The evolution of gene duplications: classifying and distinguishing between models

Hideki Innan* and Fyodor Kondrashov§

Abstract | Gene duplications and their subsequent divergence play an important part in the evolution of novel gene functions. Several models for the emergence, maintenance and evolution of gene copies have been proposed. However, a clear consensus on how gene duplications are fixed and maintained in genomes is lacking. Here, we present a comprehensive classification of the models that are relevant to all stages of the evolution of gene duplications. Each model predicts a unique combination of evolutionary dynamics and functional properties. Setting out these predictions is an important step towards identifying the main mechanisms that are involved in the evolution of gene duplications.

Our knowledge of gene duplications at the phylogenetic, functional and genomic levels is impressive. Hardly any aspect of genome evolution or function is not somehow linked to gene duplications, which occur in all kinds of life forms and have taken place since before the last universal common ancestor. Some genomes contain large numbers of genes owing to whole-genome duplications or lineage-specific gene expansions. The duplication of genes may form a cornerstone in the evolution of biological complexity. Finally, gene duplications segregate in high numbers in natural populations, and some cause disease or confer an adaptive advantage. These and other aspects of gene duplications have been continuously reviewed and their importance has been emphasized.

By contrast, our current understanding of the selective and evolutionary mechanisms involved in the emergence, maintenance and evolution of gene duplications is preliminary and fragmented. There are two main reasons for this discordance between theory and data. First, the literature contains many different models of and hypotheses about the evolution of gene duplications, which have not been described in a systematic way. Second, the development of theory on gene duplications has been driven by data, without attempts to stringently test hypotheses. Consequently, although many models for the evolution of gene duplications have been proposed, we still do not know their relative importance for describing general trends in the evolution of gene duplications or their applicability to specific gene copies.

To change this situation, we present here a classification of the different hypotheses on the emergence, maintenance and evolution of gene duplications. Following previous reviews of specific models or model types, we provide a comprehensive classification of all models of gene duplications. We include verbal and more formal models, and compare their predictions for the anticipated evolutionary scenarios in which gene duplications are retained, and for the fates of gene copies at the different stages of the evolution of a gene duplication. Capitalizing on the differences that we point out between the models, we also provide suggestions for future experiments aimed at distinguishing between them. We do not review evidence that supports or refutes any of the hypotheses. Rather, our aim is to bring order to the theoretical literature on gene duplications, which we hope will aid researchers in designing better studies to improve our understanding of the function and evolution of duplicated genes.

Classification and predictions

Every type of genetic change undergoes three main stages as it competes for evolutionary preservation: origin through mutation, a fixation phase when it segregates in the population and a preservation phase when the fixed change is maintained. Gene duplications follow this trajectory with one important addition; from the moment of emergence of a new gene copy, the acquisition of genetic differences between the copies can alter the chances of both copies being preserved. Although all of these life stages are functionally and evolutionarily important,
many models of gene duplication evolution describe the phase of acquisition of differences between gene copies as crucial in the preservation of new gene copies. Therefore, we call it the fate-determination phase (FIG. 1).

Approximately a dozen models for the evolution and maintenance of gene duplications have been proposed over the years. We describe the models as they have been articulated in the original literature (summarized in TABLE 1 and FIG. 2) and describe their development. To classify and distinguish between these models, it is convenient and useful to focus on the selective forces and evolutionary events at different stages of the life history of the duplication. However, there is substantial overlap in the descriptions and predictions of different models in the same category.

Suppose that a new duplicate gene pair (A–A) arises in a population with N random-mating diploids, in which all genomes initially have single copies of gene A. Throughout this Review, to be consistent with the models in the literature, we assume that the new duplicate has a complete set of functional motifs and is functionally indistinguishable from the original copy unless otherwise specified. The probability of the fixation of A–A and the length of the fixation phase (fixation time) theoretically depend on the relative strength of selection for the A–A and A genotypes. This is one of the most important factors that differentiate the models and we use it as the basis for our classification. When A–A confers no selective advantage or a disadvantage (that is, it is neutral), A–A will be fixed in the population at a probability of 1 / 2N and the fixation process takes on average 4N generations. This is the defining feature of the models we place in category I, including the popular neofunctionalization and subfunctionalization models (see below). By contrast, the models in categories II and III involve positive selection for the new duplicate. In these cases, the fixation probability is higher and the fixation time is shorter than in the neutral case. We place the dosage balance model in a separate category because it lacks the fixation phase and considers a pair of duplicates that are created by a whole genome duplication.

Following our systematic classification of the models, differences in the polymorphism and divergence dynamics can be used to distinguish between them. For each model we describe the most likely pattern in polymorphism levels and sequence divergence (see BOX 1 for a description of these variables), and aspects of gene function. In particular, we discuss the synonymous–non-synonymous ratios of polymorphism and divergence, ωs and ωa, which should represent the intensity of selection as described in BOX 2. Few models provide specific predictions about the long-term molecular evolution of the two duplicates in the final preservation phase. However, because most studies of the evolution of gene duplications focus on long-term divergence, based on these data we can provide the most likely predictions for this aspect of duplication evolution. Importantly, when making predictions we take into account the effect of gene conversion, which is common in many species and has a strong influence on the dynamics of the evolution.
of gene duplicates in the fixation, fate-determination and early preservation phases (BOXES 3,4).

**Category I**

This category contains three models that assume that a duplication does not affect fitness, so that the fixation of the duplicated copy is a neutral process: the neofunctionalization, duplication–degeneration–complementation (DDC) and specialization models. A general feature of these models is that a gene duplication must go through the fate-determination phase rapidly to reach the preservation phase, otherwise one of the copies can be pseudogenized because selection is relaxed before the preservation stage. The differences among these models begin at the fate-determination phase.

**Category I-a Ohno’s neofunctionalization.** Ohno’s neofunctionalization model marked the beginning of the theoretical discussion of gene duplications\(^19,20\). He reasoned that a single gene copy is enough to fulfil the function of the gene and therefore extra copies are redundant. If such a redundant copy is fixed by drift in the population, the original copy will maintain its function, and the new copy will be relieved from negative selection\(^19,20\). The new copy can therefore be pseudogenized or lost through the accumulation of neutral loss-of-function mutations. However, Ohno suggested that occasionally, as the redundant, dying gene copy accumulates substitutions, it may acquire a new gene function that will be maintained by selection. It is not clear how selection can distinguish between the new and original copies of the

### Table 1 | Summary of the models of gene-duplication evolution

<table>
<thead>
<tr>
<th>Name</th>
<th>Functional evolution</th>
<th>Fixation phase</th>
<th>Fate-determination phase</th>
<th>Preservation phase*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Function of original copy</td>
<td>Function of new copy</td>
<td>Fate-determining mutation</td>
<td>Selection on new copy</td>
</tr>
<tr>
<td>Neofunctionalization</td>
<td>Kept</td>
<td>Novel</td>
<td>Gain-of-function mutations</td>
<td>Neutral</td>
</tr>
<tr>
<td>DDC</td>
<td>Subfunctionalized</td>
<td>Subfunctionalized</td>
<td>Loss-of-function mutations</td>
<td>Neutral</td>
</tr>
<tr>
<td>Specialization or EAC</td>
<td>Subfunctionalized</td>
<td>Subfunctionalized</td>
<td>Gain-of-function mutations</td>
<td>Neutral</td>
</tr>
<tr>
<td>Positive dosage</td>
<td>Kept</td>
<td>Same as original</td>
<td>NA</td>
<td>Positive selection on duplication</td>
</tr>
<tr>
<td>Shielding against deleterious mutations</td>
<td>Kept</td>
<td>Same as original</td>
<td>NA</td>
<td>Positive selection on duplication</td>
</tr>
<tr>
<td>Modified duplication</td>
<td>Kept</td>
<td>Novel</td>
<td>Gain-of-function mutations</td>
<td>Positive selection on duplication</td>
</tr>
<tr>
<td>Permanent heterozygote</td>
<td>Subfunctionalized</td>
<td>Subfunctionalized</td>
<td>Gain-of-function mutations</td>
<td>Positive selection on pre-duplicational variation</td>
</tr>
<tr>
<td>Adaptive radiation model</td>
<td>Kept</td>
<td>Novel</td>
<td>Gain-of-function mutations</td>
<td>Positive selection on pre-duplicational variation</td>
</tr>
<tr>
<td>Diversifying selection</td>
<td>Multiple functions</td>
<td>Multiple functions</td>
<td>Gain-of-function mutations</td>
<td>Positive selection on pre-duplicational variation</td>
</tr>
<tr>
<td>Dosage balance</td>
<td>Kept</td>
<td>Original</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

*The predicted pattern of molecular evolution is indicated as α when the pattern is not different from that in the pre-duplication phase (as α’ when the selective pressure may be relaxed), as β when amino acid substitutions can be accelerated by positive selection and as o when amino acid substitutions are always accelerated by diversifying selection. DDC, duplication–degeneration–complementation; EAC, escape from adaptive conflict; NA, not applicable.

**Specialization**

A process of improvement of different aspects of gene function in each gene copy, which is driven by positive selection.
Duplica
tion
Neofunctionalization

Category I

DDC

Specialization

Fixation of duplication

Fate-determining mutation

Stable maintenance

Positive dosage

Category II

Shielding against deleterious mutations

Modified duplication

Adaptive radiation

Category III

Permanent heterozygote

Diversifying selection

www.nature.com/reviews/genetics
gene and how the new function can evolve in either copy. Nevertheless, we follow Ohno's description of the model because some of its predictions depend on the assumption that one of the copies remains under selection while the other is free from it.

This model sensu strico predicts that the rate of evolution after gene duplication will be accelerated in the duplicated copy and maintained in the original one. Because the new function in the derived copy can be improved by additional advantageous mutations in the preservation phase, it is very likely that positive selection specifically targets the derived copy. Meanwhile, as the original copy keeps the ancestral function, its selective constraint may not change much throughout the process. The parameters of this model of the time to fixation and to non-functionalization (pseudogenization) have been difficult to work out because they are based on aspects of gene function, such as the fraction of loss-of-function mutations out of all possible ones, which are yet to be accurately determined. However, because most de novo mutations are likely to be deleterious, it is widely agreed that pseudogenization is the most probable fate of the new gene copy. The rate at which a neutrally evolving gene copy may acquire a new function is the most difficult parameter to estimate and depends at least on the strength of selection in favour of the new function.

The length of time before a redundant gene copy is pseudogenized is likely to be short. The predicted behaviour of polymorphism under this model is described in BOX 5. Asymmetry of divergence is expected, such that \( \omega_c = \omega_o < 1 \) in the old copy and \( \omega_c = 1 \) in the new one until the fate-determining phase. Such asymmetry is also expected in the preservation phase, as \( \omega_c \) for the new copy can be increased by positive selection; occasionally \( \omega_c > 1 \) for the new copy and \( \omega_c = \omega_o \) for the original copy. The derived copy is not lost or pseudogenized, the evolution of a radically new function is expected. Gene conversion can inhibit the acquisition of a new function in this model (BOX 5).

Non-functionalization
The process of the accumulation of neutral mutations in a duplicated gene that renders the gene copy non-functional. Also known as pseudogenization.

Degenerate mutation
A mutation that does not affect fitness but is damaging to gene function.

Promiscuous function
A secondary, possibly neutral function that is performed by a protein with another primary function. It is an example of gene sharing.

Figure 2 | Illustration of the models in categories I, II and III for gene-duplication evolution. The processes that are expected to occur for each model from the birth of a gene duplication to its stable maintenance. Red boxes represent genes with the original function, functional evolution is shown by changes of colour, and white indicates a loss of gene function. The darker arrows indicate situations in which the process is neutral, and the light arrows indicate cases in which positive selection is involved. DDC, duplication–degeneration–complementation.

(Category I-c) Specialization or escape from adaptive conflict. Hughes proposed a different verbal model of functional specialization of gene copies. This model did not challenge the complete redundancy assumption of Ohno's neofunctionalization model and assumed that gene copies are fixed by genetic drift. However, it suggested that if the original gene was performing two functions that could not be independently improved, then after duplication each gene copy can be driven by positive selection to specialize — that is, to improve one of the two functions. The improvement could be owing to either differentiation of expression patterns or the substantial improvement of a promiscuous function. This model was recently rechristened as the escape from adaptive conflict (EAC) model.
Gene sharing
An early term describing situations in which a gene has more than one function. Modern studies describe such genes as multifunctional.

Gene dosage
The amount of product produced from a gene; broadly equivalent to gene expression.

Gene amplification
The emergence of a non-heritable extra copy of a gene in a somatic tissue. In microorganisms this term can be used interchangeably with gene duplication.

This model has not been explored formally. However, because the gene copy is fixed by drift the dynamics of the fixation phase will be the same as in the neofunctionalization and DDC models. In the fate-determination phase the divergence of the copies will be affected by positive selection for improvement of the different functions, and the adaptive mutations will act as the fate-determining events.

The predicted pattern of polymorphism for this model is explored in BOX 5. For long-term evolution after the fate-determination phase, \( \omega_k \) for both copies can be increased by positive selection, which can lead to similar functional, tissue or cellular compartmentalization to the DDC model. Because the gene functions are improved compared with the original gene, the fitness contributions of the two paralogues should be greater than those of unduplicated orthologues in closely related species. Multifunctional genes should be preferentially duplicated by this mechanism. For this model to apply widely, gene sharing and antagonistic pleiotropy must be common, such that an individual function of a multifunctional gene can be improved only by harming another functional aspect of the same gene. Gene conversion will inhibit the specialization process (BOX 3).

**Category II**
For the models in this category, the duplication itself is advantageous. There can be at least three reasons for this type of adaptation: beneficial increases in dosage, the masking of deleterious mutations and the opportunity for the immediate emergence of a new function.

**(Category II-a) Beneficial increase in dosage.** Ideas relating to the effect of gene duplication on gene dosage have been developed in parallel to the neofunctionalization model (reviewed in REFs 1,19). Early suggestions of an interplay between fitness and gene copy number tended to focus on bacteria and archaea,\(^ {1,41–45} \), and the terms gene duplication and gene amplification were often used interchangeably. However, as data from eukaryotes became more widely available, authors started to relate dosage and gene duplications in multicellular organisms.\(^ {1,43,46–48} \). The argument for the basis of the interplay between dosage and gene duplications is simple: if an increase of dosage of a particular gene is beneficial then a duplication of this gene may be fixed by positive selection. After fixation, the gene duplication will be maintained as long as the conditions that led to the selection for increased dosage remain. When the increase in fitness owing to the fixed gene duplication is small, the gene copies are expected to be under relaxed selection pressure.\(^ 1 \) The model of beneficial increase of gene dosage also intrinsically connects gene copy number differences and deleterious changes in dosage that may be caused by duplication or loss of gene copies.\(^ {48–50} \)

This model may apply to three types of genes. First, genes that mediate the interaction between the organism and the environment, such as stress-response genes, sensory genes, transport genes and genes that have a metabolism-related function.\(^ 1 \) Second, genes that have dosage-sensitive functions owing to protein–protein interaction properties of their products or the nature of the metabolic pathways in which they act.\(^ {49,50} \) Third, genes that have products that are generally required in large doses, such as ribosomal or histone genes.\(^ {51} \)

There have been few formal modelling attempts for this type of model,\(^ {1,49,50} \); however, a few general trends can be outlined. Under unvarying strong selection for increased dosage, the duplication will quickly become fixed in the population and will be preserved without undergoing a fate-determination phase. Throughout the process both copies will be under negative selection. However, if selection for the duplicated copy is weak, a null mutation might become fixed by random genetic drift, resulting in pseudogenization. This situation is similar to models in category I, in that a fate-determining mutation may be needed to reach the preservation phase. However, in contrast to category I models, weak selection for dosage increases the fixation probability and shortens the fixation phase. More importantly, the fate-determination phase can be longer, so that there is more chance of a fate-determining mutation occurring before pseudogenization takes place as a result of negative selection against null and deleterious mutations. The long-term molecular evolution in the preservation phase follows that of the models in category I.

In a variable environment, selection for increased dosage may occur but then be replaced by selection against dosage increases, leading to a cycle of gene duplication and loss.\(^ 52 \) Alternatively, if selection for increased dosage is strong and constant, there will be few gene duplications under near-equilibrium conditions.

It is possible to make predictions from this model; see BOX 5 for polymorphism patterns. If selection on dosage is strong, then \( \omega_k < 1 \) throughout the whole process. For recent duplications, the sum of the expression levels of

---

**Box 1 | Variables used throughout the Review**

- \( \mu \): per generation mutation rate
- \( N \): population size
- \( \theta \): population variability equal to \( 2N\mu \)
- \( \pi \): the average number of pairwise nucleotide differences
- \( K_s \): rate of synonymous evolution, measured as the number of synonymous substitutions divided by the number of synonymous sites
- \( K_a \): rate of non-synonymous evolution, measured as the number of non-synonymous substitutions divided by the number of non-synonymous sites
- \( \pi_s \): synonymous level of polymorphism, measured as the average number of pairwise synonymous differences divided by the number of synonymous sites
- \( \pi_n \): non-synonymous level of polymorphism, measured as the average number of pairwise non-synonymous differences divided by the number of non-synonymous sites
- \( \omega_i \): equal to \( K_k/K_s \), representing the strength of selection on non-synonymous substitutions
- \( \omega_r \): equal to \( \pi_s/\pi_n \), representing the strength of selection on non-synonymous polymorphisms
- Subscript 0: the expected measurement of any of the above variables of a duplicated gene before duplication, in the pre-duplication phase. For example, \( \pi_n \) is the expected level of synonymous polymorphisms in a gene immediately before duplication.
Box 2 | Effect of selection on polymorphism and divergence

The most important parameters that affect the pattern of polymorphism and divergence are the mutation rate and population size. We define the neutral mutation rate as per site per generation. Throughout this Review, we assume a random-mating diploid population that has a constant size $N$, so that the level of polymorphism is characterized by $\theta = 4Nu$. If the polymorphism level is measured by $\pi$, the average number of pairwise nucleotide differences, then at mutation-drift equilibrium the expectation of $\pi$ at a neutral site is given by $E(\pi) = \theta$ for a single-locus gene and also in each of the duplicated genes when they reach equilibrium (but see Box 4 when gene conversion is active). The expected rate of evolution (nucleotide substitution rate) in each copy is $\mu$, which determines the divergence at neutral sites between the two copies.

In contrast to neutral sites, other patterns are expected at functional sites owing to the actions of positive and negative selection. A common approach for evaluating selection, which we also use in our Review, is to compare the patterns of polymorphism and divergence at synonymous and non-synonymous sites. The rationale is that synonymous sites can be used as a neutral control and non-synonymous sites are more likely to be functional and subject to selection.

We define $\pi_s$ and $K_s$ as the observed levels of polymorphism and divergence, respectively, at synonymous sites, and $\pi_i$ and $K_i$ are the levels of polymorphism and divergence, respectively, at non-synonymous sites. $\omega_i$ and $\omega_s$ represent the expected non-synonymous–synonymous ratios of polymorphism and divergence $(\omega_i = E(\pi_i/K_i)$ and $\omega_s = E(\pi_s/K_s)$, respectively. Under neutrality, $\omega_i = \omega_s = 1$, so these measurements are commonly used as measures of selection. We add the subscript 0 to denote measurements of the single-copy gene in the pre-duplication phase (for example, $\pi_{0s}$ is the level of synonymous polymorphisms in a gene immediately before duplication). Such measurements can be used as a control to quantitatively measure the change in selective pressure after gene duplication.

Negative selection constantly eliminates deleterious mutations, leading to reduction of the level of polymorphism and the rate of evolution at functional sites ($\omega_i < 1$ and $\omega_s < 1$). When negative selection is relaxed, both $\omega_i$ and $\omega_s$ increase up to a maximum value of 1. Positive selection can drive a beneficial mutation to fixation, temporarily reducing the level of polymorphism in the surrounding region. If beneficial mutations are constantly fixed, there is an increased rate of evolution, resulting in an increase of $\omega_i$. Therefore, there are two possible explanations for an increase in $\omega_i$: relaxation of negative selection and the action of positive selection — and other data are usually needed to distinguish between them. However, if $\omega_i > 1$ then the action of positive selection can be inferred.

The most likely reason why the original and duplicated copies should exceed that of the non-duplicated orthologue of a closely related species. If strong or weak selection for extra dosage is caused by a change of environment, the functional repertoire of the duplicated genes should be relevant to the interaction of the gene with the changed environmental aspect. Two aspects of gene conversion can promote this process (the first two ‘effects that make preservation more likely’ in BOX 5).

(Category II-b) Shielding against deleterious mutations. Gene copies may shield one another from the accumulation of deleterious mutations, provided that they are at least partially redundant. An individual that has just obtained a new gene copy may have a slight selective advantage over those without one because new mutations will not negatively affect fitness, and the individual will be subject to a slightly lower mutational load. However, the selective advantage of such shielding is expected to be almost insignificant (that is, of the order of the mutation rate ($4\mu$), but is expected to increase with the rate of mutation, the rate of duplication or the length of duplicated segments ($K_i$). The extra gene copies are quickly destroyed by mutations and therefore duplications must occur cyclically to provide sustained sheltering against deleterious mutations. Therefore, there is no preservation phase because the only function of the gene copy is to be destroyed by mutations.

This model predicts that repeated and rapid birth and death of extra gene copies will occur, and that each copy is present for a short time in the genome, with the new copy accumulating mutations in a neutral fashion and eventually being pseudogenized. $\omega_{K_i} < \omega_K < 1$ when a copy is functional and $\omega_K = 1$ after it is pseudogenized. Gene conversion may prolong the birth and death cycle. If this model applies widely, rapid loss of synteny and many traces of pseudogenes should be observed in the genome.

(Category II-c) Gene duplication with a modified function. If the process of gene duplication itself creates a new function, the new copy can be fixed and preserved by positive selection. Partial gene duplication may fail to copy regulatory elements or other functional parts of the gene, resulting in a new function from the moment of duplication. Similarly, a new genomic location of the gene copy may introduce new functional aspects, or a retrotransposed copy may recruit regulatory elements in the new location or be integrated into an existing gene, which results in the formation of a chimeric gene. If a new and beneficial gene function emerges at the point of duplication then fixation will occur through positive selection for the new function and the gene copy will be immediately preserved. It is likely that the function of the derived copy can be improved by additional mutations, leading to a period of rapid evolution immediately after duplication. It is predicted that a selective sweep should be observed after the duplication event. In some cases $\omega_K > 1$ in the new copy owing to positive selection. A radically new function will be seen for the new copy, which may have lost or gained promoters, domains or introns.

Category III

This category considers models in which duplication occurs in a gene for which genetic variation exists in the population. In some cases, these polymorphisms can immediately become fate-determining mutations that promote fixation of the duplicated copy. The duplicate and the fate-determining mutations are fixed almost at the same time, and these models do not have a fate-determining phase. A common feature of the three models in this category is that a polymorphic allele that is present in the pre-duplication phase is fixed at the same time as the new copy, which requires recombination between the two copies. If two alleles A and B are polymorphic in the ancestral single-locus gene, then a duplication event can make only A–A or B–B haplotypes, not the advantageous A–B or B–A haplotypes. An advantageous haplotype can arise through recombination; for example, between a single-copy A and duplicated B–B copies. Positive selection can then fix A–B or B–A (FIG. 2).
Gene conversion is a non-reciprocal recombination process, which is usually described as a copy-and-paste event, and which transfers a short tract of DNA that ranges from several to (occasionally) thousands of base pairs. Gene conversion occurs not only between orthologous regions but also between paralogous regions when they have sufficient sequence identity (as shown in the figure). We focus on paralogous gene conversion events (interlocus or ectopic gene conversion) because they have a substantial effect on the molecular evolution of duplicated genes, especially during the early stages after duplication. Gene conversion shuffles the DNA variation between paralogous duplicated regions, resulting in a substantial reduction in the divergence between duplicates. When gene conversion occurs at a reasonably high rate, the duplicates co-evolve for a long time (concerted evolution).

Gene conversion can make the preservation of a duplicated gene more or less likely in the following ways.

**Effects that make preservation more likely**
- Frequent gene conversion keeps the sequence identity between paralogues high. This effect should promote the retention of both copies when a large amount of a single gene product is required; this can be applied to the model of the beneficial increase of dosage. High sequence identity may also be preferred in the dosage-balance model.
- Gene conversion allows both copies to share a beneficial mutation, and deleterious mutations can be more efficiently removed from the population. These beneficial effects hold for any model.
- Because gene conversion transfers a short DNA tract, it can create new allelic combinations that confer an advantage.

**Effects that make preservation less likely**
- The homogenizing effect of gene conversion should be disfavoured in all models in which a fate-determining mutation is involved, because selection will operate to maintain the difference between the original and duplicated copies that is created by the mutations. In such a situation, population genetic theory predicts an interesting pattern of polymorphism and divergence (Box 4).
- Gene conversion from a pseudogenized copy to a functional copy will be deleterious. In humans, there are several genetic diseases that are caused by gene conversion from pseudogenes.

**(Category III-a) Adaptive radiation model.** The model proposed by Francino emphasizes a period of ‘pre-adaptation’ in the pre-duplication phase that causes adaptive fixation of the subsequent duplication. This requires that some gene copies in the population should pre-adapt to perform new functions while still performing the original function of the gene (see Fig. 2; these copies are represented by red boxes with various coloured triangles). The original article explains pre-adaptation by using a gene that encodes a receptor for an environmental chemical. When a new chemical appears, the receptor may already have or will obtain the ability to bind the new chemical (although the affinity may be low) as well as the original chemical. Duplication of such a pre-adaptive allele (for example, the red gene with a green triangle in Fig. 2) allows the gene carrying this allele to evolve the full function (green box) by additional mutations. In the long term, the function of the duplicated copy could be improved by positive selection. Francino emphasizes that this process could occur repeatedly, resulting in a substantial increase in copy number. Except for the fixation phase, this model is similar to the specialization model. In addition, this model considers the benefits of the wider mutational target that is provided by several gene copies, the number of which would be higher for genes that are more prone to adaptive fixation of duplications. A formal model describing these conditions has not yet been articulated.

This model predicts asymmetrical evolution; \( \omega_1 \) for the new copy during the preservation phase can be increased by positive selection, and \( \omega_1 = \omega_0 \) for the original copy. The overall contribution of the original and duplicated copies to protein functionality and organismal fitness should be greater than that of the unduplicated orthologue in a closely related species. Gene conversion inhibits this process (Box 3).

**(Category III-b) Permanent heterozygote.** Several authors have modelled scenarios in which heterozygote advantage would lead to the fixation and maintenance of gene duplications, as well as putting forward verbal arguments. This model considers a population in which balancing selection maintains two alleles, A and B, of a gene (red and green boxes in Fig. 2). It is assumed that a heterozygote for a single locus is biologically equivalent to two loci that are each homoygous for a different allele. Under these conditions, the selective benefit of the two distinct loci over one heterozygous locus is the elimination of the reduced fitness of homozygotes when only a single locus is present (the segregation load). Therefore, the occurrence of a duplication in a gene for which the highest fitness is achieved in the heterozygous state is immediately beneficial and the duplication is expected to achieve fixation faster than a neutral allele. The fixation of an extra copy of a gene under balancing selection implies that a genome with A–B (the ‘permanent heterozygote’) is fixed in the population. Preservation and fixation occur simultaneously because little change in the selective constraint for A and B is expected. One crucial difference between this model and the specialization model is that in the permanent heterozygote the fitness of the heterozygote is higher than either of the homozygote states in the pre-duplication phase, but in the specialization model the fitness of each homozygote is higher than that of the heterozygote.

This model predicts that there will be a high level of polymorphism between the haplotypes that harbour the two polymorphic alleles in the pre-duplication phase. Once the gene duplication is fixed, the polymorphism level of the two copies will reflect the amounts of variation in each allele in the pre-duplication phase and will gradually increase to \( \theta \) in both copies at equilibrium. Gene conversion should be selected against in the preservation phase (Box 3). The functionality of the two copies can only be replicated by a combination of two alleles of the unduplicated orthologue in a related species. In that case, the divergence between the duplicated copies would pre-date the speciation event.
Box 4 | Effect of gene conversion on the pattern of polymorphism

Gene conversion increases the level of polymorphism\(^{91}\). This is easy to understand intuitively; duplicated copies can share polymorphisms that arose in one copy. The figure shows a pattern of polymorphism in a pair of duplicated genes X and Y when gene conversion is frequent. The regions that originated in genes X and Y are shown in green and purple, respectively. Each gene has a mosaic structure of the two colours owing to frequent exchanges of DNA sequences between X and Y through gene conversion. This creates several sites in which the same pair of allelic nucleotides is segregating in both X and Y (that is, shared polymorphism); this is the major cause for the increased polymorphism level in each gene. Theoretically, in a simple two-copy model the expected level of polymorphism in each copy (\(c\)) can be increased to a value that is twice as high as that of \(\theta\) \(^{91}\) — that is, the expectation in a single-copy gene.

One potential effect of gene conversion is that a fixed fate-determining mutation could be lost. In other words, the fixation of a fate-determining mutation does not necessarily result in a stable preservation phase when gene conversion is active. This effect of gene conversion is explained with an example of neofunctionalization. Let A and B be the original and neofunctionalizing alleles, respectively. Suppose haplotype A–B is advantageous over A–A and fixed in the population. It will then be subject to homogenization by gene conversion, which would create the deleterious haplotypes A–A and B–B. Therefore, selection has to work to eliminate the deleterious haplotypes to keep the neofunctionalizing allele. Population genetic theory\(^{86,92}\) indicates that strong selection against gene conversion is required for the stable maintenance of the fate-determining mutation (and consequently of the two duplicated copies). When such strong selection protects the neofunctionalizing allele, the polymorphism pattern is likely to show a strong signature of selection, as illustrated in the figure. Gene conversion is quickly eliminated from the population if the DNA tract that is affected includes the site that specifies the neofunctionalizing allele (blue bars for gene Y in the figure). In this situation, fixation of additional mutations occurs independently in the two genes, resulting in a local peak of divergence. This effect is restricted to a short region around the target site of selection\(^{95}\).

(Category III-c) Multi-allelic diversifying selection.
In cases in which selection favours genetic variability, gene duplications are beneficial because they provide a larger target for mutation and selection. For a single-copy gene under multi-allelic diversifying selection, such as the major histocompatibility genes, overdominant selection operates. As a result, the maximum possible number of heterozygous individuals in the population is reached\(^{86}\). Consequently, the gene that is under selection accumulates several alleles with different functions (illustrated by boxes of various colours in FIG. 2). For a duplicate of this kind of gene, the same logic holds as that for the permanent heterozygote model, and positive selection favours the fixation of a new copy\(^{91}\). The effect of selection can be increased further compared with the permanent heterozygote model because one duplicated gene can encode up to four different alleles for a diploid. The evolutionary dynamics of this model are similar to those of the permanent heterozygote model. However, unlike the permanent heterozygote model, in the preservation phase the two copies keep accumulating new alleles. This model predicts that in the pre-duplication phase a large number of different alleles are maintained in a polymorphic state by diversifying selection or multi-allelic balancing selection. The frequency of each allele is usually low and so is the level of polymorphism in each allele. However, the overall level of polymorphism is high because of the large amount of divergence between the different alleles, resulting in very high \(\pi_w\) and \(\omega_r\), typically \(\omega_r > 1\). After fixation, the new copy will accumulate mutations, especially at non-synonymous sites. Gene conversion between the two copies will accelerate the recovery of polymorphism levels and creates new allele combinations (the last of the ‘effects that make preservation more likely’ in BOX 3), which is advantageous to the host genome. Long-term evolution involves a high rate of turnover of alleles, implying that \(\omega_r\) is high, typically \(\omega_r > 1\), in both copies. New functionality is constantly gained in lineages of both duplicated and unduplicated species.

(Category IV) This category is unique in that the fixation of a duplicated copy occurs as a by-product of other events, such as whole-genome (chromosome) duplication and large segmental duplications. Therefore, the model considers the fixation of the duplicate as a precondition and there is no fixation phase.

(Category IV-a) Dosage balance. For pairs or groups of genes that have optimal dosages that are dependent on each other, the genes may be duplicated or lost only synchronously. After a whole-genome duplication event, genes involved in protein complexes are preferentially kept. This model has been derived to explain the preferential retention of some duplicated genes after whole-genome duplications\(^{91}\). The original verbal model considered a functional protein that is composed of several subunits. Consider a dosage-sensitive single-copy gene, A, which has an optimum dosage dependent on the dosage of another gene, B. Gene A cannot be duplicated in isolation because of negative selection against dosage imbalance. However, if both genes have been duplicated and fixed in a single large-scale event, such as a whole-genome duplication, they will be maintained in the population because a deletion of only one of them will also cause a dosage imbalance and may be deleterious. There may be other functional scenarios in addition to the case of protein subunits in which the fitness of a duplication of a single gene is deleterious but duplication of several is not\(^{2,3}\). Therefore, if this model applies widely, only genes that are prone to dosage imbalance should be maintained after a large-scale duplication event, and such genes will not be duplicated unless their interaction partners are also duplicated.
The processes that take place from the birth of a duplicated gene copy to its stable maintenance involve multiple rounds of drastic changes in the level of polymorphism in each copy of the gene. The general pattern of changes is briefly summarized here, and can be applied to most models except for those in category III. The first change occurs at the fixation of a duplicated copy. Whether fixation is achieved by genetic drift or selection, the level of polymorphism in the duplicated copy at the end of the fixation phase is expected to be reduced because the entire population shares the same recent origin of the duplicated copy. The same logic holds for the original copy if the two copies are tightly linked. When there is recombination between the two copies, the reduction in the level of polymorphism is relaxed in the original copy. The degree of the reduction of the level of polymorphism in the new copy positively correlates with the magnitude of selection for the duplicated copy. Therefore, a more drastic reduction is expected in the models in category II. After fixation, the two copies accumulate mutations and the level of polymorphism will recover to normal (that is, the expectation of \( \pi \) is \( \theta \)). For gene conversion, this expectation will be higher than \( \theta \) and the recovery process will be accelerated (BOX 4).

In the fate-determination phase, negative selection can be relaxed (that is, \( \omega < 1 \)) in both of the two copies in all models, except for the neofunctionalization model, in which selection is completely relaxed in the duplicated copy and the selective pressure may not change in the original copy (that is, \( \omega = \omega_o \) in the original copy and \( \omega = 1 \) in the duplicated copy).

The fixation of the fate-determining mutation will cause another round of reduction in the polymorphism level, which is specific to the copy of the gene in which the mutation arose (in the neofunctionalization model the fate-determining mutation occurs in the new copy). If fixation occurs through positive selection, the level of polymorphism will be substantially reduced in a large region around the mutation. Over time, the polymorphism level will recover to normal levels.

This model predicts that the system starts with a preservation phase in which negative selection is in operation; however, the intensity of negative selection may be relaxed (that is, \( \omega < 1 \)) in both of the two copies in all models, except for the neofunctionalization model, in which selection is completely relaxed in the duplicated copy and the selective pressure may not change in the original copy (that is, \( \omega = \omega_o \) in the original copy and \( \omega = 1 \) in the duplicated copy). Selection should also work for functional uniformity so that gene conversion is preferred (BOX 5). The functionality and fitness of unduplicated genes that function in one subunit should be the same as when all these genes are duplicated. The function of a gene duplication preserved by dosage balance should be dependent on the dosage of at least one other gene.

Conclusions and perspectives

The relative importance of the different models for the evolution and maintenance of gene duplications is hotly debated and poorly understood. As an essential part of addressing this issue, we hope that our systematic classification will guide researchers in designing rigorous studies that aim to distinguish between the mechanisms of evolution and preservation of gene duplications.

However, there are several reasons why obtaining all of the necessary data to assign specific duplications to a specific model may be an impossible task. First, some of the data are difficult to obtain or to analyse. For example, to distinguish between the DDC and specialization models, a functional characterization of two gene copies and an ancestral or orthologous single-copy gene may be needed, which is an expensive, complex and time-consuming task even for one gene duplication. Likewise, it is difficult to statistically verify deviations of the observed \( \omega \) from the expected value, such as in asymmetry of evolution between two close gene copies. Second, to distinguish all models it is necessary to obtain information from both long-term and short-term evolution of duplicated genes, which implies that specific gene duplications that are either too similar or too distant cannot be classified with confidence. Finally, in some cases more than one model may be applicable to specific genes. Therefore, our description may be more applicable to genome-wide studies of many duplicated genes that have the benefit of applying different tests to duplications in different stages of divergence.

A defining trend of gene-duplication evolution that emerges from our classification is that the differences between the models are almost entirely confined to the first two phases of gene-duplication evolution: the fixation and fate-determination phases. The duplicated gene is unstable in either of these phases, such that the probability that the new gene copy will reach fixation or avoid eventual loss is low. For models that assume fixation by drift, the fate-determination phase is key. Functionally and evolutionarily important changes also may occur in the short term before the preservation phase is reached, so the exploration of the short-term causes and consequences of gene duplications can be informative about their long-term evolution. We therefore propose two simple questions that are feasible to approach using the tools that are currently available and that capture a large fraction of the current debate on gene duplications: first, does the fixation of a duplicate involve selection? And second, what is the major factor in the initial preservation of duplicated genes?

To approach these questions, modern studies should shift away from the almost exclusive focus on the evolutionary divergence between gene copies. Measuring sequence divergence is a simple task, but many models predict identical dynamics of sequence divergence of gene copies (TABLE 1). Alternatively, the fate-determination phase may end before a substantial number of substitutions have accumulated and therefore measures of sequence divergence are likely to be uninformative. Sampling polymorphism data in the original and new copies can be more informative than measuring sequence divergence, and has the potential to answer the first question about the presence of selection in the fixation stage (for example, see REF 74). Comparative genomic approaches allow us to distinguish between the original and new copies. Studies of gene copy-number variation have already provided interesting data on the selection\(^{75-77}\) and expression\(^{78}\) of polymorphic gene duplications. However, most copy-number variation studies so far have determined the number of copies but not the sequence of each one, limiting the capacity of these studies to contribute to the debate on the evolution of gene duplications. Such studies should also measure the polymorphism pattern around the duplication and \( \omega \) in polymorphisms that segregate in the original and new copies, which will be informative about whether selection is involved in the fixation of gene copies. In addition, a comparison of the expression levels of individuals with and without a segregating duplication is a test of gene dosage as a factor that contributes to the presence of the gene duplication. Such data can be complemented by data from single-copy orthologues of closely related species.
Determining the factors that affect the likelihood of preservation of duplicated genes seems more complex. In this respect, information on sequence divergence and polymorphism do not seem to be informative in isolation. Such data should be combined with genome-wide observations of the functional and evolutionary properties of duplicated genes or functional experiments. In particular, a systematic examination of functional differences between single-copy genes and their duplicated counterparts in two closely related species may provide a partial answer to the relevance of different models of the evolution of duplications on a genomic scale.

The largest hurdle to moving forward in our understanding of gene duplications is the lack of data on the population dynamics of copy-number variation and on the SNPs in and around such copies of genes. The acquisition and analysis of such data from targeted sequencing of polymorphic gene duplications will go a long way towards resolving the long-standing debates on the evolutionary mechanisms that determine the fate of duplicated genes.

This paper considers the influence of gene conversion on the adaptive divergence of two recently duplicated genes, and establishes the conditions under which selection can overcome the homogenizing pressure of conversion.


Acknowledgements
We thank M. Lynch for insightful comments on the manuscript.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Hideki Innan’s homepage:
http://www.sendou.soken.ac.jp/esb/innan/InnanLab
Fyodor Kondrashov’s homepage:
http://kondrashovlab.crg.cat

ALL LINKS ARE ACTIVE IN THE ONLINE PDF