nature genetics

Review article https://doi.org/10.1038/s41588-024-02059-0

Functional innovation through new genes as a general evolutionary process

Received: 12 June 2024

Accepted: 15 December 2024

Published online: 28 January 2025

Check for updates

Shengqian Xia[1], Jianhai Chen[1], Deanna Arsala[1], J. J. Emerson[2] & Manyuan Long[1]

In the past decade, our understanding of how new genes originate in diverse organisms has advanced substantially, and more than a dozen molecular mechanisms for generating initial gene structures were identified, in addition to gene duplication. These new genes have been found to integrate into and modify pre-existing gene networks primarily through mutation and selection, revealing new patterns and rules with stable origination rates across various organisms. This progress has challenged the prevailing belief that new proteins evolve from pre-existing genes, as new genes may arise de novo from noncoding DNA sequences in many organisms, with high rates observed in flowering plants. New genes have important roles in phenotypic and functional evolution across diverse biological processes and structures, with detectable fitness effects of sexual conflict genes that can shape species divergence. Such knowledge of new genes can be of translational value in agriculture and medicine.

Newly evolved genes enhance the diversity of gene functions, thereby having a fundamental role in the diversity of organisms at all levels of complexity and across phenotypes. However, it was not until the early 1990s that the challenge of identifying newly evolved genes and deciphering their origination processes and functions became empirically feasible[1]. A new gene refers to a gene that has originated at a specific point in evolutionary history or within a particular lineage or species and was either previously absent or lacks detectable orthologues in closely related species or ancestral genomes. Based on this concept, we introduce the framework of gene age, which can be estimated from their evolutionary origin. Accordingly, genes can be categorized along a continuum, ranging from ancient, old and middle-aged to recently originated young genes, spanning evolutionary time scales from billions or hundreds of million years ago (Mya) to the scale of a few hundreds of Mya to tens of Mya or younger.

Although pioneers in the past century speculated on the problem of new gene evolution[2–4], the field of molecular and evolutionary biology neglected the concept of new gene evolution as it was skeptical of functional innovation in evolution. Their views were perhaps skewed by a perceived improbability of the process. Even at the turn of this century, geneticists also asserted that essential genes are not organism-specific[5]. The static view of genes and their functions in evolution was long-held[6,7] but was first challenged by the identification of *jingwei* and its encoded new dehydrogenase involved in hormone and pheromone metabolism, which were generated by exon shuffling 3 Mya in African *Drosophila*[1,8]. Propelled by advancements in genomics, gene editing and molecular biology over the recent decade, the dynamic nature and generality of new gene evolution have been unveiled through functional, mechanistic and evolutionary studies of newly emerged genes.

We review the advances made in the field of new gene evolution in main issues, including molecular mechanisms, specifically the stepwise de novo origination, functionality and phenotypes, evolutionary processes and patterns, underlying evolutionary forces and potential applications in agriculture and medicine. We provide perspectives for the future study of new genes and discussion of possible impacts of the understanding of age effects of genes in genetics and other disciplines of biology.

Molecular mechanisms to generate new genes

Historically, gene duplication with subsequent divergence was the first mechanism perceived for functional innovation in evolution[3]. It was

[1]Department of Ecology and Evolution, The University of Chicago, Chicago, IL, USA. [2]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA. e-mail: mlong@uchicago.edu

realized in the 1970s that recombination of pre-existing genetic materials between and within genes could also create new genes[4]. With the extensive studies of molecular genetics, today a dozen new mechanisms have been reported to create new genes with new functions.

Diverse molecular mechanisms for forming new gene structures

Up to 14 distinct molecular biological processes have been identified for the origination of new genes (Fig. 1 and Supplementary Table 1) in addition to the classic mechanism of gene duplication[3]. These mechanisms can be classified into the following four categories: (1) protein to protein, whereby new protein-coding genes are derived from pre-existing protein-coding genes, through exon shuffling[9], retroposition[10], gene duplication[11,12], lateral gene transfer[13–15], gene fusion[16], gene fission[17], new isoform divergence[18] and reading frameshift[19]; (2) noncoding to protein, by which protein-coding genes originate from noncoding DNAs, including de novo genes[20–25], by short repeat expansion[26], transposable element (TE) domestication[27] and bidirectional promoter use[28]; (3) protein to noncoding[29]; here long noncoding RNA (lncRNA) genes are formed by pseudogenization, for example, the *Xist* gene encodes noncoding RNAs to inactivate mammalian X chromosomes in male germlines; and (4) noncoding to noncoding[30], where lncRNA genes are derived from noncoding DNAs. It should be noted that lncRNA genes may also arise through other mechanisms, including exon shuffling, retroposition and gene duplication[31]. In addition, new gene or protein isoforms can be generated through alternative splicing or changes in start and/or stop codons or transcription start sites[18].

Typically, DNA-based duplication with divergence was considered to be the primary mechanism for new gene evolution, as such duplications were the earliest and most commonly observed mechanism[3]. However, subsequent genomic analyses showed that many nonduplicative mechanisms frequently have roles in the generation of new gene structures. For example, 19% of eukaryotic exons were estimated to be involved in exon shuffling, affecting hundreds of genes in flowering plants and numerous genes in *Drosophila* and fish, revealing clear patterns of exon recombination[32–34]. In plants, a transposable element group called Pack-Mules was found to duplicate extensively and recombine with protein-coding genes, often leading to thousands of functional chimeras[35]. Using comparative and population genomics, signatures of functional divergence can often be detected in retroposed genes, suggesting that retroposition tends to acquire new roles rather than merely partitioning existing ones[36].

The generation of new regulatory elements in new genes

One important question is how new genes gain new promoters. Studies have demonstrated that, alternatively, existing regulatory elements can be recruited as promoters for many new genes in *Drosophila* and adopt histone modifications consistent with active expression in the testis[37,38]. In the human genome, core promoters are frequently sourced from transposable elements[39]. Newly duplicated gene copies often disappear from a species population before beneficial nucleotide changes arise owing to genetic drift or negative selection, coined as 'Ohno's dilemma'[40]. To solve this paradox, the 'enhancer capture divergence' model posits that a new gene, exemplified by *Umbrea* in *Drosophila*, when copied into close proximity of an existing tissue-specific enhancer, may capture its expression pattern, avoiding being eliminated[41]. Indeed, laboratory evolution of *Escherichia coli* shorter than 50,000 generations resulted in the fixation of nine new genes in the resulting populations owing to the generation of new promotors[42].

De novo gene generation

Interest in the identification and evolutionary analysis of de novo gene origination has intensified, with related studies in humans, *Drosophila* and yeast[20–23,43,44]. Historically, de novo gene generation (Fig. 1 and Supplementary Table 1) was thought to be impossible despite a neglected discovery of a de novo protein identified from a bacterium that digested plastic material[45].

The concept of de novo gene generation can be traced to at least three lines of evidence[46,47]. The first derives from the functional and adaptive study of antifreeze proteins (AFPs) in polar fishes[47]. The molecular characterization of antifreeze glycoproteins (AFGPs) and AFPs in these fishes revealed a structurally simple protein, comprising repeats of three amino acids, which bind to ice crystals and prevent their growth within the fish body fluids[25,46,47]. The sources of the repetitive sequences are noncoding sequences from introns or intergenic regions, contributing to partial or complete de novo genes encoding AFGPs in Antarctic notothenioids and northern codfishes, respectively[25,46,47]. The second line of evidence emerged from the characterization of the yeast genome, which led to the discovery of orphan genes[48]. The final piece of evidence came from the broad identification and widespread demonstration of the translation of de novo genes in diverse organisms, such as mice[49–51], yeast[24,44,52], *Drosophila*[53,54], rice[55,56], *Arabidopsis*[57,58] and human[59,60]. The widespread translation of unannotated open reading frames (ORFs) in yeast as determined by ribosome profiling suggests that many more sequences are being translated into proteins than previously recognized[61].

De novo genes versus orphan genes

Scrutiny of orphan genes led to the realization that they could be generated by other mechanisms besides de novo origination[56,62,63]. These mechanisms include but are not limited to rapid divergence, gene loss, lateral gene transfer from rapidly evolved hosts such as viruses and bacteria, or the use of alternative reading frames[64]. Indeed, it has been estimated that 91% of nematode orphan genes might not be created through de novo mechanisms[65]. In contrast, data from rice show that 80% of orphan genes are originated de novo, and most young de novo genes comprise one exon[56]. These studies show clearly that the orphan gene is not a synonym for de novo origination. Only a small number of orphan genes have been found to be authentic de novo genes in humans[59], and a few mouse orphan genes were reported to have noncoding ancestors, such as the mouse gene *Gm13030*, which is regulated by pregnancy cycles[66]. In total, 11 out of 75 mouse de novo genes are very likely to originate from noncoding sequences[50]. Furthermore, 106 orphan genes in *Drosophila* were shown to have originated de novo[43,54]. Numerous de novo genes were also identified in nematodes[65,67] and yeasts[24,68]. Here the main challenge of identifying genes as truly de novo relies on finding evidence of noncoding ancestry[69]. For example, orthologous noncoding regions in related species need to be determined to identify putative noncoding ancestors of candidate de novo genes[70].

The prevalence of de novo genes

In *Oryza* and other grass species, 175 young genes were identified and their short stepwise evolutionary histories (<3.4 million years) reconstructed[56] (Fig. 2a,b), revealing a high rate of de novo ~50 genes per million years (g M^{-1}; Fig. 2b). The following two lines of evidence support these genes being de novo genes: the reconstructed history of stepwise evolution from recent, highly similar, noncoding ancestors (Fig. 2a,c) and the detection of proteins arising from these genes by using mass spectrometry-based targeted proteomics and ribosomal profiling[56]. Another example is *Gm13030*, a protein-coding de novo gene specific to house mice mentioned above, whose protein product was confirmed by both ribosome profiling and mass spectrometry and shows a typical stepwise origination process[66] (Fig. 2d). In bamboo plants, more than a dozen de novo genes were identified to encode proteins involved in stem development[71].

Rate of new gene evolution

Analyses of 6,794 genomes reveal a tremendous amount of variation in gene numbers and genome sizes in organisms (spanning eight orders

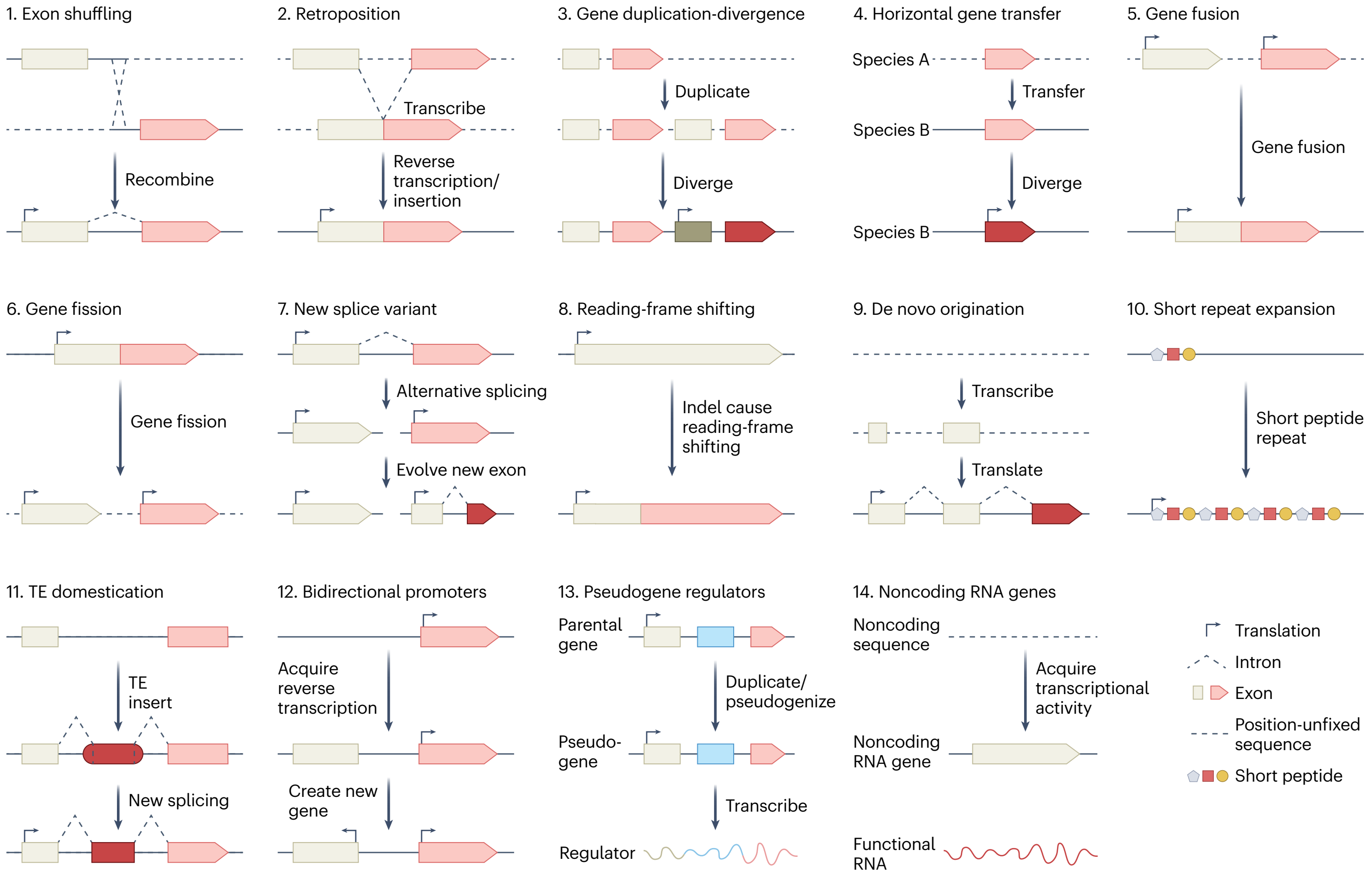


Fig. 1 | Molecular mechanisms of new gene structure origination. As illustrated here, there are multiple molecular mechanisms for generating structures of new genes, whereas different gene origination mechanisms can overlap to give rise to a single new gene. Further details on each mechanism and example genes can be found in Supplementary Table 1. Indel, insertion or deletion mutation.

of magnitude in size and five orders of magnitude in number; Supplementary Fig. 1). Comparative genomics has shown that genomic DNA content undergoes rapid turnover, with genome duplications having a substantial role in increasing genome size over time[72]. Segmental duplications also contribute to this growth and provide essential material for the formation of new genes[73,74]. In vertebrates and plants, duplicates arising from genome duplications are crucial for developing organismal complexity and adapting to environmental stresses[75–77]. However, gene loss due to adaptive genome shrinkage has been reported in metazoans[78,79]. The positive correlation between genome size and gene number suggests that new gene evolution is a general process across all organisms[20,80] (Supplementary Fig. 1).

The distribution of a new gene across an evolutionary tree reveals its age, and the availability of additional taxa permits higher resolution with short branches of one or a few million years length in the tree[1,81] (Box 1). Furthermore, the development of methods to determine orthologues and paralogues through gene synteny comparison between genomes of recently diverged species makes the age of new genes easier to date[82] (Box 1). Many organisms, such as *Saccharomyces cerevisiae*[83], *Drosophila melanogaster*[84], *Homo sapiens*[85], *Arabidopsis lyrata*[86], *Caenorhabditis nigoni*[87] and *Mus musculus*[88], were detected with the rates of gene generation events (Fig. 3 and Supplementary Table 2).

The genomes of humans, mice, *Drosophila*, nematodes, *Arabidopsis* and yeast each have a substantial proportion of genes that are lineage-specific in a short time scale, detected using the synteny-based parsimony method (Box 1). These lineages, typically of interest to researchers, have lasted a few tens of million years, during which young genes originate and become fixed in species populations (Fig. 3). The human genome contains 461 genes that originated from its ancestral primate lineage in 90 Mya[89]. Within these 461 genes, 84 human-specific genes appeared in the recent few million years[89] (Fig. 3). The mouse genome contains 2,316 rodent-specific genes[88] (<70 Mya; Fig. 3). In *D. melanogaster*, 1,124 genes originated in its ancestral *Sophophora* subgenus lineage (<62 Mya; Fig. 3), while *Drosophila virilis* acquired 652 genes in its ancestral *Drosophila* subgenus lineage (<62 Mya)[84]. In nematodes, *C. nigoni* possesses 2,424 genes that originated from the ancestral lineage within 60 Mya when the ancestor of *Caenorhabditis elegans* diverged[65] (Fig. 3). In *A. lyrata*, 1,955 new genes originated in the *Arabidopsis* lineage within 21 million years[86] (Fig. 3). *S. cerevisiae* obtained 353 new genes from its recent *Saccharomyces* ancestors (Fig. 3). These data revealed that there are large numbers of young genes in various recent lineages of metazoans, fungi and flowering plants.

Phylostratigraphic analyses show that new genes have continuously emerged on a long time scale, often throughout the entire course of evolutionary history (Box 1). This measurement allows us to compare the speeds of new gene origination between various evolutionary stages. For example, 9,660 *Arabidopsis thaliana* genes (35% of total genes) can be traced back to different evolutionary stages since their divergence from the earliest eukaryotes around 2,500 Mya (Supplementary Fig. 2). Furthermore, in both humans and *Arabidopsis*, recent lineages tend to acquire new genes more rapidly than ancient lineages (25 g M^{-1}) compared to 4–7 g M^{-1} in human, while 34 g M^{-1} compared to 5–11 g M^{-1} in *Arabidopsis*)[89,90] (Supplementary Fig. 2).

These analyses also revealed that new gene generation mainly occurs through four mechanisms of DNA-based duplication, retroposition, exon shuffling and de novo origination, while less remains

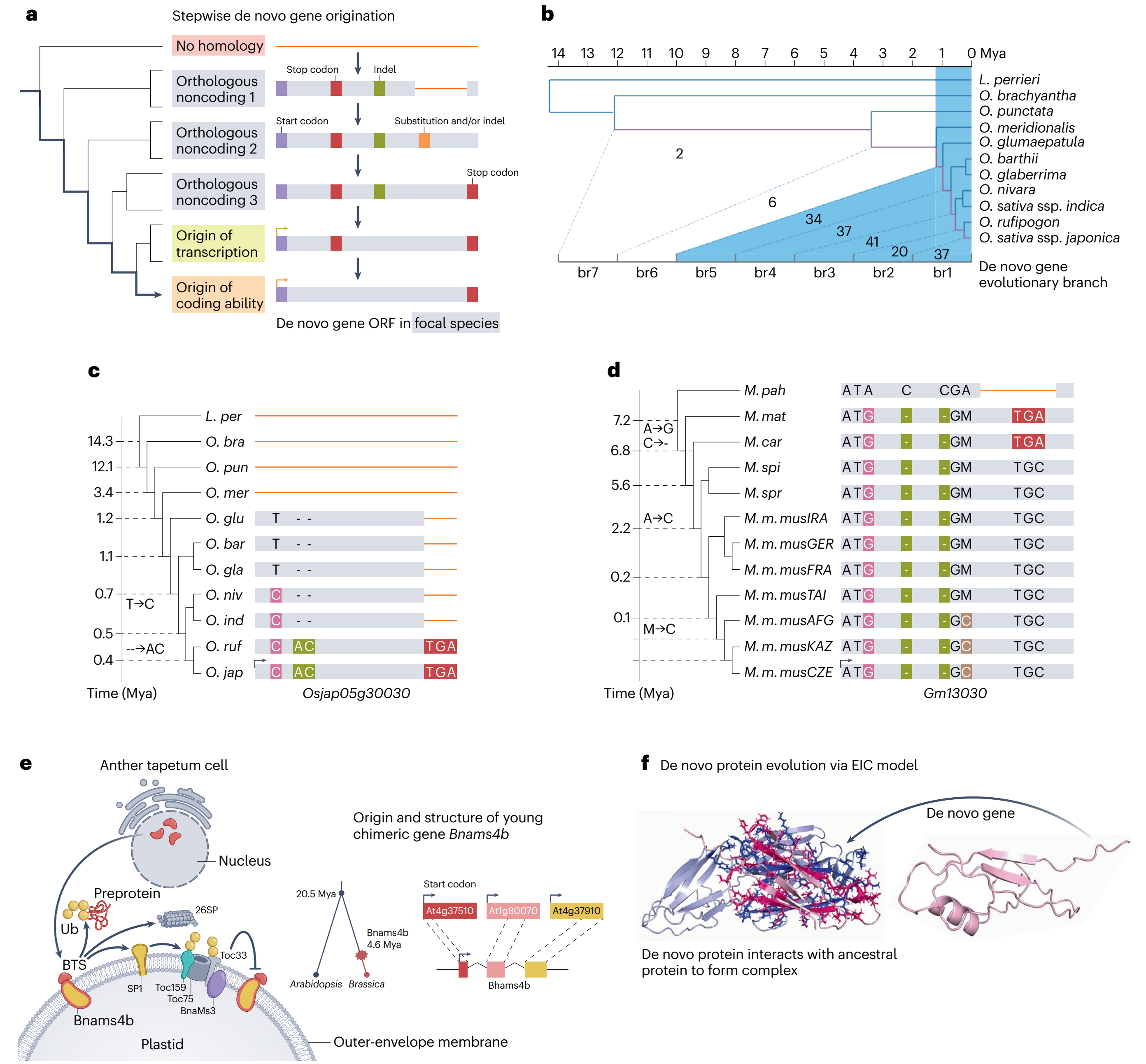


Fig. 2 | Stepwise origination of new genes and their eventual integration into protein interaction networks. a–d, Stepwise evolution of de novo genes for two examples from *Oryza* and mouse. **a**, The rapid stepwise origination of a de novo gene from a noncoding sequence to a coding ORF in less than several million years involves the gain of a start codon (purple), an indel (insertion or deletion mutation) event (green), followed by the activation of transcription (green arrow), translation (orange arrow), substitution (orange) and the removal of a premature stop codon (dark red)[56]. All de novo genes showed stepwise originations, as the examples show. **b**, A total of 175 de novo genes that originated in stepwise processes in less than 3.4 Mya in the recent ancestral lineages of *O. sativa* ssp. *japonica*[56] (the light pink indicates those that originated in <1.5 Mya). The divergence time has been retrieved from the TimeTree[170]. This scheme has been adapted from ref. 56, Springer Nature Limited. **c**, Example of stepwise origination processes for the de novo gene *Osjap05g30030* (ref. 56); here three key evolutionary steps promote the origination of de novo gene *Osjap05g30030* from ancestral noncoding intergenic sequences—the T→C substitution, the two-nucleotide 'AC' insertion and the gain of 'TGA' stop codon, with a recent origin of species-specific expression pattern in the most recent common ancestor of *O. sativa* ssp. *japonica* after its divergence from *O. rufipogon*. **d**, Exemplary stepwise origination process for the de novo mouse gene *Gm13030* (ref. 66). Here five evolutionary changes occurred in the origination of this gene indicated in color—gain of a start codon (purple), loss of two cytosine bases (green) and loss of two premature stop codons (orange)[66]. **e**,**f**, Two examples of the integration of new genes into protein interaction networks and protein complexes. **e**, The young chimeric gene *Bnams4b* (4.6 Mya) can interact with the nuclear-localized E3 ligase BRUTUS (BTS), which triggers translocation of BTS to reshape a new interaction network[115,116]. This scheme has been adapted with permission from ref. 116, Wiley-Blackwell. **f**, Illustrated here is the evolution of structural complexity in de novo proteins, which is observed in over 83% of rice stepwise de novo proteins. The example of stepwise de novo genes in rice illustrates the 3D structures of a de novo protein (right, *OSJAP01G39060*) and its immediate complexes (left)[121]. The image in **f** has been adapted with permission from ref. 121, Oxford Univ. Press. 26SP, 26S proteasome; TOC, translocon at the outer envelope membrane of chloroplasts; Ub, ubiquitin; *L. per*, *Leersia perrieri*; *O. bra*, *Oryza brachyantha*; *O. pun*, *Oryza punctata*; *O. mer*, *Oryza meridionalis*; *O. glu*, *Oryza glumaepatula*; *O. bar*, *Oryza barthii*; *O. gla*, *Oryza glaberrima*; *O. niv*, *Oryza nivara*; *O. ind*, *Oryza sativa* L. *indica*; *O. jap*; *O. sativa* ssp. *japonica*; *O. ruf*, *Oryza rufipogon*; *M.m.mus*, *Mus musculus*; *M. spr*, *Mus spretus*; *M. spi*, *Mus spicilegus*; *M. car*, *Mus caroli*; *M. mat*, *Mus matthewi*; *M. pah*, *Mus pahari*; AFG, Afghanistan; CZE, Czech Republic; FRA, France; GER, Germany; IRA, Iran; KAZ, Kazakhstan; TAI, Thailand.

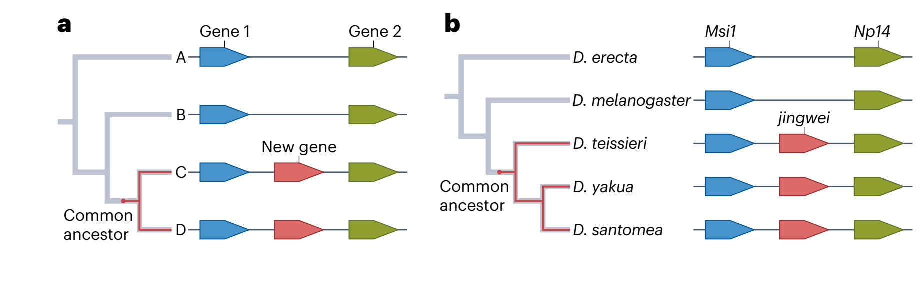
BOX 1

Methods to date gene ages

In general, the phylogenetic distribution of an orthologous gene provides age information for its common ancestor. In addition, the sequence divergence of the orthologous gene among the species cluster in the phylogeny at the DNA or protein level also offers age information. To distinguish between a gene gain to form its common ancestor or a gene loss in outgroup species, a parsimonious principle is used, in which a minimum number of evolutionary events are determined for the alternatives. The orthologous state of genes in the cluster is determined by the corresponding position in gene synteny maps of genomic sequences from high-quality, usually reciprocal, genomic alignment. In panel **a** of the figure, the new gene (red) is flanked by gene 1 (G1) and gene 2 (G2). The distribution of the red gene is observed in species C and D, but not in the two outgroup species A and B. The same color shows an orthologue in different species. Therefore, the common ancestor of the new gene appeared at a time after the ancestor diverged from the ancestor of C and D and before C and D separated. There are two conventional approaches to dating the ages of genes.

The first one was used in the determination of *jingwei*, the first new gene (illustrated in panel **b** of the figure)[1] identified and involves determining the phylogenetic distribution of orthologues of a gene, with its absence confirmed in outgroup species that have recently diverged[82]. The orthologous relationship across species groups is determined by a syntenic map surrounding the gene of interest, which is created by a high-quality reciprocal genome alignment[56]. The age of the common ancestor is taken as the age of the gene. In the case of *jingwei*, it has been known that *D. melanogaster* diverged 3 Mya from the clade of three African species, *Drosophila teissieri*, *Drosophila yakua* and *Drosophila santomea*. The common ancestor of *jingwei* appeared less than 3 Mya, thus *jingwei* is younger than 3 Mya. A genomic screen of a gene for paralogous sequences can also determine whether the gene originated from previously existing genetic material or from previously nonexisting material, for example, is a de novo gene. If a gene has a paralogous sequence, the sources of origination can be determined, such as gene duplication, TEs or genic recombination. If it has no paralogous sequence, noncoding sequences in the orthologous position in outgroup species can be examined to search for a potential de novo gene origination. This approach, often used to identify recently evolved genes in a short evolutionary time scale from a few Mya or several tens of Mya, is developed and used in the GenTree or its earlier prototype computational methods often used[84,88].

The second approach is phylostratigraphy[171,172]. It examines the species distribution of homologous genes to determine the ancestral founder gene that may have started the gene lineage over a long evolutionary time scale (often going back to the early stages of life). In panel **a** of the figure, this method requires the entire genome to contain no homologous genes to assign a common ancestor. This method is actually expected to detect orphan genes as founder genes, instead of de novo genes, because the multiple origination mechanisms can create orphan genes[171]. This method is often used to detect the genes that originated and evolved on a long evolutionary time scale, for example, the vertebrate lineages (400–700 Mya) or other lineages of life that appeared billion years ago. The statistical uncertainty of phylostratigraphy was reported to be due to the challenge of detecting distant homologs through sequence similarity[173]. Furthermore, the failure to detect homology proposed that the detectable sequence identity among homologous genes might disappear at certain evolutionary times with a molecular clock with a gene[173,174]. Nevertheless, further methodological examinations confirmed the value of the method in detecting patterns and candidates of orphan genes[68,175].



known about the extent of origination rate through most of the other mechanisms. These data also raise a new question of how genomes evolve in a dynamic process of gene gain and loss. Gene loss in birds also suggests adaptation by eliminating unnecessary or harmful genes[91].

New genes with detectable molecular functions and fertility effects

It is generally thought that new genes do not have important functions[92,93] and that the genes that are essential for the viability or development of an organism are those of ancient origin[5,92,94]. However, in the last decade, there have been many functional case studies demonstrating that many new genes actually have important roles in basic biological processes and structures, from embryonic development to neuronal organization in the brain (Table 1 and Supplementary Table 3).

We focus here on those examples with information on the functional roles of new genes in major biological processes and structures across various taxa. One such example is *Cocoon*, which was generated through retroposition 6 Mya in the *D. melanogaster*–*D. simulans* clade and was found to have an essential role in the enclosure and the proper development of the leg joints[95] (Table 1). The olfactory receptor gene *Or67a* has been detected to have duplicated seven times within the *D. melanogaster*–*D. suzukii* group species within 15 My and is co-expressed in sensory neuron populations that project to antennal lobe glomeruli[96]. Additionally, the members of the rapidly evolving *ZAD-ZNF* gene family are frequently critical for development and fertility[97]. A null mutation of the 30-million-year-old *ZAD-ZNF* gene *Nicknack* results in lethality, as it prevents *Drosophila* L1 larvae from molting into the L2 stage[97]. Indeed, 27 recent duplicates in this *Drosophila* gene family were orthologous to the mosquito gene

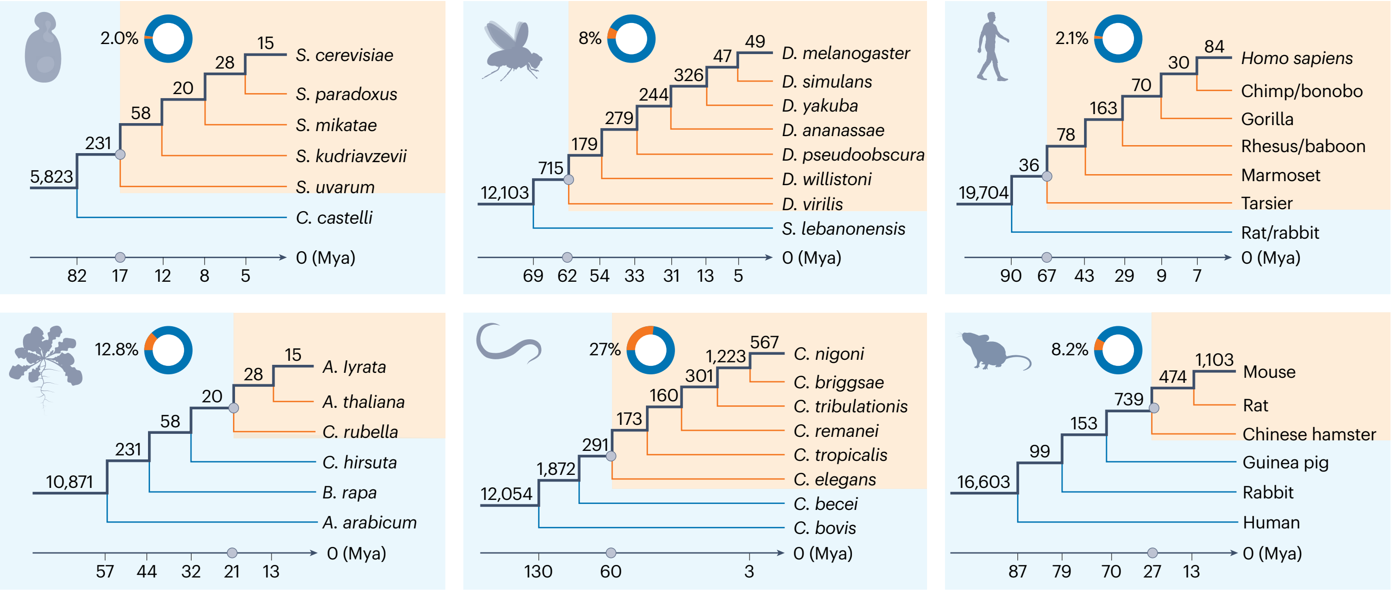


Fig. 3 | Overview of new gene origination in humans and several other species. Illustrated here are the percentages of new genes within specific lineages over a relatively short evolutionary time scale (which may vary) of less than 100 Mya. The pink shaded area represents lineage-specific new genes. The numbers in green represent the gene numbers originating during each of their evolutionary time scales. The gray circle represents the approximate time when the evolution of each lineage began. Schematics drawn from published data of *S. cerevisiae*[83], *D. melanogaster*[84], *H. sapiens*[85], *A. lyrata*[86], *C. nigoni*[87] and *M. musculus*[88] with the published time scales provided mainly by the TimeTree[170]. *S. paradoxus, Saccharomyces paradoxus*; *S. mikatae*, *Saccharomyces mikatae*; *S. kudriavzevii*, *Saccharomyces kudriavzevii*; *S. uvarum*, *Saccharomyces uvarum*; *C. castelli*, *Caenorhabditis castelli*; *D. simulans*, *Drosophila simulans*; *D. yakuba*, *Drosophila yakuba*; *D. ananassae*, *Drosophila ananassae*; *D. pseudoobscura*, *Drosophila pseudoobscura*; *D. willistoni*, *Drosophila willistoni*; *D. virilis*, *Drosophila virilis*; *S. lebanonensis*, *Scaptodrosophila lebanonensis*; *C. hirsuta*, *Cardamine hirsuta*; *A. arabicum*, *Adenium arabicum*; *C. briggsae*, *Caenorhabditis briggsae*; *C. tribulationis*, *Caenorhabditis tribulationis*; *C. remanei*, *Caenorhabditis remanei*; *C. tropicalis*; *Caenorhabditis tropicalis*; *C. becei*, *Caenorhabditis becei*; *C. bovis*, *Corynebacterium bovis*.

cucoid that determines the embryo polarity in early development[98]. As an eutherian-specific gene, the human *CATACOMB* (<125 Mya) was shown to antagonize epigenetic modification of H3K27me2/H3K27me3, which may have a role in placental development[99]. In *A. thaliana*, two species-specific duplicates from partial and tandem duplication evolved remarkable morphological traits throughout the developmental process[100,101]. The tandem duplicates, *AT5G12950* and *AT5G12960*, that originated 16 Mya acquired new phenotypic effects to change flowering time through neofunctionalization[100]. The partial duplicate, *EXOV*, that appeared 3.5 Mya evolved with a selection-driven sequence change into a major effector gene that affects morphology and development[101].

New genes with the evolution of functional essentiality

The relationship between gene age and functional essentiality is a complex and ongoing area of research. It was shown that, contrary to the belief that gene essentiality was conserved[5], sequences of essential genes can evolve rapidly under positive selection (Supplementary Fig. 3a). In *Drosophila*, an unexpected pattern was observed in knockdown experiments that silenced a large number of new genes, in that the probability of a gene evolving an essential function was shown to be independent of its evolutionary age, with about 30% of new genes causing organismal lethality when they were constitutively silenced[102]. A similar proportion (25–28%) was also reported to be functionally indispensable in the young duplicate members of the *ZAD-ZNF* families[97]. The detected essentiality of new genes is likely an underestimate due to lower knockdown efficiency, leading to a much higher false-negative than false-positive rate[103]. A whole-genome knockdown experiment in *D. melanogaster* generated an independent dataset to screen for genes required for intestinal stem cell maintenance and differentiation[104]. From this dataset, new genes in various ancestral branches across the entire *Drosophila* lineage since divergence from the most recent common ancestor (MRCA) had a similar proportion of lethality (Supplementary Fig. 3b). Finally, independent functional analyses of young *Drosophila* genes identified critical germline functions, such as *COX4L*, a nuclear mitochondrial duplicate (63 Mya), and *Zeus*, a young gene that originated through retroposition (3 Mya), that revealed indispensable effects on male fertility[105–107]. These experiments provided evidence that new genes can evolve essential functions quickly.

While the abovementioned analyses demonstrate that new genes can evolve quickly in *Drosophila*, in other organisms, the gain of important functions of new genes might be a more gradual process[66,87]. Other reports have argued that newly duplicated genes provide genetic robustness, as deletions of one paralog yield only mild effects, whereas only double deletions have substantial effects in *S. cerevisiae*, as well as may also be critical for stress responses in *A. thaliana*[108]. A recent study in *C. elegans* demonstrates that new genes exhibit an overall lower percentage of essential genes compared to older genes[87].

Although previous findings in *Drosophila* have shown that new genes can evolve important functions rapidly[102,104,106,109], little is known about how this occurs. A recent functional study of *Apollo* (*Apl*) and *Artemis* (*Arts*), a young pair of tandem gene duplicates as homologs of importins, which arose 200,000 years ago, provided insight into the evolution of essential gametogenesis functions through sexually antagonistic selection[109,110]. A paralog-specific CRISPR–Cas9 experiment revealed that *Apl* and *Arts* were essential for male and female fertility, respectively, in the studied population[111]. A sexual conflict drive model was proposed to interpret the rapid evolution of these genes in a rapidly evolving germline environment[110]. In vitro biochemical assays and in vivo mutational analysis revealed that *Apl* and *Arts* evolved divergent molecular functions, with *Apl* being involved in the simultaneous deposition of a protamine-like protein, Mst77F, on DNA and the dissociation of histone–DNA complexes in germline cells, whereas *Arts* may regulate nuclear transport of essential components of actin networks in the embryo[111].

Table 1 | Examples of new genes with documented phenotypes and functions

Species	Mechanism	Gene name[a]	Ages (million years)	Phenotypes and functions
Development and morphogenesis				
Drosophila	Retrotransposition	*Cocoon*	<6	Knockdown causes fused leg joint
Drosophila	Retrotransposition	*Desr*	<25	Foraging ability related
Drosophila	DNA duplication	*ms(3)K81*	<44	Zygote viability
Drosophila	DNA duplication	*Umbrea*	<11	Essential centromere function
Drosophila	Duplication	*bicoid*	<150	Patterning the anterior embryo
Drosophila	Duplication	*ZAD-ZNF* gene	<40	Heterochromatin functions
Arabidopsis	Partial gene duplication	*EXOV*	<3.5	Multiple morphological traits
Arabidopsis	De novo	*sORF2146*	<8.9	Controls floral transition
Chironomus	DNA duplication	*panish*	<180	Embryo polarity
Zebrafish	Retrotransposition	*chirons*	<4	Regulating NAD^+ levels
Eutherians	Pseudogenization	*XIST*	<160	X-chromosome inactivation
H. sapiens	Retroviral genes	*Syncytin 1 (ERVW1)*	<160	Human placental morphogenesis
Placental mammals	Gene duplication	*INSL4*	<45	Define the relaxin family repertoire
Dogs	Retrotransposition	*FGF4*	~0.01	Breed-defining chondrodysplasia
Salamander	Orphan gene	*Prod1*	<151	Preaxial digit formation
O. sativa	Orphan gene	*JAUP1*	<3	Regulates jasmonate biosynthesis and signaling to promote root development
Adaptation				
Drosophila	Exon shuffling	*jingwei*	<3	Produced a new dehydrogenase
Drosophila	DNA-based duplication	*Adh-Finnega*	<3	Influence the efficiency of alcohol metabolism
Arabidopsis	Retrotransposition	*CYP98A8/CYP98A9*	<28	Phenolic pathway
Arabidopsis	Gene duplication	*CYP84A4*	<8	Pathway for α-pyrone biosynthesis
Arabidopsis	Orphan gene	*QQS*	<39	Starch metabolic network
Arabidopsis	Orphan gene	*AtEWR*	<34	Resistance to vascular wilt pathogens
Brassica rapa	Orphan gene	*BrOG1*	<20	Affect soluble sugar metabolism
Potato	Exon shuffling	*GapC*	<60	The mitochondrial targeting function
Human	Segmental duplications	*TCAF1/TCAF2*	<7	Antagonistically regulate the cold-sensor protein TRPM8
Owl monkey	Retrotransposition	*TRIMCyp*	<43	Resistance to HIV-1
Acorn barnacles	Retrotransposition	*bcs-6*	<1	Adapt to a sessile lifestyle
Codfish	De novo gene	AFGP	<15	Antifreeze adaption
Winter flounder	Partially de novo	GIG2	<8	Antifreeze adaption
Brain and nervous system				
Drosophila	Retrotransposition	*Xcbp1*	<6	Participate in the foraging circuit
H. sapiens	Segmental duplication	*SRGAP2B/SRGAP2C/SRGAP2D*	1.0–3.4	Modulate SRGAP2A-dependent synaptic development
H. sapiens	Segmental duplication	*ARHGAP11B*	<5	Promotes basal progenitor amplification and neocortex expansion
H. sapiens	Segmental duplication	*NOTCH2NLA, NOTCH2NLB, NOTCH2NLC*	<4	Affect notch signaling and cortical neurogenesis
H. sapiens	De novo	*ENSG00000205704*	<7	Unique human brain developmental functionality
Mammalian	Gene duplication	Olfactory receptor genes	<90	Olfactory receptor
Primate	Tandem duplication	*KRAB–ZNF* family	35–40	Influence human neuronal differentiation
Primate	Gene duplication	*Primate opsin genes*	<60	Trichromatic vision
Speciation				
Drosophila	Gene duplication	*Odysseus (OdsH)*	<40	Altered heterochromatin binding involving hybrid sterility
Drosophila	Transposed duplication	*JYAlpha*	<3	Hybrid sterility in *Drosophila*

Table 1 (continued) | Examples of new genes with documented phenotypes and functions

Species	Mechanism	Gene name[a]	Ages (million years)	Phenotypes and functions
Rice	Orphan genes	*iORF3* and *iORF4*	<3	Confers reproductive isolation in rice
Rice	Transposon-mediated duplicated	*HWS1* and *HWS2*	<3	Causes segregation distortion
Monkeyflowers	Partial duplication	*YELLOW UPPER(YUP)*	<15	Ecological speciation in monkeyflower
Platyfish	Chimeric	*Xmrk-2*	<3	Linked to melanoma-determining Tu loci
Reproduction and sexuality				
Drosophila	DNA duplication	*Apl*	0.02	Resolving sexual conflict
Drosophila	DNA duplication	*nsr*	<6	Regulating Y-linked male fertility genes
Drosophila	DNA duplication	*Sflc*	<11	Survival and reproduction
Drosophila	Putative de novo	*Goddard, Saturn, Atlas*	<40	Essential in spermatogenesis
Drosophila	DNA duplication	*Sdic* family	<3	Contributes to the differential reproductive success
Drosophila	DNA duplication	*p24-2*	<3	Development and reproduction
Drosophila	Retrotransposition	*Zeus*	<5	Male reproduction
Drosophila	Retrotransposition	*Pros28.1A*	<11	Male-specific functions
Drosophila	Retrotransposition	*Poseidon*	<62	Compensates for meiotic X chromosomal inactivation
Drosophila	Retrotransposition	*Sphinx*	<3	Male courtship
Diamondback moth	Orphan gene	*Tssor-3* and *Tssor-4*	<15	Male reproductive regulation
Mouse	De novo	*POLDI/Pldi*	2.5–3.5	Sperm differentiation
B. napus	DNA duplication/chimeric	*Bnams4b*	4.6	Anther tapetum development
Therian mammals	Gene duplication	*INSL3*	<160	Testicular descent

[a] See the corresponding reference of each gene in Supplementary Table 3.

New genes integrate into and change existing gene networks

It may be expected that any new gene or protein will also integrate into the overall cellular gene or protein network and be less likely to act in complete isolation. Indeed, statistical and network analyses have predicted interactions between new and pre-existing genes, but little is known about how the interactions occur. New genes in vertebrates (human genome) and invertebrates (*Drosophila* and *C. elegans*) were detected to integrate into pre-existing expression networks with faster rates of new connection incorporation in invertebrates than humans[67,112,113]. One example is the RNA-based duplication from *Caf40*, a key component of the CCR4–NOT deadenylase complex[105,107,114] into *Zeus*, which resides in the *D. melanogaster*–*D. simulans* clade that diverged 3 Mya, and *Poseidon*, found in the *Sophophora* subgenus species that diverged 40 Mya[105,107,114]. Co-immunoprecipitation assays showed that Poseidon proteins bind to NOT1 of CCR4–NOT1, whereas *Zeus* shows little to no detectable binding. Assays to measure mRNA repression revealed that *Poseidon* is capable of degrading a reporter mRNA, suggesting that it retained similar functions to its X-linked parental copy, *Caf40* (ref. 105). However, unlike *Poseidon*, *Zeus* seems to have acquired new functions, as a mutation in this gene decreased male fertility up to 80%[107]. Zeus also was shown to bind and regulate the expression of 193 genes[107] and is experiencing rapid protein sequence evolution under positive selection[114]. Another example is the *Bnams4b* gene, which is involved in anther tapetum development[115,116]. The young chimeric protein (4.6 Mya) encoded by *Bnams4b* can interact with the nuclear-localized E3 ligase BRUTUS (BTS) as confirmed by yeast-two-hybrid, pull-down, bimolecular fluorescence complementation and co-immunoprecipitation assays[116]. This indicates that Bnams4b triggers translocation of BTS protein to reshape a new interaction network[116] (Fig. 2e). Younger de novo proteins in rice were found to form new gene networks by forming new protein complexes with old unrelated proteins (Fig. 2f). Over 83% of de novo proteins in rice were observed to form new protein complexes[117]. These observations reveal that new genes usually form interactions with old genes to carry gene functions, rather than function in isolation.

Patterns of new gene evolution

The large numbers of new genes in various organisms, especially those new genes detected in model species, provide opportunities to characterize patterns of new gene evolution and their statistical features. These patterns reflect the underlying evolutionary forces that occur at various layers, including the structure of genes and proteins (Fig. 4a), cellular and tissue expression, gene function and interaction, locations of topologically associating domains (TADs, self-interacting genomic regions) and coding sequence length. Below, we discuss such well-described and recently reported patterns[82] (Fig. 4).

The rapid evolution of the nascent structure of de novo protein

How the structure of de novo proteins evolves is an interesting problem that has been debated[118–120]. Various structures that are simpler than well-folded proteins in isolation have been observed in de novo genes in *Oryza* that originated within the last million years[121], whereas the older de novo genes in this species more frequently exhibit more complex and well-folded structures[121], as well as form protein complexes with other old proteins[121]. In fact, *Oryza* de novo genes showed a steady evolution of structure that is accompanied by increasing complexity in a relatively short time of 2.4 million years[121] (Fig. 4a). However, such linear progress in the evolution of structural complexity was not observed in putative de novo genes in *Drosophila*[70], which showed a higher degree of intrinsic disorder[70,122]. Observations based on analyzing the average Grantham distances between residues (a proxy for their evolutionary distance) within each gene age stratum indicate that substitutions in young genes are more likely to occur between less biochemically similar residues, implying that these substitutions have larger physicochemical impacts compared to those in older genes[117].

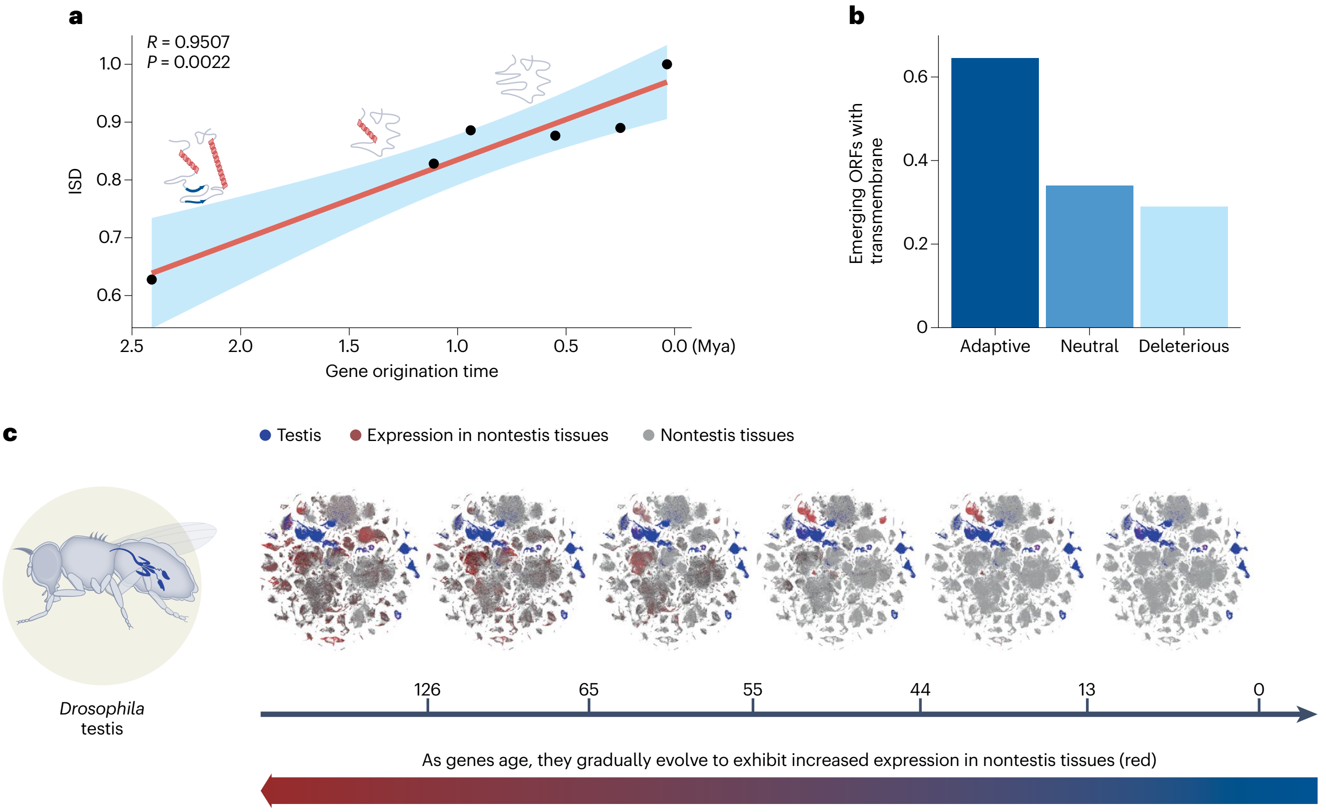


Fig. 4 | Evolutionary patterns of new genes at molecular and cellular levels. As three examples illustrated here, new genes exhibit a pattern in evolution in terms of their ages with respect to protein structure, cellular structure and cell population. **a**, Tertiary structure of de novo proteins in *Oryza*. Linear regression analysis shows a substantial decrease in ISD as the evolutionary ages of de novo genes increase in a short evolutionary period of less than 2.5 Mya[121], suggesting rapid evolution toward increasing stable and distinct domain structures. The shaded area represents the 95% confidence interval. The image in **a** has been adapted with permission from ref. 121, Oxford Univ. Press. **b**, Adaptive de novo sequences, which are evolutionarily very young, tend to exhibit a pattern of encoding putative transmembrane domains in yeasts[134]. The image in **b** has been adapted from ref. 134, Springer Nature Limited. **c**, 'Out of testis' expression pattern of new genes at the single-cell level in *Drosophila*. As genes become older, they gradually accumulate expression in nontestis tissues (nontestis cells shown in red). Gray dots indicate all ~570,000 cells from 15 tissues. Blue dots indicate ~44,621 cells from the testis. Red dots indicate the expression of ten represented genes in each branch in other tissues except testis[128]. The image in **c** has been designed with permission from ref. 128, AAAS, by using the published single-cell database. ISD, intrinsic structural disorder.

Increasing structural complexity of de novo genes in evolution

De novo genes in *Oryza* have evolved quickly (that is, in a few million years) with regard to gene size, exon number, exon size and protein size. For example, the de novo genes in *Oyrza* emerged as small ORFs that gradually increased in length with evolutionary age[56]. Such patterns of increasing sizes of structural elements in evolution were also observed in the putative de novo genes in *Drosophila* and humans[70,123]. These observations in metazoans and flowering plants suggest that de novo gene origination may be generally an incremental process in structure.

Expression and interaction patterns of new genes

The majority of new genes exhibit testis expression, as reported in gene traffic analyses in *Drosophila* and humans[124,125], giving rise to the 'out of testis' pattern[126,127], with their expression in other tissues increasing over time, as confirmed in both *Drosophila* and humans by recent single-cell transcription analyses or extensive RNA-sequencing data[128–131] (Fig. 4c). In accordance with this, new genes in *Oryza* and *Arabidopsis* are also often found to be more highly expressed in anthers and other male flower tissues, in much the same way that the testis produces sperm in metazoans[56], giving rise to the 'out of pollen' hypothesis[132,133]. Such a pattern of expression of new genes in male reproductive tissues suggests a potential link to the evolution of pleiotropy[85]. TADs in human and mouse genomes have been shown to align with clusters of genes that share similar evolutionary ages. At the subcellular level, most of the adaptive emerging de novo genes in *S. cerevisiae*, which originated from thymine-rich ancestral intergenic regions, were detected to encode putative transmembrane domains[134] (Fig. 4b). In *Drosophila* cells, younger genes are also observed to be expressed in fewer subcellular structures than old genes[84]. These results echo the pattern of protein–protein interactions increasing with gene age; specifically, as the divergence time between new and corresponding parental genes increases, the average connectivity in terms of the protein–protein interaction network gradually rises. These data revealed a time-dependent expression and protein interaction patterns of new genes.

Evolutionary forces acting on new genes

The above-given sections have reviewed molecular processes and consequences of new gene evolution, which are governed by evolutionary forces. These forces include conventionally expected natural selection for functional innovation and also recently detected sexually antagonistic selection. Various approaches were used to detect these evolutionary forces, as we will discuss below.

Functional and phenotypic evidence for adaptation and sexual conflicts

New functions conferred by new genes often help organisms to adapt to environmental changes, such as AFPs in polar fishes mentioned above[46,47], which helped them to survive oceanic glaciation[25,135].

In *Oryza*, new genes originated by gene fusion and exon recombination[32] have been experimentally shown to control traits, such as seed germination, shoot length and root length, reflecting drought adaptation[136]. However, adaptation is not the only force that drives new gene evolution, as exemplified by the aforementioned finding of two species-specific genes, *Apl* and *Arts*, in *D. melanogaster*, whose evolution was driven by sexual conflict[109,110]. Young de novo genes found to be expressed in the somatic accessory gland and testis reveal major differences between the two tissues in terms of gene abundance, expression level and *cis*-regulatory mechanism under positive selection[137].

Sequence evolution and molecular population genetics

Changes in the sequences of new genes throughout evolution also provide important information on the underlying evolutionary forces that can be detected by using molecular population genetic analyses. Nucleotide substitution analysis by comparing substitution rates in synonymous and nonsynonymous sites detected a prevalent role of positive selection in the evolution of most new genes[22,82]. Furthermore, the strength of positive selection can be determined by using molecular population genetics tools. For instance, phenotypic effects and underlying evolutionary forces have been analyzed in a number of new gene duplicates that originated in various stages of the evolution of the *Sophophora* subgenus toward *D. melanogaster* (3–40 Mya)[95,97]. A link between gene essentiality and positive selection was observed in seven of the nine duplicates that showed lethality or sterility with a high proportion of amino acid substitutions (measured as α values of 62–91%[97,138], the proportion of all amino acid substitutions changed by positive selection; Supplementary Fig. 3a). In addition, in de novo genes of *Oryza* and a species-specific duplicate gene of *A. thaliana*, positive selection was also detected in the earliest protogene stage and later stages by using population genetics approaches[139,140] and comparison of substitution rates between synonymous and nonsynonymous sites[56,100]. The use of these population genetics approaches detected strong positive selection on the putative de novo genes in *Drosophila*[54] and other types of new genes as well[102]. In addition, positive selection on young gene duplicates was also detected through analysis of linkage disequilibrium, reduced nucleotide diversity and haploid states in *Drosophila*[141,142].

Molecular population genomics

Systematic surveys of population variation across entire genomes can help to investigate the natural selection in the genetic variants that lead to the formation of new genes. An investigation in *Drosophila* focused on frequencies of copy number variants (CNVs) in northern and southern populations established that repeated patterns of latitudinal clinal variation affect duplicated genes in Australian and North American continents[143]. A duplication of *Ace* is associated with clinal differentiation across the continents[143]. These data suggest that the spatially varying selection may determine the evolutionary fate of these polymorphic duplicates.

Genome annotation, similar to environmental context, can also be used as a proxy for function. Population studies in *Drosophila* and humans concluded that among the different types of CNVs, complete gene duplications are the least prevalent, followed by intronic, and then by exonic duplications[144], which were shown to be a consequence of dosage sensitivity of CNVs[145]. One of the most powerful outcomes of polymorphism studies is the ability to use population data to estimate evolutionary parameters, such as the distribution of fitness effects, and this approach has been used in a few early CNV studies in *Drosophila* and humans[146,147]. These studies applied the Poisson random-field model to the histogram of allele frequencies to estimate the distribution of fitness effects of duplications[148,149]. Both *Drosophila* and human duplicate polymorphisms experience purifying selection in general and positive selection, as well in duplicates of seven *Drosophila* genes as shown in high frequencies and functional resistance to dichlorodiphenyltrichloroethane (DTT) and toxins[146,147,150]. The young duplicates (0.3–0.8 Mya) of the salivary amylase gene *AMY1* in human populations were found to appear to help humans adapt to the increasing amount of starch in their daily diets[151,152]. Remarkably, the average number of *AMY1* duplicates in Eurasian populations increased from four to more than seven since agricultural civilization started in the Eurasian continent 12,000 years ago, when starch-rich crops such as rice and wheat were domesticated[152]. These studies quantitatively measured and detected the joint forces of weak selection and genetic drift on duplicate variants (Supplementary Fig. 4).

Studies of polymorphism can be combined with an analysis of divergence between species to determine the rate of fixation of duplicates between species, from which the rate of adaptive substitution of mutations can be estimated[138,153]. Applying such an approach to *Drosophila*, it has been shown that genomic sites composed of duplications that overlap the boundaries of TADs of self-interacting genomic regions exhibit a greater divergence between species, suggesting a role of adaptation on the genomic sites[144].

The application of new genes in agriculture and medicine

Numerous new genes have been reported in various crop plants and analyzed for their role in their genetic improvement. Moreover, new human genes were also found to be associated with disease phenotypes and, in some cases, could be correlated with the molecular mechanisms underlying oncogenesis. Furthermore, the rise of agricultural civilization in crop plant domestications also changed the genetic structures of human populations in the use of diet components[151,152].

It has been shown that evolutionarily young genes can directly serve as primary sources of agricultural trait innovation and divergence in crop breeding. For example, the full ORF of de novo gene *GSE9*, functioning as a regulator of cell proliferation and cell expansion in rice, originated at 0.3 Mya, just before the divergence between *Oryza rufipogon* and *Oryza sativa* ssp. *japonica*[154]. Indeed, deleting *GSE9* in the *japonica* variety Zhonghua 11 resulted in longer and narrower grains, whereas its overexpression substantially increased the 1,000-grain weight[154]. The *Zea* genus-specific micropeptide RPG, which originated de novo from a noncoding sequence after the divergence between the genera *Zea* and *Tripsacum* approximately 0.65 Mya, reduces the kernel dehydration rate in maize[155]. Furthermore, the male sterile gene, *Bnams4b*, used in the hybrid development of *Brassica napus* for hybrid vigor usage is a typical chimeric gene originating via exon shuffling ~4.6 Mya[115,156]. It encodes a long chimeric protein consisting of a plastid signal peptide, partial spliceosome sequence and heat shock protein domain[115]. The *A. thaliana* gene *QQS* is a species-specific orphan gene[57]. Ectopic expression of this gene in soybean leads to decreased levels of starch in the leaf and increased leaf protein content, demonstrating that even species-specific young genes can have the potential for genetic improvement in another crop species. Finally, the rice young genes such as *OsDR10* (ref. 55), *JAUP1* (ref. 157), *Xa7* (ref.158), *OsPHT3* (ref. 159), *OsPDX3* (ref. 160) and *Xio1* (ref. 161) are all involved in multi-stress tolerance, adding genetic sources of crop plant breeding.

New genes are associated with a number of conditions, in which their mutational disruptions were found to lead to neurological or cognitive disorders, cancer or reproductive diseases[139]. New genes are particularly interesting from a human disease perspective because they often originated in a lineage-specific manner and are only expressed in certain specific tissues and cell types[139]. Examining ~4,000 Mendelian disease genes with biomedically relevant phenotypes revealed that new genes steadily integrate at a rate of ~0.07% per million years into the human genome over evolutionary time[85]. New genes exhibit accelerated sexual selection and human-specific adaptive innovations due to lower pleiotropy, whereas older genes are under stronger selective constraints because of their higher pleiotropic burden that impacts a greater number of anatomical systems[85]. For instance, increased pathogenic virulence by *Candida* in humans is found to be a consequence of recently expanding duplicate copies of the adhesin gene that were divergent under positive selection[162].

A list of de novo genes in humans has been compiled with a re-evaluation of their role in diseases[163]. In total, 39 out of 82 de novo genes have been reported to be associated with various cancers (31), Parkinson's disease (2), Alzheimer's disease (1), schizophrenia (1), reproduction (2), ulcerative colitis (1) and type 2 diabetes (1)[164]. The human-specific, partial duplicated gene *NOTCH2NL* has been shown to affect notch signaling and cortical neurogenesis[163,165]. *ARHGAP11B*, another human-specific gene, is critical for promoting basal progenitor amplification and neocortex expansion; its overexpression in mice results in neocortex expansion and increased memory flexibility[166]. In addition, reproductive disorders were reported to be caused by genetic disruption of new genes[85], and young genes with mutations more frequently damage reproductive organs than other tissues in humans[85]. Furthermore, molecular biochemistry research and structural characterization of the de novo gene *NCYM* as an antisense transcript of *MYCN* oncogene pointed to the potential of this young gene as a drug target for cancer treatment[167,168].

Conclusions and perspectives

Genes have been conventionally viewed as static and conservative genetic elements in the genomes of organisms, but this view is beginning to change. As discussed here, it is now clear that new genes have been continuously emerging and changing genomes, leading to present-day organisms. Besides the classic and better-understood mechanism of gene duplication, either of individual genes or genomes, a number of additional distinct mechanisms were found to also have important roles in shaping the initial structure of new genes. Despite being thought of as unlikely, other routes of de novo gene origination have been detected that generate new protein-coding genes out of noncoding ancestral sequences. Furthermore, new genes were shown to have evolved various functions in various biological processes and structures with peculiar patterns.

However, much more remains to be determined and provides rich opportunities for further scientific exploration. Most studies have focused on a few mechanisms of new gene generation, such as gene duplication and de novo gene origination, whereas the remaining ten mechanisms (Fig. 1), from TE domestication to gene fission, await further exploration with regard to their mechanistic details and frequencies; it is also likely that even more thus far unknown mechanisms will be discovered in the future. In addition, variations in the rate of new gene generation have been studied to a lesser extent, despite the fact that the basis of these patterns could lead to new insights. The description of the molecular functions and phenotypic effects of new genes has already led to a new understanding of evolutionary innovations and may add new conceptual and technical research problems to molecular biology[169]. Therefore, we anticipate that efforts to understand the many aspects of new genes discussed here will increase in the near future and might result in new concepts and help solve outstanding issues in the field that range from defining the origin of species to what makes *H. sapiens* human, by incorporating the effect of gene age. In addition, our knowledge of evolutionarily new genes will likely also bring benefits for translational studies in medicine and agriculture.

References

1. Long, M. Y. & Langley, C. H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
2. Muller, H. J. Bar duplication. *Science* **83**, 528–530 (1936).
3. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
4. Gilbert, W. Why genes in pieces. *Nature* **271**, 501 (1978).
5. Ashburner, M. et al. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. *Genetics* **153**, 179–219 (1999).
6. Lewin, B., Krebs, J., Kilpatrick, S. T. & Goldstein, E. S. *Lewin's Genes X* (Jones & Bartlett Learning, 2011).
7. Watson, J. et al. *Molecular Biology of the Gene* Vol. 364 (John Inglis, 2014).
8. Zhang, J. M., Dean, A. M., Brunet, F. & Long, M. Y. Evolving protein functional diversity in new genes of *Drosophila*. *Proc. Natl Acad. Sci. USA* **101**, 16246–16250 (2004).
9. Gilbert, W. The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 901–905 (1987).
10. Brosius, J. Retroposons—seeds of evolution. *Science* **251**, 753 (1991).
11. Muller, H. J. Pilgrim trust lecture—the gene. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **134**, 1–37 (1947).
12. Ohno, S. *Evolution by Gene Duplication* (Springer Science & Business Media, 2013).
13. Freeman, V. J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol.* **61**, 675–688 (1951).
14. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
15. Syvanen, M. Cross-species gene transfer; implications for a new theory of evolution. *J. Theor. Biol.* **112**, 333–343 (1985).
16. Leffers, H., Gropp, F., Lottspeich, F., Zillig, W. & Garrett, R. A. Sequence, organization, transcription and evolution of RNA polymerase subunit genes from the archaebacterial extreme halophiles *Halobacterium halobium* and *Halococcus morrhuae*. *J. Mol. Biol.* **206**, 1–17 (1989).
17. Riley, M. & Labedan, B. Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* **268**, 857–868 (1997).
18. Xu, W. et al. Evolution of Yin and Yang isoforms of a chromatin remodeling subunit precedes the creation of two genes. *eLife* **8**, e48119 (2019).
19. Roth, J. R. Frameshift mutations. *Annu. Rev. Genet.* **8**, 319–346 (1974).
20. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).
21. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
22. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).
23. Alba, M. M. & Castresana, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
24. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
25. Zhuang, X., Yang, C., Murphy, K. R. & Cheng, C.-H. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl Acad. Sci. USA* **116**, 4400–4405 (2019).
26. DeVries, A. L. Biological antifreeze agents in coldwater fishes. *Comp. Biochem. Physiol. A Physiol.* **73**, 627–640 (1982).
27. Makałowski, 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).
28. Beck, C. & Warren, R. Divergent promoters, a common form of gene organization. *Microbiol. Rev.* **52**, 318–326 (1988).
29. Jacq, C., Miller, J. & Brownlee, G. A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* **12**, 109–120 (1977).

30. Holley, R. W. et al. Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).
31. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
32. Wang, W. et al. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**, 1791–1802 (2006).
33. Rogers, R. L. & Hartl, D. L. Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. *Mol. Biol. Evol.* **29**, 517–529 (2012).
34. Fang, C., Gan, X., Zhang, C. & He, S. The new chimeric chiron genes evolved essential roles in zebrafish embryonic development by regulating NAD$^+$ levels. *Sci. China Life Sci.* **64**, 1929–1948 (2021).
35. Hanada, K. et al. The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. *Plant Cell* **21**, 25–38 (2009).
36. Casola, C. & Betrán, E. The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome Biol. Evol.* **9**, 1351–1373 (2017).
37. Zhang, J. & Zhou, Q. On the regulatory evolution of new genes throughout their life history. *Mol. Biol. Evol.* **36**, 15–27 (2019).
38. Su, Q., He, H. & Zhou, Q. On the origin and evolution of *Drosophila* new genes during spermatogenesis. *Genes* **12**, 1796 (2021).
39. Makałowski, W., Gotea, V., Pande, A. & Makałowska, I. *Transposable Elements: Classification, Identification, and Their Use as a Tool for Comparative Genomics* (Springer, 2019).
40. Bergthorsson, U., Andersson, D. I. & Roth, J. R. Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl Acad. Sci. USA* **104**, 17004–17009 (2007).
41. Lee, U. et al. The 3-dimensional genome drives the evolution of asymmetric gene duplicates via enhancer capture-divergence. *Sci. Adv.* **10**, eadn6625 (2024).
42. Uz-Zaman, M. H., D'Alton, S., Barrick, J. E. & Ochman, H. Promoter recruitment drives the emergence of proto-genes in a long-term evolution experiment with *Escherichia coli*. *PLoS Biol.* **22**, e3002418 (2024).
43. Begun, D. J., Lindfors, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172**, 1675–1681 (2006).
44. Cai, J., Zhao, R. P., Jiang, H. F. & Wang, W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–496 (2008).
45. Ohno, S. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proc. Natl Acad. Sci. USA* **81**, 2421–2425 (1984).
46. Cheng, C.-H. C. & Chen, L. Evolution of an antifreeze glycoprotein. *Nature* **401**, 443–444 (1999).
47. Chen, L., DeVries, A. L. & Cheng, C.-H. C. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl Acad. Sci. USA* **94**, 3811–3816 (1997).
48. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
49. Heinen, T., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**, 1527–1531 (2009).
50. Murphy, D. N. & McLysaght, A. De novo origin of protein-coding genes in murine rodents. *PLoS ONE* **7**, e48650 (2012).
51. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**, 1–13 (2013).
52. Ekman, D. & Elofsson, A. Identifying and quantifying orphan protein sequences in fungi. *J. Mol. Biol.* **396**, 396–405 (2010).
53. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2007).
54. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
55. Xiao, W. F. et al. Arice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE* **4**, e4603 (2009).
56. Zhang, L. et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* **3**, 679–690 (2019).
57. Li, L. et al. Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves. *Plant J.* **58**, 485–498 (2009).
58. Li, Z. W. et al. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biol. Evol.* **8**, 2190–2202 (2016).
59. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759 (2009).
60. Wu, D. D., Irwin, D. M. & Zhang, Y. P. De novo origin of human protein-coding genes. *PLoS Genet.* **7**, 1002379 (2011).
61. Ingolia, N. T. et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8**, 1365–1379 (2014).
62. Prabh, N. & Rödelsperger, C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics* **17**, 226 (2016).
63. Schlötterer, C. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
64. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
65. Prabh, N. & Rodelsperger, C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes. *G3 (Bethesda)* **9**, 2277–2286 (2019).
66. Xie, C. et al. A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife* **8**, e44392 (2019).
67. Zhang, W. Y., Gao, Y. X., Long, M. Y. & Shen, B. R. Origination and evolution of orphan genes and *de novo* genes in the genome of *Caenorhabditis elegans*. *Sci. China Life Sci.* **62**, 579–593 (2019).
68. Vakirlis, N., Carvunis, A. R. & McLysaght, A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500 (2020).
69. Levy, A. How evolution builds genes from scratch. *Nature* **574**, 314–317 (2019).
70. Peng, J. & Zhao, L. The origin and structural evolution of de novo genes in *Drosophila*. *Nat. Commun.* **15**, 810 (2024).
71. Jin, G. H. et al. New genes interacted with recent whole-genome duplicates in the fast stem growth of bamboos. *Mol. Biol. Evol.* **38**, 5752–5768 (2021).
72. Wolfe, K. H. & Li, W. H. Molecular evolution meets the genomics revolution. *Nat. Genet.* **33**, 255–265 (2003).
73. Dennis, M. Y. & Eichler, E. E. Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* **41**, 44–52 (2016).
74. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
75. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
76. Li, Z. et al. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28**, 326–344 (2016).
77. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* **9**, 938–950 (2008).
78. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
79. Guijarro-Clarke, C., Holland, P. W. & Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.* **4**, 519–523 (2020).

80. Chen, S. D., Krinsky, B. H. & Long, M. Y. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **14**, 645–660 (2013).
81. Begun, D. J. Origin and evolution of a new gene descended from alcohol dehydrogenase in *Drosophila*. *Genetics* **145**, 375–382 (1997).
82. Long, M. Y., VanKuren, N. W., Chen, S. D. & Vibranovski, M. D. New gene evolution: little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
83. Vakirlis, N. et al. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).
84. Dong, C. et al. New gene evolution with subcellular expression patterns detected in PacBio-sequenced genomes of *Drosophila* genus. *Mol. Biol. Evol.* (in press).
85. Chen, J. et al. Evolutionarily new genes in humans with disease phenotypes reveal functional enrichment patterns shaped by adaptive innovation and sexual selection. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.14.567139 (2023).
86. Gan, X. et al. The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. *Nat. Plants* **2**, 16167 (2016).
87. Ma, F., Lau, C. Y. & Zheng, C. Young duplicate genes show developmental stage-and cell type-specific expression and function in *Caenorhabditis elegans*. *Cell Genom.* **4**, 100467 (2024).
88. Shao, Y. et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res.* **29**, 682–696 (2019).
89. Trigos, A. S., Pearson, R. B., Papenfuss, A. T. & Goode, D. L. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. *Proc. Natl Acad. Sci. USA* **114**, 6406–6411 (2017).
90. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends Plant Sci.* **19**, 698–708 (2014).
91. Zhou, Q. et al. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**, 1246338 (2014).
92. Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
93. Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance* (Harvard Univ. Press, 1982).
94. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
95. Lee, Y. C. G. et al. Rapid evolution of gained essential developmental functions of a young gene via interactions with other essential genes. *Mol. Biol. Evol.* **36**, 2212–2226 (2019).
96. Auer, T. O., Alvarez-Ocana, R., Cruchet, S., Benton, R. & Arguello, J. R. Copy number changes in co-expressed odorant receptor genes enable selection for sensory differences in drosophilid species. *Nat. Ecol. Evol.* **6**, 1343–1353 (2022).
97. Kasinathan, B. et al. Innovation of heterochromatin functions drives rapid evolution of essential *ZAD-ZNF* genes in *Drosophila*. *eLife* **9**, e63368 (2020).
98. Li, M., Kasan, K., Saha, Z., Yoon, Y. & Schmidt-Ott, U. Twenty-seven *ZAD-ZNF* genes of *Drosophila melanogaster* are orthologous to the embryo polarity determining mosquito gene cucoid. *PLoS ONE* **18**, e0274716 (2023).
99. Piunti, A. et al. CATACOMB: an endogenous inducible gene that antagonizes H3K27 methylation activity of polycomb repressive complex 2 via an H3K27M-like mechanism. *Sci. Adv.* **5**, eaax2887 (2019).
100. Tao, F., Sollapura, V., Robert, L. S. & Fan, C. Neofunctionalization of tandem duplicate genes encoding putative β-L-arabinofuranosidases in *Arabidopsis*. *Plant Physiol.* **192**, 2855–2870 (2023).
101. Huang, Y. et al. Species-specific partial gene duplication in *Arabidopsis thaliana* evolved novel phenotypic effects on morphological traits under strong positive selection. *Plant Cell* **34**, 802–817 (2022).
102. Chen, S. D., Zhang, Y. E. & Long, M. Y. New genes in *Drosophila* quickly become essential. *Science* **330**, 1682–1685 (2010).
103. Xia, S. Q. et al. Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in *Drosophila* development. *PLoS Genet.* **17**, e1009654 (2021).
104. Zeng, X. K. et al. Genome-wide RNAi screen identifies networks involved in intestinal stem cell regulation in *Drosophila*. *Cell Rep.* **10**, 1226–1238 (2015).
105. Xia, S. Q. et al. Rapid gene evolution in an ancient post-transcriptional and translational regulatory system compensates for meiotic X chromosomal inactivation. *Mol. Biol. Evol.* **39**, msab296 (2022).
106. Eslamieh, M., Mirsalehi, A., Markova, D. N. & Betrán, E. COX4-like, a nuclear-encoded mitochondrial gene duplicate, is essential for male fertility in *Drosophila melanogaster*. *Genes* **13**, 424 (2022).
107. Chen, S. D. et al. Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. *EMBO J.* **31**, 2798–2809 (2012).
108. Zou, C., Lehti-Shiu, M. D., Thomashow, M. & Shiu, S.-H. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genet.* **5**, e1000581 (2009).
109. VanKuren, N. W. & Long, M. Y. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat. Ecol. Evol.* **2**, 705–712 (2018).
110. VanKuren, N. W., Chen, J. & Long, M. Sexual conflict drive in the rapid evolution of new gametogenesis genes. *Semin. Cell Develop. Biol.* **159–160**, 27–37 (2024).
111. Emelyanov, A. V., Barcenilla-Merino, D., Loppin, B. & Fyodorov, D. V. APOLLO, a testis-specific *Drosophila* ortholog of importin-4, mediates the loading of protamine-like protein Mst77F into sperm chromatin. *J. Biol. Chem.* **299**, 105212. (2023).
112. Zhang, W., Landback, P., Gschwend, A. R., Shen, B. & Long, M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* **16**, 1–14 (2015).
113. Zu, J. et al. Topological evolution of coexpression networks by new gene integration maintains the hierarchical and modular structures in human ancestors. *Sci. China Life Sci.* **62**, 594–608 (2019).
114. Krinsky, B. H. et al. Rapid cis-trans coevolution driven by a novel gene retroposed from a eukaryotic conserved CCR4–NOT component in *Drosophila*. *Genes* **13**, 57 (2022).
115. Xia, S. et al. Altered transcription and neofunctionalization of duplicated genes rescue the harmful effects of a chimeric gene in *Brassica napus*. *Plant Cell* **28**, 2060–2078 (2016).
116. Zhang, Z. et al. Two young genes reshape a novel interaction network in *Brassica napus*. *New Phytol.* **225**, 530–545 (2020).
117. Moutinho, A. F., Eyre-Walker, A. & Dutheil, J. Y. Strong evidence for the adaptive walk model of gene evolution in *Drosophila* and *Arabidopsis*. *PLoS Biol.* **20**, e3001775 (2022).
118. Bornberg-Bauer, E., Hlouchova, K. & Lange, A. Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183 (2021).
119. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).
120. Schmitz, J. F., Ullrich, K. K. & Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2**, 1626–1632 (2018).
121. Chen, J. et al. The rapid evolution of de novo proteins in structure and complex. *Genome Biol. Evol.* **16**, evae107 (2024).

122. Middendorf, L., Ravi Iyengar, B. & Eicholt, L. A. Sequence, structure, and functional space of *Drosophila* de novo proteins. *Genome Biol. Evol.* **16**, evae176 (2024).
123. Grandchamp, A. et al. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in *Drosophila melanogaster*. *Genome Res.* **33**, 872–890 (2023).
124. Betrán, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* **12**, 1854–1859 (2002).
125. Emerson, J. J., Kaessmann, H., Betrán, E. & Long, M. Y. Extensive gene traffic on the mammalian X chromosome. *Science* **303**, 537–540 (2004).
126. Dai, H. Z., Yoshimatsu, T. F. & Long, M. Y. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. *Gene* **385**, 96–102 (2006).
127. Vinckenbosch, N., Dupanloup, I. & Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl Acad. Sci. USA* **103**, 3220–3225 (2006).
128. Li, H. et al. Fly cell atlas: a single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).
129. Villanueva-Cañas, J. L. et al. New genes and functional innovation in mammals. *Genome Biol. Evol.* **9**, 1886–1900 (2023).
130. Ruiz-Orera, J. et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721 (2015).
131. Soumillon, M. et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).
132. Wu, D. D. et al. 'Out of pollen' hypothesis for origin of new genes in flowering plants: study from *Arabidopsis thaliana*. *Genome Biol. Evol.* **6**, 2822–2829 (2014).
133. Cui, X. et al. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol. Plant* **8**, 935–945 (2015).
134. Vakirlis, N. et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
135. Zhuang, X. & Cheng, C.-H. C. Propagation of a de novo gene under natural selection: antifreeze glycoprotein genes and their evolutionary history in codfishes. *Genes* **12**, 1777 (2021).
136. Zhou, Y. et al. Gene fusion as an important mechanism to generate new genes in the genus *Oryza*. *Genome Biol.* **23**, 1–23 (2022).
137. Cridland, J. M., Majane, A. C., Zhao, L. & Begun, D. J. Population biology of accessory gland-expressed de novo genes in *Drosophila melanogaster*. *Genetics* **220**, iyab207 (2022).
138. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024 (2002).
139. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
140. Hudson, R. R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).
141. Cardoso-Moreira, M. et al. Evidence for the fixation of gene duplications by positive selection in *Drosophila*. *Genome Res.* **26**, 787–798 (2016).
142. Raices, J. B., Otto, P. A. & Vibranovski, M. D. Haploid selection drives new gene male germline expression. *Genome Res.* **29**, 1115–1122 (2019).
143. Schrider, D. R., Hahn, M. W. & Begun, D. J. Parallel evolution of copy-number variation across continents in *Drosophila melanogaster*. *Mol. Biol. Evol.* **33**, 1308–1316 (2016).
144. Liao, Y., Zhang, X., Chakraborty, M. & Emerson, J. Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Res.* **31**, 397–410 (2021).
145. Zhang, D. et al. Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in *Drosophila* and humans. *Nat. Ecol. Evol.* **6**, 273–287 (2022).
146. Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. & Long, M. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* **320**, 1629–1631 (2008).
147. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
148. Williamson, S. H. et al. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl Acad. Sci. USA* **102**, 7882–7887 (2005).
149. Boyko, A. R. et al. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.* **8**, e1000451 (2010).
150. Groza, C. et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat. Commun.* **15**, 657 (2024).
151. Yılmaz, F. et al. Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science* **386**, eadn0609 (2024).
152. Bolognini, D. et al. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* **634**, 617–625 (2024).
153. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
154. Chen, R. et al. A de novo evolved gene contributes to rice grain shape difference between *indica* and *japonica*. *Nat. Commun.* **14**, 5906 (2023).
155. Yu, Y. et al. A *Zea* genus-specific micropeptide controls kernel dehydration in maize. *Cell* **188**, 1–16 (2025).
156. Deng, Z. et al. Map-based cloning reveals the complex organization of the *BnRf* locus and leads to the identification of *BnRf b*, a male sterility gene, in *Brassica napus*. *Theor. Appl. Genet.* **129**, 53–64 (2016).
157. Muzaffar, A. et al. A newly evolved rice-specific gene JAUP1 regulates jasmonate biosynthesis and signalling to promote root development and multi-stress tolerance. *Plant Biotechnol. J.* **22**, 1417–1432 (2024).
158. Wang, C. Y. et al. *Xa7*, a small orphan gene harboring promoter trap for AvrXa7, leads to the durable resistance to *Xanthomonas oryzae* Pv. *oryzae*. *Rice (NY)* **14**, 48 (2021).
159. Fang, H. et al. A monocot-specific hydroxycinnamoylputrescine gene cluster contributes to immunity and cell death in rice. *Sci. Bull.* **66**, 2381–2393 (2021).
160. Shen, S. Q. et al. An *Oryza*-specific hydroxycinnamoyl tyramine gene cluster contributes to enhanced disease resistance. *Sci. Bull.* **66**, 2369–2380 (2021).
161. Moon, H., Jeong, A. R., Kwon, O. K. & Park, C. J. *Oryza*-specific orphan protein triggers enhanced resistance to *Xanthomonas oryzae* pv. *oryzae* in rice. *Front. Plant Sci.* **13**, 859375 (2022).
162. Smoak, R. A., Snyder, L. F., Fassler, J. S. & He, B. Z. Parallel expansion and divergence of an adhesin family in pathogenic yeasts. *Genetics* **223**, iyad024 (2023).
163. Fiddes, I. T. et al. Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369 (2018).
164. Broeils, L. A., Ruiz-Orera, J., Snel, B., Hubner, N. & vanHeesch, S. Evolution and implications of de novo genes in humans. *Nat. Ecol. Evol.* **7**, 804–815 (2023).
165. Suzuki, I. K. et al. Human-specific NOTCH2NL genes expand cortical neurogenesis through delta/notch regulation. *Cell* **173**, 1370–1384 (2018).
166. Florio, M. et al. Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).

167. Yamamoto, S. et al. Structural characterization of human de novo protein NCYM and its complex with a newly identified DNA aptamer using atomic force microscopy and small-angle X-ray scattering. *Front. Oncol.* **13**, 1213678 (2023).
168. Suenaga, Y., Nakatani, K. & Nakagawara, A. De novo evolved gene product NCYM in the pathogenesis and clinical outcome of human neuroblastomas and other cancers. *Jpn. J. Clin. Oncol.* **50**, 839–846 (2020).
169. Gilbert, W. Towards a paradigm shift in biology. *Nature* **349**, 99 (1991).
170. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
171. Domazet-Loso, T., Brajkovic, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
172. Domazet-Loso, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–U107 (2010).
173. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* **32**, 258–267 (2015).
174. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862 (2020).
175. Domazet-Lošo, T. et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* **34**, 843–856 (2017).

Acknowledgements

This review was motivated and supported by the John Simon Guggenheim Memorial Fellowship for Natural Sciences (Biology), US and Canada (2022), with support from the Division of Biological Sciences at the University of Chicago (to M.L.). M.L. was also recently supported by NSF (MCB2020667) and NIH (1R01GM116113-01A1). J.J.E. was supported by NIH under grants R01GM123303 and R35GM153327, and D.A. was also supported by NIH under grant F32GM146423. We are indebted to many friends, colleagues and Long lab members, present and past, for valuable discussion.

Author contributions

M.L. and S.X. organized and composed the review with D.A., J.C. and J.J.E.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-02059-0.

Correspondence and requests for materials should be addressed to Manyuan Long.

Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.