originated, it is more parsimonious to infer a recent convergent origin for Antarctic AFGP.

Recently evolved genes in human

What are the genetic differences between humans and non-humans? Nucleotide substitutions have been found in

every gene examined to date that have homologues in other organisms; human genome structure differs in numerous details from other organisms, although the total number of genes in the human genome does not appear as big as was speculated [37,38]. Several investigations have revealed that novel genes originated in the human lineage, implying that humans may have human-specific genes.

Nahon and co-workers [39••,40,41] not only found a new mechanism to translocate an intact intron–exon structure to new positions within the genome, as described above, they also revealed two chimerical genes *PMCHL1* and *PMCHL2* in Hominidae. *PMCHL1* evolved 25 million years ago, before the old-world monkeys (*Catarrhini*) diverged, by a complex mechanism of retroposition described above, coupled with *de novo* creation of the intron–exon boundaries in the 3′ coding region recruited from non-coding genomic DNA, leading to a chimerical gene structure. A large duplication of the region encompassing *PMCHL1* in the Hominidae five million years ago yielded *PMCHL2*, a young gene appearing in humans and chimpanzees. RNA expression experiments revealed that *PMCHL1* and *PMCHL2* are specifically and differentially regulated in testis; *PMCH1* expresses in human fetus and brain, showing that these genes are tightly regulated [40,41].

Thompson *et al.* [73••] have reported a new protein encoded in the human genome by chimeric transcripts from two adjacent genes. The fusion protein — with its amino terminus deriving from the *Kua* gene and the carboxyl terminus deriving from the *UEV* gene — originated from a read-through transcript from the two genes, which are several kilobases apart. These genes are unlinked in *Caenorhabditis elegans* and far apart in *D. melanogaster*. Both genes also produce separate transcripts and separate, highly conserved proteins. The chimerical protein may have evolved a new function by acquiring new intracellular locations. It is unclear when the fusion event took place but it was found that the mouse counterparts of the two genes may be closely proximal but generated different hybrid transcripts, suggesting the fused protein and its functions may have evolved recently in humans or their primate ancestors [73••,74].

Conclusions

Significant progress has been made in understanding the initial stage in the evolution of new genes, the creation of new gene structure. Detailed molecular mechanisms underlying exon shuffling have been found. The classic candidate mechanism for exon shuffling, illegitimate recombination, was demonstrated under laboratory conditions, revealing further detail of how an intereaction between topoiosmerase and a specific recognition sequence signal can lead to the generation of exon duplication. A quite different type of illegitimate recombination, a new mechanism called L1 element mediated 3′ transduction, may not be a rarely used mechanism to recombine exons in genomes with high copy numbers of L1 elements, such as the human genome. Several other features of processes that create new genes have been detected or better understood as well, such as the generality of retroposition and gene duplication in the origin of new genes, and exaption of transposable elements into genes encoding cellular proteins.

The experimental search for young genes has emerged as a direct approach to observation of early stages in evolutionary process. It is noteworthy that adaptive selection is involved in the early history of new genes (e.g. convergent evolution of antifreeze proteins in Antarctic notothenioid fishes and several northern cods or the origin of *Jingwei* and other new *Drosophila* genes). Furthermore, these new genes, exemplified by *Sdic*, revealed 'tinkerism' among the processes shaping the early course of evolution of new genes. However, the tinkers in evolution of new genes do not, as speculated originally, work slowly and inefficiently; rather, under Darwinian selection, they have formed new genes rapidly and efficiently. New chimerical genes and fusion genes have been found in humans and our evolutionary relatives, adding a new element for consideration as part of the genetic divergence between human and non-human organisms.

Although major recent progress has been focused on the early molecular processes creating new genes, explorations of evolutionary forces that govern the fixation of new gene structures and subsequent sequence evolution have emerged, using young genes with novel functions as model systems. It can be anticipated that in the near future, more young genes in various organisms will be identified and the further investigation of such new genes will elaborate which evolutionary forces underlie their genesis.

Update

Nekrutenko and Li [75] have found that 4% of human genes contain TE-derived protein-coding regions and that TE integration may affect gene fucntion.

Acknowledgements

I thank all members, past and present, in my laboratory for their efforts to solve various challenging problems related to the origin of new genes, their stimulating discussions and questions. I am also indebted to Janice Spofford for valuable discussion and suggestions when I prepared this manuscript. This work is supported by grants from the National Science Foundation and the Packard Fellowship in Science and Engineering.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Darwin C: ***The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life.*** New York: Washington Square Press, Inc.; 1859.

2. Long M, de Souza SJ, Gilbert W: **Evolution of the intron-exon structure of eukaryotic genes.** *Curr Opin Genet Dev* 1995, **5**:774-778.

3. Gilbert W, de Souza SJ, Long M: **Origin of genes.** *Proc Natl Acad Sci USA* 1997, **94**:7698-7703.

4. De Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W: **Toward a resolution of the introns early/late debate: only phase zero**

introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA* 1998, **95**:5094-5099.

5. Gilbert W: **Why genes in pieces?** *Nature* 1978, **271**:501.

6. Sudhof TC, Goldstein JL, Brown MS, Russell DW: **The LDL receptor gene: a mosaic of exons shared with different proteins.** *Science* 1985, **228**:815-822.

7. Patthy L: **Modular exchange principles in proteins.** *Curr Opin Struct Biol* 1991, **1**:351-361.

8. Patthy L: **Genome evolution and the evolution of exon-shuffling — a review.** *Gene* 1999, **238**:103-114.

9. Long M, de Souza SJ, Rosenberg C, Gilbert W: **Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor.** *Proc Natl Acad Sci USA* 1996, **93**:7727-7731.

10. Domon C, Steinmetz A: **Exon shuffling in anther-specific genes from sunflower.** *Mol Gen Genet* 1994, **244**:312-317.

11. Jones JM, Huang JD, Mermall V, Hamilton BA, Mooseker MS, Escayg A, Copeland NG, Jenkins NA, Meisler MH: **The mouse neurological mutant flailer expresses a novel hybrid gene derived by exon shuffling between Gnb5 and Myo5a.** *Hum Mol Genet* 2000, **9**:821-928.

12. Chen JJ, Janssen BJ, Williams A, Sinha N: **A gene fusion at a homeobox locus: alterations in leaf shape and implications for morphological evolution.** *Plant Cell* 1997, **9**:1289-1304.

13. Gilbert W: **The exon theory of genes.** *Cold Spring Harbor Symp Quant Biol* 1987, **52**:901-905.

14. Brosius J: **Retroposons — seeds of evolution.** *Science* 1991, **251**:753.

15. Gilbert W, Long M, Rosenberg C, Glynias M: **Tests of the exon theory of genes**. In *Tracing Biological Evolution in Protein and Gene Structures, Proceedings of the 20th Taniguchi International Symp, Division Of Biophysics*. Edited by Go M, Schimmel P. Elsevier Science 1995:237-247.

16. Van Rijk AF, van den Hurk MJ, Renkema W, Boelens WC, de Jong WW, Bloemendal H: **Characteristics of super αA-crystallin, a product of *in vitro* exon shuffling**. *FEBS Lett* 2000, **480**:79-83.

17. Van Rijk AA, de Jong WW, Bloemendal H: **Exon shuffling mimicked in cell culture.** *Proc Natl Acad Sci USA* 1999, **96**:8074-8079.

18. Moran JV, DeBerardinis RJ, Kazazian HH Jr: **Exon shuffling by L1 retrotransposition.** *Science* 1999, **283**:1530-1534.

19. Kazazian HH Jr: **L1 retrotransposons shape the mammalian genome.** *Science* 2000, **289**:1152-1153.

20. Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH Jr: **A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion.** *Nat Genet* 1994, **7**:143-148.

21. Boeke JD, Pickeral OK: **Genome structure — retroshuffling the genomic deck.** *Nature* 1999, **398**:108.

22. Eickbush T: **Transcription: exon shuffling in retrospect.** *Science* 1999, **283**:1465.

23. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome.** *Nature* 200, **409**:847-849.

24. • Pickeral OK, Makalowski W, Boguski MS, Boeke JD: **Frequent human genomic DNA transduction driven by LINE-1 retrotransposition.** *Genome Res* 2000, **10**:411-415.

The authors, with [25•], report a computational analysis of the human genome for the frequency of *LINE-1*-mediated movement of human genomic DNAs. Surprisingly, it was found that the frequency was as high as 1%.

25. • Goodier JL, Ostertag EM, Kazazian HH: **Transduction of 3′-flanking sequences is common in L1 retrotransposition.** *Hum Mol Genet* 2000, **9**:653-657.

See annotation [24•].

26. Long M, Rosenberg C, Gilbert W: **Intron phase correlations and the evolution of the intron/exon structure of genes.** *Proc Natl Acad Sci USA* 1995, **92**:12495-12499.

27. Long M, de Souza SJ, Gilbert W: **Evolution of the intron–exon structure of eukaryotic genes.** *Curr Opin Genet Dev* 1995, **5**:774-778.

28. Dibb NJ, Newman AJ: **Evidence that introns arose at proto-splice sites.** *EMBO J* 1989, **8**:2015-2021.

29. Dibb NJ: **Proto-splice site model of intron origin.** *J Theor Biol* 1991, **151**:405-416.

30. Logsdon JM Jr: **The recent origins of spliceosomal introns revisited.** *Curr Opin Genet Dev* 1998, **8**:637-648.

31. Logsdon JM Jr, Stoltzfus A, Doolittle WF: **Molecular evolution: recent cases of spliceosomal intron gain?** *Curr Biol* 1998, **8**:R560-R563.

32. Long M, de Souza SJ, Rosenberg C, Gilbert W: **Relationship between 'proto-splice sites' and intron phases: evidence from dicodon analysis.** *Proc Natl Acad Sci USA* 1998, **95**:219-223.

33. •• Long M, Rosenberg C: **Testing the 'proto-splice sites' model of intron origin: evidence from analysis of intron phase correlations.** *Mol Biol Evol* 2000, **17**:1789-1796.

This statistical analysis for the relationship between intron phases and various hypothetical proto-splice sites [32] is a companion work to this one. Using the large database of GenBank in exhaustive tests, these analyses rejected the model of proto-splice sites as insertion sites of introns and further confirmed the underlying mechanism of non-random distribution of intron phases – exon shuffling.

34. Tiedge H, Chen W, Brosius J: **Primary structure, neural-specific expression, and dendritic location of human BC200 RNA.** *J Neurosci* 1993, **13**:2382-2390.

35. Brosius J: **RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements.** *Gene* 1999, **238**:115-134.

36. Rogers JH: **The origin and evolution of retroposons.** *Int Rev Cytol* 1985, **93**:187-279.

37. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.

38. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.

39. •• Courseaux A, Nahon JL: **Birth of two chimeric genes in the Hominidae lineage.** *Science* 2001, **291**:1293-1297.

This report, together with [40], describes two human genes that originated 5 and 20 million years ago, respectively. A novel molecular process that created these chimerical genes was reported: the initial gene structure evolved by retroposition from antisense MCH messenger RNA with newly recruited coding regions and splice sites.

40. Viale A, Courseaux A, Presse F, Ortola C, Breton C, Jordan D, Nahon JL: **Structure and expression of the variant melanin-concentrating hormone genes: only PMCHL1 is transcribed in the developing human brain and encodes a putative protein.** *Mol Biol Evol* 2000, **17**:1626-1640.

41. Viale A, Ortola C, Richard F, Vernier P, Presse F, Schilling S, Dutrillaux B, Nahon JL: **Emergence of a brain-expressed variant melanin-concentrating hormone gene during higher primate evolution: a gene 'in search of a function'.** *Mol Biol Evol* 1998, **15**:196-214.

42. Ohno S: *Evolution by Gene Duplication*. Berlin: Springer-Verlag; 1970.

43. Walsh JB: **How often do duplicated genes evolve new functions?** *Genetics* 1995, **139**:421-428.

44. Ohta T: **Simulating evolution by gene duplication.** *Genetics* 1987, **115**:207-213.

45. Gu X: **Statistical methods for testing functional divergence after gene duplication.** *Mol Biol Evol* 1999, **16**:1664-1674.

46. •• Malik HS, Henikoff S: **Adaptive evolution of cid, a centromere specific histone in *Drosophila*.** *Genetics* 2001, **157**:1293-1298.

This paper provides a clear example of how new centromeric DNA-binding function in the *cid* gene arose as a consequence of positive Darwinian selection. DNA sequence variation between and within species in the *Cid* gene was analyzed in *D. melanogaster* and its close siblings Remarkably, a rapid amino acid substitution was identified in the protein region that likely mediates binding to centromeric DNA, a fast-moving satellite repeat region. This investigation provides an explicit example of how function of a duplicate member evolves as a response to selelction.

47. Brown CJ, Todd KM, Rosenzweig RF: **Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment.** *Mol Biol Evol* 1998, **15**:931-942.

48. Hall BG: **Evolution of new metabolic functions in laboratory organisms.** In *Evolution of Genes and Proteins*. Edited by Nei M, Koehn K. Sunderland, MA: Sinauer Associates; 1983:234-257.

49. Hartley BS: **Experimental evolution of ribotol dehydrogenase.** In *Microorganisms as Model Systems for Studying Evolution*. Edited by Mortlock RP. New York: Plenum Press; 1984.

50. Makalowski W, Boguski MS: **Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences.** *Proc Natl Acad Sci USA* 1998, **95**:9407-9412.

51. • Makalowski W: **Genomic scrap yard: how genomes utilize all that junk.** *Gene* 2000, **259**:61-67.
In this report, an interesting survey of the TEs contributing to protein diversity (i.e. the portion of TEs recruited into proteins of nuclear genes) is reviewed. Makalowski's useful website (http://www.ncbi.nlm.nih.gov/Makalowski/ScrapYard) reports >200 cases that were identified recently.

52. Caras IW, Davitz MA, Rhee L, Weddell G, Martin DW Jr, Nussenzweig V: **Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins.** *Nature* 1987, **325**:545-549.

53. Lawrence JG, Lawrence JG: **Gene transfer, speciation, and the evolution of bacterial genomes.** *Curr Opin Microbiol* 1999, **2**:519-523.

54. Groisman EA, Ochman H, Groisman EA, Ochman H: **Pathogenicity islands: bacterial evolution in quantum leaps.** *Cell* 1996, **87**:791-794.

55. • de Koning AP, Brinkman FS, Jones SJ, Keeling PJ: **Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*.** *Mol Biol Evol* 2000, **17**:1769-1773.
The authors report an interesting case of the origin of a novel gene with new function by a bacteria→eukaryote gene transfer. A likely functional change brought by the structural evolution of the created gene in the protozoan organism *T. vaginalis* was analyzed.

56. Nugent JM, Palmer JD: **RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution.** *Cell* 1991, **66**:473-481.

57. Jacob F: **Evolution and tinkering.** *Science* 1977, **196**:1161-1166.

58. Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL: **Selective sweep of a newly evolved sperm-specific gene in *Drosophila*.** *Nature* 1998, **396**:572-575.

59. Long M, Langley CH: **Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*.** *Science* 1993, **260**:91-95.

60. Long M, Wang W, Zhang J: **Origin of new genes and source for N-terminal domain of the chimerical gene, *jingwei*, in *Drosophila*.** *Gene* 1999, **238**:135-141.

61. •• Wang W, Zhang J, Alvarez C, Llopart A, Long M: **The origin of the *Jingwei* gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*.** *Mol Biol Evol* 2000, **17**:1294-1301.
A report of an ancestral gene of a newly evolved chimerical gene, *Jingwei*, in *Drosophila*. Indicative of a complex organization, the structure of *Yellow emperor* comprises a nested gene, *Musashi*, different isoforms from alternative splicing, and a module for exon shuffling. This complex modular gene structure was not predicted by the genome annotation in *D. melanogaster*, representing a serious hurdle to gene discovery in computational analyses of genome sequence.

62. Charlesworth B, Charlesworth D: **How was the *Sdic* gene fixed?** *Nature* 1999, **400**:519-520.

63. Nurminsky DI, Hartl DL: **How was the *Sdic* gene fixed? Reply.** *Nature* 1999, **400**:520.

64. •• Nurminsky D, Aguiar DD, Bustamante CD, Hartl DL: **Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*.** *Science* 2001, **291**:128-130.
To test the chromosomal effect of the rapid evolution in *Sdic*, polymorphisms in 10 genes adjacent to *Sdic* in the X chromosome in *D. melanogaster* and their 10 homologues in *D. simulans* that do not contain this gene were compared – across the chromosome and between the two species. The contrast patterns of polymorphisms between the species, with *D. melanogaster* having a 'trough' in the region including *Sdic*, were presented as evidence for the positive selection of this gene.

65. Begun DJ: **Origin and evolution of a new gene descended from alcohol dehydrogenase in *Drosophila*.** *Genetics* 1997, **145**:375-382.

66. Ohta T: **Further examples of evolution by gene duplication revealed through DNA sequence comparisons.** *Genetics* 1994, **138**:1331-1337.

67. Wistow G: **Lens crystallins: gene recruitment and evolutionary dynamism.** *Trends Biochem Sci* 1993, **18**:301-306.

68. Piatigorsky J, Horwitz J: **Characterization and enzyme activity of argininosuccinate lyase/delta-crystallin of the embryonic duck lens.** *Biochim Biophys Acta* 1996, **1295**:158-164.

69. Chen L, DeVries AL, Cheng CH: **Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish.** *Proc Natl Acad Sci USA* 1997, **94**:3811-3816.

70. Chen L, DeVries AL, Cheng CH: **Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod.** *Proc Natl Acad Sci USA* 1997, **94**:3817-3822.

71. Cheng CH, Chen L: **Evolution of an antifreeze glycoprotein.** *Nature* 1999, **401**:443-444.

72. Logsdon JM Jr, Doolittle WF: **Origin of antifreeze protein genes: a cool tale in molecular evolution.** *Proc Natl Acad Sci USA* 1997, **94**:3485-3487.

73. •• Thomson TM, Lozano JJ, Loukili N, Carrió R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M *et al.*: **Fusion of the human gene for the polyubiquitination coeffector UEV1 with *Kua*, a newly identified gene.** *Genome Res* 2000, **10**:1743-1756.
The authors of this study report a clear case that a newly evolved chimerical protein in humans evolved by gene fusion; the new proteins are located in novel intracellular positions, suggesting the development of an original protein function.

74. Long M: **A new function evolved from gene fusion.** *Genome Res* 2000, **10**:1655-1657.

75. Nekrutenko A, Li WH: **Transposable elements are found in a large number of human protein-coding genes.** *Trends Genet* 2001, **17**:619-621.