THE ORIGIN OF NEW GENES: GLIMPSES FROM THE YOUNG AND OLD

Manyuan Long[*,‡], *Esther Betrán*[§], *Kevin Thornton*[‡] *and Wen Wang*[||]

Genome data have revealed great variation in the numbers of genes in different organisms, which indicates that there is a fundamental process of genome evolution: the origin of new genes. However, there has been little opportunity to explore how genes with new functions originate and evolve. The study of ancient genes has highlighted the antiquity and general importance of some mechanisms of gene origination, and recent observations of young genes at early stages in their evolution have unveiled unexpected molecular and evolutionary processes.

NEW GENE
A gene that has originated recently in the relevant evolutionary timescale.

ECTOPIC RECOMBINATION
Recombination between nonhomologous sequences.

*[*]Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA.
[‡]Committee on Genetics, The University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA.
[§]Biology Department, University of Texas, Arlington, Texas 76019, USA.
[||]Chinese Academy of Sciences (CAS)–Max Planck Junior Scientist Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, CAS, Kunming 650223, China.
Correspondence to M.L.
e-mail: mlong@uchicago.edu*

doi:10.1038/nrg1204

Although interest in evolutionary novelties can be traced back to the time of Darwin, studies of the origin and evolution of genes with new functions have only recently become possible and attracted increasing attention. The available molecular techniques and rapidly expanded genome data from many organisms mean that searching for and characterizing NEW GENES is no longer a formidable technical challenge. Also, molecular evolution and molecular population genetics have provided useful analytical tools for the detection of the processes and mechanisms that underlie the origin of new genes.

The variation in gene number among organisms indicates that there is a general process of new gene origination. Two levels of questions about this process can be defined. First, at the level of individual new genes, what are the initial molecular mechanisms that generate new gene structures? Once a new gene arises in an individual genome in a natural population, how does it spread throughout an entire species to become fixed? And, how does the young gene subsequently evolve? Second, at the level of the genome, how often do new genes originate? If new gene formation is not a rare event, are there patterns that underlie the process? And, what evolutionary and genetic mechanisms govern any such patterns?

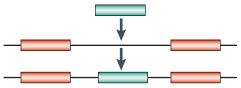
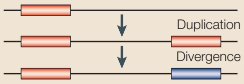
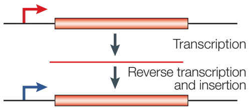
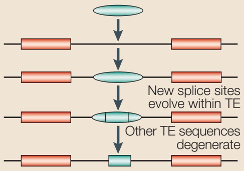
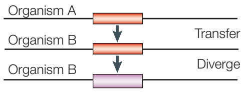
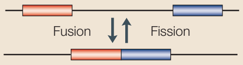
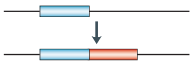
Previous efforts to study the origin of new genes have been sporadic and have focused on the evolution of duplicate and chimeric genes, even though these are often hundreds of millions of years old. Analyses of genes that have been identified by such an approach have provided some exciting insights. However, a more efficient approach is the direct observation of young genes when they are at an early stage in their evolution. This is advantageous because many details of the origin of a gene are lost over longer periods of time. Also, the study of ancient genes has established the antiquity of some of the molecular mechanisms used to generate new genes. This review discusses insights into the creation and evolution of new genes. We introduce the processes by which individual genes originate, and analyse the rate and pattern of new gene evolution at the genomic level.

Sources of new genes

Several molecular mechanisms are known to be involved in the creation of new gene structures (for a summary, see TABLE 1), the details of which are understood to varying degrees.

Exon shuffling. Two or more exons from different genes can be brought together ectopically, or the same exon can be duplicated, to create a new exon–intron structure[1]. Two mechanisms are known to lead to the ECTOPIC RECOMBINATION of exons: ILLEGITIMATE RECOMBINATION[2,3] and

Table 1 | **Molecular mechanisms for creating new gene structures**

Mechanism	Process	Examples	Comments	References
Exon shuffling: ectopic recombination of exons and domains from distinct genes		*fucosyltransferase*, *jingwei*, *Tre2*	~19% of exons in eukaryotic genes have been formed by exon shuffling	8,32,40,62, 65–68,105
Gene duplication: classic model of duplication with divergence		*CGβ*, *Cid* , *RNASE1B*	Many duplicates have probably evolved new functions	9–11,29,35,39, 47,48,106
Retroposition: new gene duplicates are created in new genomic positions by reverse transcription or other processes		*PGAM3*, *Pgk2*, *PMCHL1*, *PMCHL2*, *Sphinx*	1% of human DNA is retroposed to new genomic locations	23,43,61,76, 80–82,107–110
Mobile element: a mobile element, also known as a transposable element (TE), sequence is directly recruited by host genes	 New splice sites evolve within TE Other TE sequences degenerate	*HLA-DR-1*, human *DAF*, *lungerkine* mRNA, *mNSC1* mRNA	Generates 4% of new exons in human protein-coding genes	16,78,111,112
Lateral gene transfer: a gene is laterally (horizontally) transmitted among organisms	 Organism A Organism B Organism B Transfer Diverge	*acytylneuraminate lysase, Escherichia coli mutU* and *mutS*	Most often reported in prokaryotes and recently reported in plants	18–20,113
Gene fusion/fission: two adjacent genes fuse into a single gene, or a single gene splits into two genes		Fatty-acid synthesis enzymes, *Kua-UEV, Sdic*	Involved in the formation of ~0.5% of prokaryotic genes	21,22,42, 114,115
De novo origination: a coding region originates from a previously non-coding genomic region		*AFGPs*, *BC1RNA*, *BC200RNA*	Rare for whole gene origination; might not be rare for partial gene origination	52–53,116,117

AFGP, antifreeze glycoprotein; *CGβ*, chorionic gonadotropin β polypeptide; *Cid*, centromere identifier; *DAF,* decay-accelerating factor; *HLA-DR-1*, major histocompatibility complex DR1; *PGAM3*, phosphoglycerate mutase 3; *Pgk2*, phosphoglycerate kinase 2; *PMCHL*, pro-melanin-concentrating hormone-like; *RNASE*, ribonuclease; *Sdic*, sperm-specific dynein intermediate chain; *UEV*, tumour susceptibility gene.

retroposed exon insertions[4,5]. There is genomic evidence that exon shuffling, which is also known as domain shuffling, often recombines sequences that encode various protein domains to create mosaic proteins[6,7]. Using direct sequence comparison, Patthy[8] identified numerous genes that were created by exon shuffling.

Gene duplication. This classical model creates a duplicate gene that can evolve new functions, whereas the ancestral copy maintains its original functions[9,10] (TABLE 1). Many new gene functions have evolved through gene duplication and it has contributed tremendously to the evolution of developmental programmes in various organisms (for a recent review, see REF. 11). Also, duplications at the segmental chromosomal and genome levels, which are abundant in plants[9], have been shown to contribute to the evolution of new functions in humans[12,13].

Retroposition. This mechanism creates duplicate genes in new genomic positions through the reverse transcription of expressed parental genes (TABLE 1). As a retroposed gene copy does not usually retropose a promoter copy from its parental gene, it has to recruit a new regulatory sequence to be functional or it will die out as a processed pseudogene. So, a functional retroposed gene has a chimeric structure — either a retroposed coding region with a new regulatory sequence or a retroposed coding region with a new protein fragment that is recruited from the targeted site — that leads to it having a different function to its parental gene. In mammals, the L1 RETRO-ELEMENT is responsible for retroposing nuclear genes[4,5].

Mobile elements. Makalowski *et al.*[14] were the first to describe the integration of an ALU ELEMENT into the coding portion of the human decay-accelerating factor (*DAF*) gene. They found that mobile element-derived diversity was not limited to the human genome or to the Alu family[15] (TABLE 1). Further analyses of human genome sequences[16] and vertebrate genes[17] have shown that the integration of MOBILE ELEMENTS into nuclear genes to generate new functions is a general phenomenon.

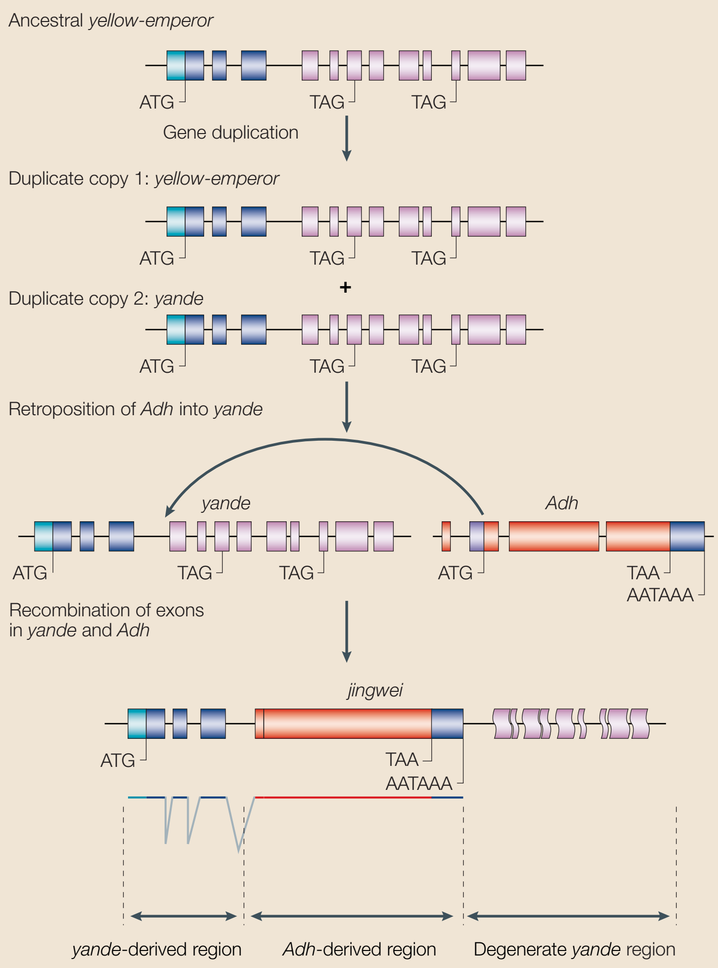
Lateral gene transfer. In prokaryotes, genes are often transferred between organisms (TABLE 1). Although this lateral, or horizontal, gene transfer can lead to the

ILLEGITIMATE RECOMBINATION
Nonhomologous sequence recombination at the genomic DNA level.

L1 RETRO-ELEMENT
A member of the long interspersed transposable element (LINE) family, which is a type of large repetitive DNA sequence that inserts itself throughout the genome through retroposition. L1 retro-elements are ~6,400 base pairs long and are abundant in the human genome.

ALU ELEMENT
An interspersed DNA sequence of ~300 base pairs (bp) that is found in the genomes of primates, which can be cleaved by the restriction enzyme *Alu*I. They are composed of a head-to-tail dimer, with the first monomer ~140-bp long and the second ~170-bp long. In humans, there are 300,000–600,000 copies of Alu elements.

Box 1 | The origin of *jingwei*


Ancestral *yellow-emperor*
ATG
TAG
TAG
Gene duplication
Duplicate copy 1: *yellow-emperor*
ATG
TAG
TAG
+
Duplicate copy 2: *yande*
ATG
TAG
TAG
Retroposition of *Adh* into *yande*
yande
Adh
ATG
TAG
TAG
ATG
TAA
AATAAA
Recombination of exons in *yande* and *Adh*
jingwei
ATG
TAA
AATAAA
yande-derived region
Adh-derived region
Degenerate *yande* region

In the early 1990s, the first young gene to be described was *jingwei* in a group of African *Drosophila* species[32]. It provided enough details for the molecular mechanism underlying its origination to be deduced. A portion of *jingwei* was found to be a homologue of the *Adh* gene that encodes alcohol dehydrogenase[98] and was later characterized as a retrosequence of *Adh*[99]. Further population genetic, molecular biological and comparative phylogenetic analyses showed it to be a new processed functional gene that originated around 2 million years ago in the common ancestor of two African *Drosophila* species, *Drosophila yakuba* and *Drosophila teissieri*.

In the ancestral species, there were two single copy genes, *yellow-emperor* (*ymp*) and *Adh*. *yellow-emperor* was duplicated into two copies: one also called *yellow-emperor* and the other called *yande* (*ynd*)[100,101]. Whereas *yellow-emperor* maintained its original functions, *yande* was further involved in the origin of *jingwei*. In the short time before the speciation event, *Adh* mRNA retroposed into the third intron of *yande* as a fused exon and recombined with the first three *yande* exons. This formed *jingwei*, which is a gene that is translated into a chimeric protein.

In the figure, the splicing pattern of *jingwei* is shown underneath its structure: the blue regions are from the first three *yande* exons and the red regions are from the *Adh*-derived exon. The insertion of the *Adh*-retroposed sequence rendered the nine downstream *yande* exons (shown in pink) degenerate. The boxes in the figure represent exons, and TAG and TAA are stop codons. For simplicity, the RNA stage of *Adh* retroposition is not shown.

This process provided the answers to several related problems. The creation of a new gene does not destroy previous functions. First, although *yande* recombined with *Adh*, the previously encoded functions were maintained by *yellow-emperor*. Second, chimeric structures can easily provide new protein diversity. This is the principle of exon shuffling[1]: new genes can be readily created by recombining previously existing domains or modules, which differs from the classic process of gene duplication followed by successive substitutions. The first three *yande* exons that were recruited by *jingwei* seem to form an essential subunit for the functions of two distinct isoforms of testis-specific protein[100]. Third, the inserted *Adh* exon does not passively use the regulatory signals of the recipient gene; this means that the associated transcription terminating signals (AATAAA) of *Adh* terminate the readthrough transcription and downstream *yande* exons degenerate. These degenerate exons could not have been observed if the origination event was not recent, as *Drosophila* pseudogenes have a short half-life (<12 million years)[102]. Fourth, *jingwei* clearly inherited the regulatory sequence of *yande* from the previous duplication of *yellow-emperor*, as *D. teissieri jingwei* has the same testis-specific expression as its parental gene *yellow-emperor*[100]. This seems to be the way in which the promoter-lacking *Adh* retrosequence was rescued from the fate of many retrosequences as processed pseudogenes[103,104]. Fifth, and finally, the protein sequence encoded by *jingwei* evolved rapidly both before and after the divergence of *D. teissieri* and *D. yakuba*, which points to an ADAPTIVE EVOLUTIONARY process that was driven by positive Darwinian selection.

exchange of homologous genes, there is evidence that it can also recruit new genes and provide new phenotypes; for example, converting benign bacteria into pathogens[18]. Recently, lateral gene transfer was observed in the protozoan *Trichomonas vaginalis*[19] and, surprisingly, in flowering plants, in which five such transfer events were observed that involved mitochondrial genes[20]. This indicates that lateral gene transfer might be important in the evolution of eukaryotic genes. Also, the horizontal transfer of transposable elements has been observed, which might contribute or carry linked genes or gene fragments during the formation of new genes in recipient organisms.

Gene fusion/fission. Two adjacent genes can fuse into a single gene by readthrough transcription, through the deletion or mutation of the translation stop codon and the transcription termination signal in the upstream gene (TABLE 1).Conversely, a single gene can be split into two separate genes, although the underlying mechanism by which such fission occurs remains unclear. Many cases of gene fusion and fission have been identified in prokaryotic genomes and gene fusions have also been reported in higher eukaryotes. For example, Thompson *et al.*[21] identified a human fusion gene, *KUA-UEV*, in which the ubiquitin E2 variant domain of tumour susceptibility gene (*UEV*) and a newly identified gene known as *KUA* were fused together by

MOBILE ELEMENTS
Also known as transposable elements. DNA sequences in the genome that replicate and insert themselves into various positions in the genome.

readthrough transcription and alternative splicing of their coding sequences.

***De novo* origination.** Although the true *de novo* origination of new genes from previously non-coding sequences is rare, there are genes with a portion of coding-region sequence that has originated *de novo*. For example, in the *Drosophila* sperm-specific dynein intermediate chain gene *Sdic*, a previously intronic sequence has been converted into a coding exon[22].

Combined mechanisms. New genes can be created by the mechanisms discussed above, either individually or in combination, as in the case of *jingwei*, which was the first young gene to be described (BOX 1,2). The origin of *jingwei* has highlighted the creative roles of several molecular processes acting in combination: exon shuffling, retroposition and gene duplication. In *jingwei*, all of these molecular mechanisms were identified by sequence and functional comparison, taking advantage of the similarity between *jingwei* and its parental genes.

This is not the only young gene system that has been directly observed. Other examples are the *sphinx* gene[23,24] and the *Sdic* gene[22], which are both present in the single *Drosophila* species *Drosophila melanogaster,* and so are younger than the divergence time between *D. melanogaster* and its sibling species, no more than 3 million years ago. Like *jingwei*, *sphinx* was also created by retroposition, in this case from the *ATP synthase F* gene, which recruited nearby intron and exon sequences to form a standard chimeric structure. *sphinx* also evolved rapidly (FIG. 1). Besides exon shuffling and retroposition, a mobile element (S element) participated in the creation of a new splice site and coding region of the *sphinx* gene. Further examples of new genes in *Drosophila* and other organisms can be found in TABLE 2.

Evolution of new genes

The fate of a newborn gene. The classical view of the fate of gene duplications dates back to the work of J. B. S. Haldane[25] and R. A. Fisher[26]. They believed that, in the presence of recurrent mutation, one member of a duplicate pair eventually becomes nonfunctional; that is, most duplicates should eventually die out as pseudogenes. However, a recent model of SUBFUNCTIONALIZATION[27,28] provided one probable interpretation of the fact that there are many duplicate genes in the coding genomes of higher eukaryotes (for a review, see REF. 11), although NEOFUNCTIONALIZATION is also important.

Kimura[10] extended the classical model in an effort to describe how gene duplicates could acquire new functions and ultimately be preserved in a lineage. In this model, after gene duplication, PURIFYING SELECTION against deleterious mutations is relaxed on one or both copies of the gene; this is attributable to the extra sheltering of recessive mutations when there are extra copies of a locus. Consequently, mutations that would normally be eliminated by selection could accumulate at one or both loci. In most cases, the fixation of null alleles results in the loss of one copy; however, under certain conditions, such as environmental change, some of the mutant alleles that encode a new function could become beneficial and therefore be preserved by natural selection. Kimura's model is pertinent because it is based on an obvious intuition that new gene duplicates might provide redundant functions in their early stages and, therefore, the mutations in the duplicates are neutral or nearly neutral — we call this the 'waiting model', because it requires time to reach the moment of functional distinction.

A more straightforward conjecture is that adaptive evolution might have had a principal role throughout the creation and subsequent evolution of new genes — we call this the 'immediate model', because it requires no waiting time for the evolution of a new function. Several case studies and theoretical works (for example, see REFS 29,30) have shown that the evolution of recently created genes involves accelerated changes in both protein-coding sequences and gene structures from the onset, which supports the immediate model. An important role of positive Darwinian selection has been detected in these processes and these studies have uncovered some interesting results. For example, the initial functions of new genes are rudimentary and further improvement under selection might be crucial. So, new gene functions that are created by altering a sequence that encodes one or a few amino acids might be special cases rather than the general situation. Also, the rapid changes in well-defined new genes with new functions could help to explain a past

SUBFUNCTIONALIZATION
The process of 'partitioning' the ancestral functions of a locus among its duplicates. For example, if a single-copy gene that is normally expressed in two tissues subsequently duplicates, and each duplicate is then expressed in a different tissue, subfunctionalization has occurred.

NEOFUNCTIONALIZATION
The evolution of a new function by a duplicate gene.

PURIFYING SELECTION
Selection against deleterious alleles.

ADAPTIVE EVOLUTION
An evolutionary process that is directed by natural selection, which makes a population better adapted to live in an environment.

Box 2 | The names of genes and two ancient legends

jingwei. In an ancient Chinese legend, the first Chinese emperor Yande (3,000 BC) brother to the Yellow-Emperor, had a pretty princess named Jingwei. Like other legendary southern Chinese goddesses, Jingwei liked to swim. Unfortunately, she drowned in the East China Sea. She was reincarnated into a beautiful bird, who, to save others from possible tragedy, carried soil and stones in an attempt to fill in the ocean. The new gene was named *jingwei* because it was first thought to be a pseudogene and was then 'reincarnated' as a new functional gene with a new structure. To be consistent, other related genes were named following this legend.

sphinx. According to ancient Greek legend, the sphinx was a creature with a human head, the body of a lion and the wings of an angel, which loved to ask riddles to those who dared to guess an answer. Unfortunately, those brave but unlucky people who failed to answer the riddles correctly often became part of the sphinx's diet. The extremely young *sphinx* gene was so-called because it has a chimeric exon–intron structure with the participation of a third mobile-element component, an S element, and it also presented a riddle with its unusual functionality. Fortunately, the investigators who took on the challenge of investigating the formation of *sphinx* had the luck of Oedipus, who strangled the sphinx.

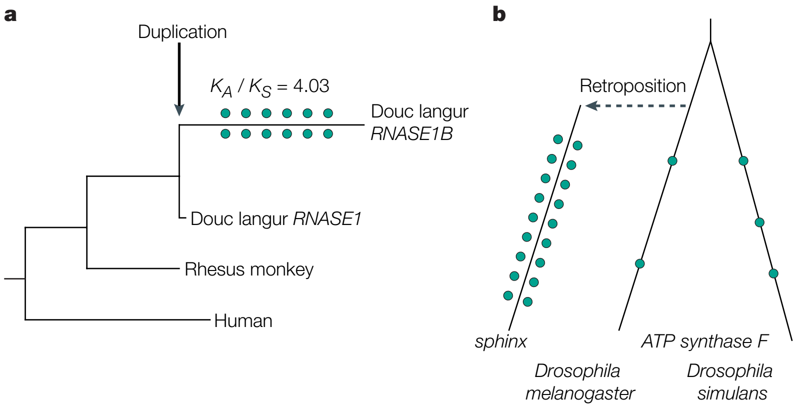


Figure 1 | **Two examples of the accelerated evolution of new genes with new functions.** **a** | The ribonuclease *RNASE1B* gene in the leaf-eating colobine monkey *Pygathrix nemaeus* (the douc langur)[35]. The green circles indicate nucleotide substitutions in the coding regions of the genes. In the *RNASE1B* lineage, 12 sites have been substituted, most of which are replacement changes; however, its sibling copy *RNASE1* has no sequence changes. **b** | The *sphinx* RNA gene in *Drosophila melanogaster*[23]. Similar to the protein-coding gene shown in part **a**, the RNA gene *sphinx* has also undergone accelerated substitution since its formation: there are 18 substitutions in *sphinx* versus 2 substitutions in the parental gene *ATP synthase F*.

conjecture in molecular evolution studies: that rapid sequence evolution in many old genes might reflect a diverged function under selection[31].

Accelerated evolution of new genes. New genes usually evolve with rapid changes in their sequence, structure and expression (for example, see FIG. 1). In its early stage, the retroposed protein-coding sequence of *jingwei* evolved at least an order of magnitude faster than normal proteins[32]. Ribonuclease and lysozyme that are specific to primates (see TABLE 2) also have significant changes in their protein sequence, which are associated with the rapid emergence of new biological functions that are driven by positive Darwinian selection in several primate species[33–35]. Ohta[36] observed that several new gene duplicates in the human genome that evolved new expression patterns to accommodate new biological functions all evolved more rapidly in their protein-coding sequences than in those of the orthologous copies of parental genes. Recent studies[23,37–39] showed that rapid evolution of new genes is a general phenomenon. In some cases, the changes take place in a sequence that encodes a newly acquired peptide that is important for the development of a new function; for example, the formation of the TBC domain of the *TRE2* (*USP6*) hominoid-specific gene[40] and the mitochondrial target domain of cytochrome c1 in plants[41]. The *Sdic* gene in *D. melanogaster*[22,42] is a notable example of a rapid change in gene structure (that is, a selective sweep). In this case, the two halves of the *Sdic* gene fused together from two parental genes: the intron of one parental gene was transformed into exon sequence, a former exon sequence was changed into promoter and the regulatory sequences acquired a new function in sperm tails. All of these structural innovations took place in the short evolutionary period after *D. melanogaster* split from its siblings. A new expression pattern also rapidly emerged for *jingwei* and the recently evolved primate processed gene *PGAM3*, which was retroposed from the phosphoglycerate mutase gene[43].

Evolutionary forces that drive changes in new genes. One implication of the frequent observations of rapid change in new genes is that initial gene functions are not sufficient, and continuous changes in sequence and structure are vital if a further diverged function is to be established. So, many changes in the new genes seem to be needed for adequate function, and only one or a few changes leading to new functions might be the exception. However, one problem remains: what evolutionary forces drive these changes?

Positive Darwinian selection could be an important force driving the evolution of new genes. Early speculations were that new genes might be an outcome of adaptive evolution and therefore ought to be subject to positive selection[9,10,29,36]. Experimental evolution in microbial organisms has shown that selection can rapidly generate new gene functions[44,45]. For example, after 450 generations of glucose-limited growth, *Saccharomyces cerevisiae* strains evolve three new chimeric genes from the hexose transport loci *HXT6* and *HXT7*; these chimaeras increase the ability of *S. cerevisiae* to scavenge glucose at low substrate concentrations[44]. Several lines of analyses in new genes have shown the role of positive natural selection.

The first test for selection was to analyse the K_A/K_S ratio in new gene lineages. Zhang *et al.*[35] showed that a new duplicate ribonuclease gene *RNASE1B*, which arose 4.2 million years ago in the leaf-eating colobine monkey *Pygathrix nemaeus* (the douc langur), has a K_A/K_S ratio that is significantly higher than unity (0.0310/0.0077 = 4.03), whereas its paralogous copy *RNASE1* has no evolutionary changes (FIG. 1), which indicates that there is strong Darwinian selection on the new gene. Furthermore, it was consistently found that seven out of nine amino-acid substitutions in the new duplicate result in charge changes and increase the negative charge of the protein, which predicts that there is an altered optimal pH for RNASE1B. To test this, the authors conducted a protein-overexpression experiment and biochemical analysis. They confirmed that RNASE1B has a lower optimal pH value, which is correlated with a digestive system that has evolved for leaf-eating and foregut fermentation in the colobine monkey.

Second, a population genetic analysis of molecular sequences that compares two levels of variation — between-species divergence and within-species polymorphism — using the neutral theory of molecular evolution[10], which assumes that molecular variation in these levels is free from natural selection, has provided a sensitive probe to detect selection. *Drosophila jingwei* is subject to strong selective constraint, most of its polymorphism is synonymous and it has a low proportion of replacement changes. By contrast, most of the between-species substitutions were replacement substitutions. A McDonald-Kreitman test[46] showed excess amino-acid replacement substitutions in comparison with the neutrality prediction that the two levels of variation should be positively correlated[10]; this indicates that

REPLACEMENT CHANGES
(Substitutions). Changes in the nucleotide sequences of coding genes that result in changes in the peptide sequence (that is, the replacement of an amino acid). These contrast with silent (or synonymous) changes in coding sequences, which do not result in changes in the peptide.

K_A/K_S
K_A is the rate of substitution at non-synonymous sites and K_S is the rate of substitution at synonymous sites. The ratio between the two (K_A/K_S) is often used to infer selection: a K_A/K_S that is <1 indicates a functional constraint; a K_A/K_S that is equal to 1 indicates a lack of functional constraint; and a K_A/K_S that is >1 indicates positive Darwinian selection.

POPULATION GENETIC ANALYSIS
The process of making inferences about the evolutionary and demographic history of a gene (or organism) on the basis of data on genetic variation in a species.

SELECTIVE CONSTRAINT
A limit on evolutionary change.

MCDONALD-KREITMAN TEST
A statistical test that is commonly used for the comparison of between-species divergence and within-species polymorphism at replacement and synonymous sites to infer adaptive protein evolution.

Table 2 | **Examples of new genes of known age**

Genes	Age	Evolutionary features	References
Drosophila			
jingwei	2.5 my	A standard chimeric structure with rapid sequence evolution	32
Sdic	<3 my	Rapid structural evolution for a specific function in sperm tails	22
sphinx	<3 my	A non-coding RNA gene that rapidly evolved new splice sites and sequence	23
Cid	Function diverged in the past 3 my	Co-evolved with centromeres under positive Darwinian selection	48
Dntf-2r	3–12 my	Origin of new late testis promoter for its male-specific functions	49
Adh-Finnegan	30 my	Recruited a peptide from an unknown souce and evolved at a faster rate than its parent gene	38
Primates			
FOXP2	100,000 y	A selective sweep in this gene, which has language and speech function, took place recently	118,119
RNASE1B	4 my	Positive seletion detected, which corresponds with new biological traits in leaf-eating monkeys	35
PMCHL2	5 my	Expression is specifically and differentially regulated in testis	107
PMCHL1	20 my	A new exon–intron in the 3′ coding region created *de novo* and an intron-containing gene structure created by retroposition	107
Morpheus	12–25 my	Strong positive selection in human–chimpanzee lineages	37
TRE2	21–33 my	A hominoid-specific chimeric gene with testis-specific expression	40
FUT3/FUT6	35 my	New regulatory untranslated exons created *de novo* in new gene copies; the family has been shaped by exon shuffling, transposation, point mutations and duplications	105
CGβ	34–50 my	One of two subunits of placentally expressed hormone; the rich biological data clearly detail its function	39
BC200	35–55 my	A non-coding RNA gene that is expressed in nerve cells	110,120
Rodents			
4.5Si RNA	25–55 my	A non-coding RNA gene that is expressed ubiquitously	121
BC1 RNA	60–110 my	A neural RNA that originated from an unusual source: tRNA^{Ala}	110,120
Fish			
Arctic *AFGP*	2.5 my	Convergent evolution; antifreeze protein created from an unexpected source driven by the freezing environment	52,53
Antarctic *AFGP*	5–14 my	Convergent evolution; antifreeze protein created from an unexpected source driven by the freezing environment	52,53
Plants			
Sanguinaria rps1	<45 my	A chimeric gene structure created by lateral gene transfer	20
Cytochrome c1	110 my	Origin of mitocondrial-targeting function by exon shuffling	41
Protozoa			
N-acetylneuraminate lyase	<<15 my	A laterally transferred gene from proteobacteria that recruited a signal peptide	19

AFGP, antifreeze glycoprotein; *CGβ*, chorionic gonadotropin β polypeptide; *Cid*, centromere identifier; *Dntf-2r*, *Drosophila* nuclear transport factor-2-related; *FOXP2*, forkhead box P2; *FUT3*, fucosyltransferase 3; my, million years; *PMCHL*, pro-melanin-concentrating hormone-like; *RNASE*, ribonuclease; *rps1*, ribosomal protein small subunit 1; *Sdic*, sperm-specific dynein intermediate chain; y, years.

there has been strong selection throughout its evolution[32]. A similar approach has detected Darwinian positive selection in several new *Drosophila* genes[40,47–50].

A third line of evidence stems from the findings of convergent evolution. A remarkable example is the acquisition of antifreeze glycoproteins (AFGPs) in Antarctic notothenioid fish and northern cod[51–53]. AFGPs in both types of fish are polymers of a simple glycotripeptide monomer (Thr-Ala-Ala), which binds ice crystals that form in fish tissues in freezing environments. Although the two kinds of AFGPs have nearly identical sequences, they have different ancestors. The Antarctic *AFGP* gene recruited a signal peptide region and a 3′ untranslated region (UTR) from an ancestral trypsinogen gene, and amplified a tripeptide (Thr-Ala-Ala)-coding region that occurred in the same gene. It is likely that the Antarctic *AFGP* evolved from nine nucleotides spanning the splice junction between intron 1 and exon 2 of trypsinogen, which translate into Thr-Ala-Ala, and subsequently expanded through replication slippage or unequal crossing over. It acquired independence from the ancestral gene by shedding unrelated

CONVERGENT EVOLUTION
Independent evolution from different ancestors that leads to similar characteristics.

NOTOTHENIOID
The most abundant group of Antarctic fish.

TRYPSINOGEN
A large diverse protein family of serine peptidases.

Table 3 | **Intron-phase correlation in eukaryotic genomes**

	Symmetrical			Asymmetrical					
	(0,0)	**(1,1)**	**(2,2)**	**(0,1)**	**(0,2)**	**(1,2)**	**(1,0)**	**(2,0)**	**(2,1)**
Observed number*	3,051	1,303	620	1,321	1,184	749	1,408	1,219	704
Expected number‡	2,709	1,013	558	1,657	1,230	752	1,657	1,229	752

*The frequencies in an exon database extracted form GenBank. ‡Calculated as a product of E(i,j) = Pi x Pj × N, assuming that the association of two introns in the same gene is random: Pi is the proportion of intron phase i actually observed (P_0 = 0.48; P_1 = 0.30; P_2 = 0.22); P_0, P_1 and P_2 are the frequencies of phase zero introns (between two codons), phase one introns (after the first nucleotide within a codon) and phase two introns (after the second nucleotide in a codon); N is the total observed number of intron associations (i,j) (N = 11,559). When i = j the association is called symmetrical exon; when i ≠ j the association is called asymmetrical exon. The observed intron-phase frequencies are significantly different from the expected distribution. Modified with permission from REF. 122 © (2000) Nature Publishing Group.

trypsinogen exons and introns, as supported by the existence of a chimeric gene that encodes the AFGP polymer and trypsinogen[54]. However, although the Arctic *AFGP* gene also encodes a polymer of repeated tripeptide, it shares no sequence identity with the trypsinogen gene. The distinct exon–intron structures and phylogenetic distribution support a separate ancestry for these genes: a common freezing environment shaped Antarctic and Arctic *AFGP* genes 5–14 and 2.5 million years ago, respectively. Indeed, the mechanism that created AFGPs — the selection plus amplification of a Thr-Ala-Ala monomer — differs from other mechanisms (TABLE 1).

These data and analyses detected the force of positive Darwinian selection from the beginning of the new genes through to the evolution of more sophisticated functions, and support the immediate model of adaptive evolution. However, testing for a general role for the adaptive evolution of new genes apparently cannot rely only on analysis of individual cases, and an exhaustive comparison for all new genes is not feasible. A comparison of the fixation probabilities of mutations among different parts of genomes — for example, the X chromosome and autosomes — under various models of genetic evolutionary parameters should detect general forces that drive the evolution of gene duplicates. This genomic approach to population genetic analysis has provided further insights into the roles of various evolutionary forces in the origin of new genes.

Taking advantage of *D. melanogaster* annotated gene sequences, Thornton and Long[55] compared the substitution rates of 107 pairs of paralogues on the X chromosome with 1,734 paralogues on the same or different autosomes, or between the X chromosome and autosomes. The average K_A/K_S between X-linked duplicates (0.4701) is nearly double that of duplicates involving autosomes (0.2581–0.2740). The rapid divergence of X-linked paralogues is most likely to be a consequence of recessive advantageous mutations, as presented in the model of Charlesworth *et al.*[56], rather than a result of changed environmental conditions or a genetic background that has converted previously deleterious variation into beneficial variation[57]. However, in genomic analyses of single-copy X-linked genes, no accelerated substitution was seen[58], which indicates that different forces might act on single-copy genes and gene duplicates. This is understandable because new duplicates are more likely to be evolving new functions under positive selection and the dominance effects might be different between single-copy genes and duplicates[55,59]. In fact, the distribution of K_S in the paralogues seems to indicate that the proportion of young duplicates is higher on the X chromosome than on the autosomes[55], which is consistent with positive selection for evolving new functions and with high substitution rates of X-linked paralogues.

Frequency of origin of new genes

The origination of new genes was previously thought to be a rare event at the level of the genome. This is understandable because, for example, only 1% of human genes have no similarity with the genes of other animals[59], and only 0.4% of mouse genes have no human homologues[60], although it is unclear whether these orphan genes are new arrivals, old survivors or genes that lost their identity with homologues in other organisms. However, it does not take many sequence changes to evolve a new function. For example, with only 3% sequence changes from its paralogues, *RNASE1B* has developed a new optimal pH that is essential for the newly evolved digestive function in the leaf-eating monkey[35]. Although it will take a systematic effort to pinpoint the rate at which new genes evolve, there is increasing evidence from *Drosophila* and mammalian systems that new genes might not be rare[61]. Patthy[62] compiled 250 metazoan modular protein families that were probably created by exon shuffling. Todd *et al.*[63] investigated 31 diverse structural enzyme superfamilies for which structural data were available, and found that almost all have functional diversity among their members that is generated by domain shuffling as well as sequence changes.

The general role of exon shuffling has been investigated by examining the distribution of INTRON PHASES[64,65]. In a large-scale genomic analysis of intron phases using the GenBank database, two unexpected phenomena were observed. First, there was a great excess of phase zero introns. Second, there was a significant excess of symmetric intron associations for those genes that contained more than one intron (TABLE 3). The observed numbers of all three symmetrical exons (0,0), (1,1) and (2,2) in the table are greater than the expected numbers. These peculiar features of non-random intron distribution could not be adequately interpreted as biased intron insertions[66–68]. Rather, they are all probably signatures of exon shuffling, which requires that the length of an inserted exon is a multiple of three and that there is

INTRON PHASE
The relative position of an intron within or between codons. Phase zero, one and two are defined by the position of an intron between two codons or after the first or second nucleotide of a codon, respectively.

			Expected		Observed	
			%	No.	No.	Excess
X	→	A	23.3	37.1	63	70
A	→	X	20.3	32.2	24	–25
A	→	A	56.4	89.7	72	–20

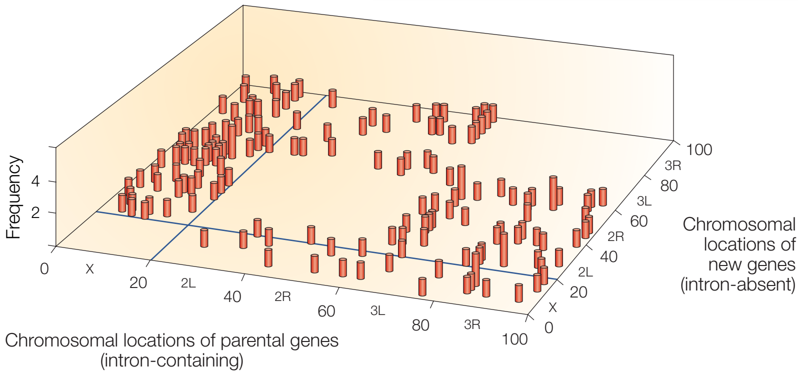


Figure 2 | **The biased distribution of retroposition events in the *Drosophila melanogaster* genome.** A sample is defined as 50% identity between parental and new genes. The cytological regions of five main chromosomal arms (X, 2L, 2R, 3L and 3R) are labelled 0–100. Chromosome 4 was omitted because of its small size. The two blue lines separate the X chromosome and the autosomes. The distribution of the retroposition events in the genome between different chromosomes is summarized in three directions: from the X chromosome to the autosomes (X → A); from the autosomes to the X chromosome (A → X); and from one autosome to another (A → A). The expected percentage (%) of retroposition was calculated on the basis of an assumption of random generation and insertion[85], and the excess was calculated as (observation – expectation)/expectation x 100. A statistical comparison between the observed and expected values (χ^2 = 23.66, degrees of freedom = 2) indicated a significant probability ($p<10^{-5}$).

an identical intron phase at the two ends of an inserted exon and its recipient intron to avoid frameshift mutation[69]. The positions and phases of introns in ancient genes were found to be more correlated with the boundaries of protein modules than they were in recently created genes[7,70,71]. Kaessmann *et al.*[6] observed that introns at the boundaries of domains show a high excess of symmetrical phase combinations, whereas non-boundary introns show no excess symmetry, which indicated that exon shuffling involved the rearrangement of structural and functional domains as a whole. The analysis of ancient gene regions, which are shared by both eukaryotes and prokaryotes, showed a similar signature of exon shuffling, which indicates that this mechanism might have had a role in the early days of life even before the origin of eukaryotes[7,65,70,71]. It should be noted that there is some debate about exon shuffling and the evolution/age of introns; however, recent data (for example, see REFS 70,71) have indicated that the actual picture is more complicated than was implied by the early debate between the 'introns old, exon shuffling old' versus 'introns young, exon shuffling young' schools of thought — this is outside the focus of this review and is not discussed further here.

The deviation from the random prediction of intron-phase distribution to the observed distribution provides an indirect estimate of the contribution of exon shuffling: 19% of eukaryotic exons might have been involved in exon shuffling[65]. Given the distribution of exon number in human genes[72] and general eukaryotic genes[73,74], this observation indicates that the evolution of most eukaryotic genes involved exon shuffling, which is consistent with the conclusion from a detailed structural analysis of protein superfamilies[63].

All of these estimates were based on exon-shuffling mechanisms. TABLE 1 shows that several other mechanisms contribute to the evolution of new gene functions. Remarkably, many of these mechanisms are not observed sporadically, but seem to have shaped a considerable portion of the coding regions of genomes. Retropositions caused by L1 elements contribute 1% of the human genome[75,76]. There are ~10,000 gene sequences in the human genome that have been created by retropositions[77], many of which have been shown to be functional and have expression patterns that are different from those of their parental genes, which indicates that they might have possible new functions[61]. Various mobile-element fragments are translated into 4% of mammalian proteins[16] and 5% of human alternatively spliced exons are *Alu*-derived[78]. Gene duplication has been shown to be a principal source of new gene evolution, although how many of the duplications simply maintain redundant functions is unclear. So, it is safe to conclude that new genes with new functions are not as rare as was previously thought.

X-chromosome bias in the origin of new genes

A recurrent pattern in the distribution of new genes or in the distribution of the parental coding sequences from which they evolved, might provide clues to the evolutionary forces that underlie new gene evolution at the genomic level. Understanding how such a pattern arose would help to understand genome structure as well as the functional implications of this process.

Among early efforts to detect a pattern in the origin of genes was the genetic analysis of Lifschytz and Lindsley[79]. They proposed that male GERMLINE INACTIVATION of X chromosomes in *Drosophila* would have an effect on the chromosomal distribution of genes. Investigation of human phosphoglycerate kinase (PGK) glucose metabolism genes attracted attention to the relationship between chromosome location and function[80,81]. The X-linked *PGK1* is ubiquitously expressed, except in male germline cells owing to male germline X-chromosome inactivation, whereas the retroposed autosomal *PGK2*, which originated from *PGK1*, is expressed only during late spermatogenesis[82]. Apparently, *PGK2* provides PGK function when *PGK1* is silent after X-chromosome inactivation. This study, and others[83,84], led to a large-scale test of gene movement from the X chromosome by retroposition in the human genome. It failed to detect any patterns of gene movement[72], which showed that in humans some retroposed genes moved from the autosomes to the X chromosome as well.

However, a clear pattern in the origin of new genes was identified recently in *Drosophila*: there is a significant excess of retrogenes that are generated by X-linked parental genes escaping to autosomes[85] (FIG. 2). This biased distribution is not predicted by a model of the random generation and insertion of new retroposed genes.

GERMLINE INACTIVATION
The early inactivation of the sex chromosomes in germline cells in the heterogametic sex.

Considering the gene number and euchromatin size of chromosomes, and the relative population sizes of the X chromosome and autosomes, only 23.3% of retroposed genes should be expected to derive from X-linked parents, which is far lower than the numbers observed: 40–50% of retrogenes were derived from the X chromosome. Conversely, the autosome-generated new genes were more disproportionately inserted into the autosomes than into the X chromosome, which indicated that these genes tended to avoid X-linkage. So, it becomes obvious that retroposed genes are escaping from X-linkage; that is, retroposition in the *Drosophila* genome is an asymmetric process between the X chromosome and the autosomes (FIG. 2).

What forces are driving this asymmetric distribution of retroposition events? At first glimpse, it might easily be assumed that retroposition is a biased mutation process that would generate more X-chromosome-to-autosome insertions. Several factors, such as a hypothetical difference in transcription rates, recombination rates, mutation rates and negative selection against insertion, have all been tested and found to be irrelevant as they do not differ significantly between the X chromosome and the autosomes[85,86]. Another hypothesis for this bias is that it might not be immediately possible for DOSAGE COMPENSATION to be established for the newly inserted retroposed copy in the male *Drosophila* X chromosome[87]. However, this hypothesis cannot explain the similar bias that is seen in the human genes that are expressed during male meiosis[85]. A different dosage compensation mechanism is used in the latter — one X chromosome in females is inactivated by *XIST* TRANSCRIPTS[88], rather than the X-linked genes being hypertranscribed as they are in *Drosophila* males[89].

Two forms of selection are likely to be responsible for the asymmetric movement of new genes. The first is that X-chromosome inactivation in early spermatogenesis, which is well documented in humans, mice and *Drosophila*[79,90,91], might create different selective pressures on the X chromosome and autosomes for those genes that are expressed during X-chromosome inactivation. So, autosomal retroposed genes that are functional during X-chromosome inactivation would be favoured over those that are X linked. The strong correlation between the timing of expression and the chromosomal locations of spermatogenesis genes[91,92] supports this selective model, as these genes seem to avoid inactivation. Also, X-linked genes might be subject to the effect of sexual antagonism, which would favour one sex at the cost of the other[93], and so might be redistributed in autosomes[91]. This could account for the excess male-specific autosomal genes — for example, accessory gland proteins[87] and other male somatic-expressed autosomal genes[94] — that are not involved in the meiotic process with X-chromosome inactivation in spermatogenesis. These two forms of selection make a similar prediction: that X-chromosome-derived autosomal genes should be related to male-specific functions. This is confirmed by the observed high percentage (91%) of X-chromosome-derived retrogenes on autosomes that are expressed in the testis[85].

An inescapable consequence of this biased movement would be that most male-specific genes in *Drosophila* would be located in autosomes — they would shun the X chromosome. Interestingly, genome-wide analyses of the location of male and female genes in *Drosophila* have shown an excess of male genes on the autosomes[94,95]. Using microarray and expressed sequence tag analyses in *D. melanogaster*[94], and comparative microarray analysis in *D. melanogaster* and *D. simulans*[95], many genes that are expressed in male germline and somatic tissues were identified and most were found to be located on autosomes — the X-linked male genes were underrepresented. Also, most sterility genes that function late in spermatogenesis in *Drosophila* are autosomal[91]. A similar excess of autosomal male-specific genes was observed in *Caenorhabditis elegans*[96], in which X-chromosome inactivation is also known to occur[97].

Conclusions and perspectives

By analysing young genes, progress has been made in understanding the molecular mechanisms and evolutionary forces that underlie the origin of new genes. Analyses of individual new genes have provided many details of the processes that are involved and have defined new problems for research. Extending the studies of individual new genes to the genomic analysis of the rates and patterns of new gene formation has further highlighted the generality of this process.

The classical model of new gene origination by duplication has been tested by examining young duplicates of ribonuclease and other genes, and has shown a clear link between new biological traits and the evolved functions of new proteins. The expansion of protein diversity through exon shuffling and, therefore, domain shuffling has been shown to be unexpectedly common. Also, retroposition and transposable elements have been found to shape the structures of protein-coding genes in an impressive portion of mammalian genomes. Lateral gene transfer and gene fusion/fission not only have an effect on prokaryotic genomes but have also been identified in higher eukaryotic genomes.

An obvious feature that is associated with the formation of new genes by these mechanisms is the rapid evolution of new sequences, structures and expression. Most of these evolutionary changes are probably adaptive changes made under positive Darwinian selection, as tests in several new genes and selective laboratory experiments in yeast and *Escherichia coli* have shown. These studies have also shown that new genes are not as rare as was previously thought. Furthermore, the finding that new retroposed genes in fruitflies 'escape' from the X chromosome has shown a general pattern of new gene evolution.

Nonetheless, little is known about the genomic process of new gene evolution because of the challenge of identifying an adequate number of young genes with identified new functions. The origination rate of new gene functions, which is an exciting measurement that will show how quickly organisms have changed their genetic diversity and functional complexity to adapt to

DOSAGE COMPENSATION
The phenomenon whereby the expression levels of sex-linked genes are made equal in males and females of heterogametic species.

XIST TRANSCRIPT
A non-coding RNA that is transcribed by an X-linked gene known as *Xist* (X-inactive-specific transcription), which has a role in the somatic transcriptional inactivation of one X chromosome in female mammals. This is believed to occur through the interaction of transcripts from *Xist* and the related gene *Tsix*.

variable environments, has yet to be observed. Also, the phylogenetic distribution of new genes will provide an overview of new gene origination, its rate and speed variation. Stochastic-process modelling of the origination rates would be useful to help understand the formation process, and both experimental and computational genomic analyses will be productive in identifying new gene functions. Meanwhile, it is imperative to experimentally test the biochemical or phenotypic functions of new genes and to understand how selection works during their formation and evolution. Although further investigation and understanding of these questions will continue to depend on theoretical and careful functional analyses of individual genes — demanding the kind of effort shown by the heroic endeavours of generations of evolutionary and molecular biologists in the past half century — we are now much better equipped in terms of sequence databases and technology than we were only a decade ago.

1. Gilbert, W. Why genes in pieces? *Nature* **271**, 44 (1978).
2. van Rijk, A. & Bloemendal, H. Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica* **118**, 245–249 (2003).
3. van Rijk, A. A., de Jong, W. W. & Bloemendal, H. Exon shuffling mimicked in cell culture. *Proc. Natl Acad. Sci. USA* **96**, 8074–8079 (1999).
4. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
5. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367 (2000).
6. Kaessmann, H., Zollner, S., Nekrutenko, A. & Li, W. H. Signatures of domain shuffling in the human genome. *Genome Res.* **12**, 1642–1650 (2002).
Based on analysis of human and *C. elegans* genomes, this paper provides new evidence that intron-phase correlation is a sign of exon shuffling that recombines protein-coding domains to form a new gene.
7. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert, W. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl Acad. Sci. USA* **93**, 14632–14636 (1996).
8. Patthy, L. Exon shuffling and other ways of module exchange. *Matrix Biol.* **15**, 301–310, 311–312 (1996).
9. Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
10. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
11. Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nature Rev. Genet.* **3**, 827–837 (2002).
12. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
13. Samonte, R. V. & Eichler, E. E. Segmental duplications and the evolution of the primate genome. *Nature Rev. Genet.* **3**, 65–72 (2002).
14. Makalowski, W., Mitchell, G. A. & Labuda, D. Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.* **10**, 188–193 (1994).
15. Makalowski, W. in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (ed. Maraia, R. J.) 86–104 (Landes Company, Austin, 1995).
16. Nekrutenko, A. & Li, W. H. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* **17**, 619–621 (2001).
17. Lorenc, A. & Makalowski, W. Transposable elements and vertebrate protein diversity. *Genetica* **118**, 467–477 (2003).
18. Ochman, H. Lateral and oblique gene transfer. *Curr. Opin. Genet. Dev.* **11**, 616–619 (2001).
19. de Koning, A. P., Brinkman, F. S., Jones, S. J. & Keeling, P. J. Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. *Mol. Biol. Evol.* **17**, 1769–1773 (2000).
20. Bergthorsson, U., Adams, K. L., Thomason, B. & Palmer, J. D. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424**, 197–201 (2003).
21. Thomson, T. M. *et al.* Fusion of the human gene for the polyubiquitination coeffector UEV1 with *Kua*, a newly identified gene. *Genome Res.* **10**, 1743–1756 (2000).
22. Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D. & Hartl, D. L. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* **396**, 572–575 (1998).
In this study, a young *D. melanogaster* gene showed unusually rapid changes in its structure, from seemingly unrelated resources, which led to unexpected new functions of the gene in sperm tails.
23. Wang, W., Brunet, F. G., Nevo, E. & Long, M. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **99**, 4448–4453 (2002).
24. Wang, W., Thornton, K., Berry, A. & Long, M. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science* **295**, 134–137 (2002).
25. Haldane, J. B. S. The part played by recurrent mutation in evolution. *Am. Nat.* **67**, 5–19 (1933).
26. Fisher, R. A. The sheltering of lethals. *Am. Nat.* **69**, 446–455 (1935).
27. Hughes, A. L. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B* **256**, 119–124 (1994).
28. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
29. Walsh, J. B. How often do duplicated genes evolve new functions? *Genetics* **139**, 421–428 (1995).
30. Walsh, B. Population-genetic model of the fates of duplicate genes. *Genetica* **118**, 279–294 (2003).
31. Gu, X. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18**, 453–464 (2001).
32. Long, M. & Langley, C. H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
33. Messier, W. & Stewart, C. B. Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154 (1997).
34. Zhang, J., Rosenberg, H. F. & Nei, M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA* **95**, 3708–3713 (1998).
35. Zhang, J., Zhang, Y. P. & Rosenberg, H. F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genet.* **30**, 411–415 (2002).
In this study, the authors combine sequence analysis and recombination protein technology to detect positive selection in the adaptive evolution of a new gene duplicate to the changed digestive system in the leaf-eating colobine monkey.
36. Ohta, T. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. *Genetics* **138**, 1331–1337 (1994).
37. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
38. Begun, D. J. Origin and evolution of a new gene descended from *alcohol dehydrogenase* in *Drosophila*. *Genetics* **145**, 375–382 (1997).
39. Maston, G. A. & Ruvolo, M. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol. Biol. Evol.* **19**, 320–335 (2002).
This paper discusses how an important function (involving pregnancy establishment in human and other primates) arose by gene duplication, supported by ample biological data.
40. Paulding, C. A., Ruvolo, M. & Haber, D. A. The *Tre2* (*USP6*) oncogene is a hominoid-specific gene. *Proc. Natl Acad. Sci. USA* **100**, 2507–2511 (2003).
41. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc. Natl Acad. Sci. USA* **93**, 7727–7731 (1996).
42. Ranz, J. M., Ponce, A. R., Hartl, D. L. & Nurminsky, D. Origin and evolution of a new gene expressed in the *Drosophila* sperm axoneme. *Genetica* **118**, 233–244 (2003).
43. Betrán, E., Wang, W., Jin, L. & Long, M. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol. Biol. Evol.* **19**, 654–663 (2002).
44. Brown, C. J., Todd, K. M. & Rosenzweig, R. F. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* **15**, 931–942 (1998).
This study of experimental evolution shows that a new gene function can arise rapidly as a response to a changed environment.
45. Hall, B. G. The EBG system of *E. coli*: origin and evolution of a novel β-galactosidase for the metabolism of lactose. *Genetica* **118**, 143–156 (2003).
46. McDonald, J. H. & Kreitman, M. Adaptative protein evolution at the *Adh* locus in *Drosophila. Nature* **351**, 652–654 (1991).
47. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
48. Malik, H. S. & Henikoff, S. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. *Genetics* **157**, 1293–1298 (2001).
49. Betrán, E. & Long, M. *Dntf-2r:* a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. *Genetics* **164**, 977–988 (2003).
50. Llopart, A., Comeron, J. M., Brunet, F. G., Lachaise, D. & Long, M. Intron presence–absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc. Natl Acad. Sci. USA* **99**, 8121–8126 (2002).
51. Cheng, C. H., Chen, L., Near, T. J. & Jin, Y. Functional antifreeze glycoprotein genes in temperate-water New Zealand nototheniid fish infer an Antarctic evolutionary origin. *Mol. Biol. Evol.* 28 July 2003 (doi:10.1093/molbev/msg208).
52. Chen, L., DeVries, A. L. & Cheng, C. H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl Acad. Sci. USA* **94**, 3817–3822 (1997).
This study clearly showed how a similar environmental challenge in two different geographic locations created similar genes with the same function from different DNA materials.
53. Chen, L., DeVries, A. L. & Cheng, C. H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl Acad. Sci. USA* **94**, 3811–3816 (1997).
54. Cheng, C. H. & Chen, L. Evolution of an antifreeze glycoprotein. *Nature* **401**, 443–444 (1999).
55. Thornton, K. & Long, M. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol. Biol. Evol.* **19**, 918–925 (2002).
56. Charlesworth, B., Coyne, J. A. & Barton N. H. The relatives rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (1987).
57. Orr, H. A. & Betancourt, A. J. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157**, 875–884 (2001).
58. Betancourt, A. J., Presgraves, D. C. & Swanson, W. J. A test for faster X evolution in *Drosophila*. *Mol. Biol. Evol.* **19**, 1816–1819 (2002).
59. The Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
60. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
61. Betrán, E. & Long, M. Expansion of genome coding regions by acquisition of new genes. *Genetica* **115**, 65–80 (2002).
62. Patthy, L. *Protein Evolution by Exon-shuffling* (Springer, New York, 1995).
63. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
A detailed structural and sequence analysis of 31 enzyme superfamilies for which structural data are available. Interestingly, almost all of the families were subject to domain shuffling, which highlights the generality of the mechanism in new gene origination.
64. Long, M., de Souza, S. J. & Gilbert, W. Evolution of the intron–exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.* **5**, 774–778 (1995).
65. Long, M., Rosenberg, C. & Gilbert, W. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA* **92**, 12495–12499 (1995).

66. Gilbert, W., de Souza, S. J. & Long, M. Origin of genes. *Proc. Natl Acad. Sci. USA* **94**, 7698–7703 (1997).
67. Long, M. Evolution of novel genes. *Curr. Opin. Genet. Dev.* **11**, 673–680 (2001).
68. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. *Proc. Natl Acad. Sci. USA* **95**, 219–223 (1998).
69. Patthy, L. Intron-dependent evolution: preferred types of exons and introns. *FEBS Lett.* **214**, 1–7 (1987).
70. Roy, S. W., Lewis, B. P., Fedorov, A. & Gilbert, W. Footprints of primordial introns on the eukaryotic genome. *Trends Genet.* **17**, 496–501 (2001).
The authors described an unexpectedly significant correlation between the distribution of intron phases and the age of the host genes, which supported a model of ancient exon–intron structure with gradual intron addition.
71. Fedorov, A., Roy, S., Cao, X. & Gilbert, W. Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. *Genome Res.* **13**, 1155–1157 (2003).
72. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
73. Deutsch, M. & Long, M. Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**, 3219–3228 (1999).
74. Long, M. & de Souza, S. J. Intron–exon structures: from molecular to population biology. *Adv. Genome Biol.* **5A**, 143–178 (1998).
75. Kazazian, H. H. J. L1 retrotransposons shape the mammalian genome. *Science* **289**, 1152–1153 (2000).
76. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000).
77. Harrison, P. M. *et al.* Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280 (2002).
78. Sorek, R., Ast, G. & Graur, D. Alu-containing exons are alternatively spliced. *Genome Res.* **12**, 1060–1067 (2002).
79. Lifschytz, E. & Lindsley, D. L. The role of X-chromosome inactivation during spermatogenesis (*Drosophila*-allocycly-chromosome evolution-male sterility-dosage compensation). *Proc. Natl Acad. Sci. USA* **69**, 182–186 (1972).
80. McCarrey, J. R. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. *Gene* **61**, 291–298 (1987).
81. McCarrey, J. R. Molecular evolution of the human Pgk-2 retroposon. *Nucleic Acids Res.* **18**, 949–955 (1990).
82. McCarrey, J. R. Evolution of tissue-specific gene expression in mammals: how a new phosphoglycerate kinase was formed and refined. *BioScience* **44**, 20–27 (1994).
83. Pan, Y., Decker, W. K., Huq, A. H. & Craigen, W. J. Retrotransposition of glycerol kinase-related genes from the X chromosome to autosomes: functional and evolutionary aspects. *Genomics* **59**, 282–290 (1999).
84. Elliott, D. J. *et al.* An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. *Hum. Mol. Genet.* **9**, 2117–2124 (2000).
85. Betrán, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* **12**, 1854–1859 (2002).
86. Kaminker, J. S. *et al.* The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* **3**, 0084 (2002).
87. Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F. & Aquadro, C. F. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98**, 7375–7379 (2001).
88. Avner, P. & Heard, E. X-chromosome inactivation: counting, choice and initiation. *Nature Rev. Genet.* **2**, 59–67 (2001).
89. Franke, A. & Baker, B. S. Dosage compensation rox! *Curr. Opin. Cell Biol.* **12**, 351–354 (2000).
90. Richler, C. *et al.* Splicing components are excluded from the transcriptionally inactive XY body in male meiotic nuclei. *Mol. Biol. Cell* **5**, 1341–1352 (1994).
91. Wu, C-I. & Xu, E. Y. Sexual antagonism and X inactivation — the SAXI hypothesis. *Trends Genet.* **19**, 243–247 (2003).
92. Wang, P. J., McCarrey, J. R., Yang, F. & Page, D. C. An abundance of X-linked genes expressed in spermatogonia. *Nature Genet.* **27**, 422–426 (2001).
93. Rice, W. R. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**, 735–742 (1984).
94. Parisi, M. *et al.* Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science* **299**, 697–700 (2003).
95. Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D. & Hartl, D. L. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300**, 1742–1745 (2003).
96. Reinke, V. *et al.* A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **6**, 605–616 (2000).
97. Kelly, W. G. *et al.* X-chromosome silencing in the germline of *C. elegans*. *Development* **129**, 479–492 (2002).
98. Langley, C. H., Montgomery, E. & Quattlebaum, W. F. Restriction map variation in the *Adh* region of *Drosophila*. *Proc. Natl Acad. Sci. USA* **79**, 5631–5635 (1982).
99. Jeffs, P. & Ashburner, M. Processed pseudogenes in *Drosophila*. *Proc. R. Soc. Lond. B* **244**, 151–159 (1991).
100. 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.* **17**, 1294–1301 (2000).
101. Long, M., Wang, W. & Zhang, J. Origin of new genes and source for N-terminal domain of the chimerical gene, *jingwei*, in *Drosophila*. *Gene* **238**, 135–141 (1999).
102. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349 (1996).
103. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55**, 631–661 (1986).
104. Brosius, J. Retroposons — seeds of evolution. *Science* **251**, 753 (1991).
105. Javaud, C., Dupuy, F., Maftah, A., Julien, R. & Petit, J-M. The fucosyltransferase gene family: an amazing summary of the underlying mechanisms of gene evolution. *Genetica* **118**, 157–170 (2003).
106. Hughes, A. *Adaptive Evolution of Genes and Genomes*, (Oxford Univ. Press, Oxford, 2000).
107. Courseaux, A. & Nahon, J. L. Birth of two chimeric genes in the Hominidae lineage. *Science* **291**, 1293–1297 (2001).
This paper reported two chimeric genes — *PMCHL1* and very young *PMCHL2* — in Hominidae, which showed that an intron-containing gene structure can be retroposed to another genomic location.
108. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Jr. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
109. Brosius, J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**, 115–134 (1999).
110. Brosius, J. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118**, 99–116 (2003).
111. Makalowski, W. Genomic scrap yard: how genomes utilize all that junk. *Gene* **259**, 61–67 (2000).
112. Lorenc, A. & Makalowski, W. Transposable elements and vertebrate protein diversity. *Genetica* **118**, 183–191 (2003).
113. Ragan, M. A. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* **201**, 187–191 (2001).
114. McCarthy, A. D. & Hardie, D. G. Fatty acid synthase — an example of protein evolution by gene fusion. *Trends Biochem. Sci.* **4**, 60–63 (1984).
115. Snel, B., Bork, P. & Huynen, M. Gene fusion versus gene fission. *Trends Genet.* **16**, 9–11 (2000).
116. Martignetti, J. A. & Brosius, J. Neural BC1 RNA as an evolutionary marker: guinea pig remains a rodent. *Proc. Natl Acad. Sci. USA* **90**, 9698–9702 (1993).
117. Martignetti, J. A. & Brosius, J. BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. *Proc. Natl Acad. Sci. USA* **90**, 11563–11567 (1993).
118. Zhang, J., Webb, D. M. & Podlaha, O. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. *Genetics* **162**, 1825–1835 (2002).
119. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
120. Brosius, J. & Gould, S. J. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc. Natl Acad. Sci. USA* **89**, 10706–10710 (1992).
121. Gogolevskaya, I. K. & Kramerov, D. A. Evolutionary history of 4.5SI RNA and indication that it is functional. *J. Mol. Evol.* **54**, 354–364 (2002).
122. Long, M. Protein-coding segments: evolution of exon–intron gene structure. Nature Encyclopedia of Life Sciences [online], <http://www.els.net> (doi:10.1038/npg.els.0000887) (2000).

Acknowledgements

We thank J. Sporfford, J. Zhang and the referees for their helpful suggestions. We also thank all members of the M.L. laboratory, past and present, for their devoted contributions to the studies on the origin of new genes and their evolution. M.L. is funded by a David and Lucile Packard Foundation Fellowship, a National Science Foundation CAREER Award and a National Institutes of Health grant.

Online links

DATABASES

The following terms in this article are linked online to:

FlyBase: http://flybase.bio.indiana.edu

Adh | *Sdic* | *sphinx* | *ymp* | *ynd*

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink

DAF | *PGK1* | *PGK2* | *USP6*

***Saccharomyces* Genome Database:**

http://www.yeastgenome.org

HXT6 | *HXT7*

FURTHER INFORMATION

GenBank: http://www.ncbi.nlm.nih.gov/Genbank/index.html

Institute of Experimental Pathology/Molecular Neurobiology:

http://zmbe2.uni-muenster.de/expath/frames.htm

Manyaun Long's laboratory:

http://pondside.uchicago.edu/~longlab/longlab.html

Molecular Population Genetics: http://www.molpopgen.org

The Genomic Scrap Yard:

http://www.ncbi.nlm.nih.gov/CBBresearch/Makalowski/ScrapYard

Access to this interactive links box is free online.