Phylogenetic Analysis: Gene Duplication and Speciation

Tiratha R Singh and Ankush Bansal, Jaypee University of Information Technology, Solan, Himachal Pradesh, India

Background

Significant progress has been made in the past few decades in understanding Darwin's theory of the origin of species. Different genomic methods have helped to understand the genetic variations and gene flow that makes new species (Magadum *et al.*, 2013; Shapiro *et al.*, 2016). Although new methods have not changed the previous speculations about how species formed, they have quickened the pace of information gathering (Reams and Roth, 2015). Compiling studies on hereditary investigations would be useful to answer queries of the upcoming generations on the relative occurrence and significance of various procedures that occur during speciation (Liu *et al.*, 2016).

Many 20th century biologists viewed genes as traits of species, exquisitely tuned to current utility. This resulted in the assumption that each species should possess different genes. Gene duplication was recognized, but was implicitly assumed to have occurred recently (Rose and Oakley, 2007). Many biologists now assume that most genes have their origins in gene duplication events, which happen throughout evolutionary history. As a result, many genes form families that have persisted for hundreds of millions of years.

Gene duplication events and results of such events play a crucial role in determination of the function of novel genes. There have been various models and theories that have emerged to support the concept of gene copies (Liu *et al.*, 2016; Singh *et al.*, 2009). However, a clear picture of gene duplication events is still not clear and needs more information to come to any conclusion. Different prediction software and tools based models such as hidden Markov models give insight to evolutionary functional properties and dynamics (Singh and Pardasani, 2009). Hence, understanding the gene duplication events and speciation is an essential step towards understanding and identifying the major mechanisms that are involved in the evolution (Seehausen *et al.*, 2014).

Gene Duplication

The evolutionary understanding of gene duplication events was first performed by Haldane and John (1932), who suggested that a redundant duplicate(s) of a gene may acquire divergent mutations and eventually emerge as a new gene. A gene duplication event was first noted by Bridges (1936) in the Bar locus in *Drosophila*. A substantial increase in the number of copies of a DNA segment can be brought by various types of gene duplication (White, 1977; Raj Singh, 2008). There are various studies where gene duplication and deletion events are being systematically observed (Ma *et al.*, 2014; Schacherer *et al.*, 2004; Simillion *et al.*, 2002; Stephens, 1951). Many types of duplications are recognized: First, partial gene duplication; second, complete gene duplication; third, partial chromosomal duplication; fourth, complete chromosomal duplication; fifth, genome duplication (Innan and Kondrashov, 2010; Conant and Wolfe, 2008; Panchy *et al.*, 2016). The first four are treated as regional duplicates as they do not alter the haploid set of chromosomes. The main reason for gene duplication includes uneven crossing over (Iñiguez and Hernández, 2017; Levasseur and Pontarotti, 2011; Qian and Zhang, 2014). Uneven crossing over in two nonaligned sequences gives a duplicated region on one chromosome along with deletion on second on the basis of the size of the nonaligned region. DNA sequence duplication in tandem results in a progressive increase in uneven crossing over, which ultimately increases the number of duplicate copies (Mendivil Ramos and Ferrier, 2012).

Partial Gene Duplication

Tandem duplication events in DNA sequences may provide information about genetic evolution events in terms of the complete gene while tandem duplication in a small region, or maybe in part of gene, ultimately results in mutations and is the cause of various diseases (Hu and Worton, 1992; Hu *et al.*, 1988). The duplication arrangement can be understood by taking the inference from molecular information lying in the sequence (Toll-Riera *et al.*, 2011). For instance, structure level changes induced by changes in nucleotide and amino acid sequence level influences the protein evolution. Toll-Riera *et al.* (2011) have demonstrated this by considering a large dataset of human and mouse orthologs protein and later mapping with PDB structures. Evidence from literature suggests that duplication may arise from either homologous (Alu-Alu) recombination or nonhomologous recombination, the latter possibly mediated by topoisomerases. For the dystrophin gene, in which most duplications have been identified, these recombination events are intrachromosomal, which suggests that unequal sister chromatid exchange is the major mechanism (Toll-Riera *et al.*, 2011).

Domain Duplication and Gene Elongation

A domain is a well-defined region within a protein that either performs a specific function within a protein, such as substrate binding, or constitutes a stable, independently folding, compact structural unit within the protein that can be distinguished from

all the other parts" (Li and Makova, 2001). Theoretically, several possible relationships may be envisioned between the structural domains and the arrangements of the exons in the gene, e.g., in many globular proteins, a more or less exact correspondence exists between exons of gene and the structural domains of the protein product. Alternate splicing is one of the main reasons for exon shuffling in a duplication event as there are always chances for repetition of a similar exon set again. In a considerable number of cases, several adjacent models were found to be encoded by the same exon (Li and Makova, 2001).

The vertebrate hemoglobin α and β chains, consist of four domains, whereas their genes consist of only three exons, the second of which encodes two adjacent domains. In *Caenorhabditis elegans*, a globin-encoding gene, during the evolution of a globin gene family from a four exon ancestral gene, several lineages lost some or all of their three introns, thereby, generating panoply of exon–intron permutations (Vogel *et al.*, 2005). In the majority of cases, a domain duplication at the protein level indicates that an exon duplication has occurred at the DNA level. Moreover, many proteins of present day organisms show internal repeats often correspond to functional or structural domains within the proteins. A survey of modern genes in eukaryotes shows that the internal duplications have occurred frequently in evolution. This gene duplication is one of the most important steps in the evolution of complex genes from the simple ones (Vogel *et al.*, 2005).

Theoretically, elongation of genes can also occur by other means; for example, a mutation change converting a stop codon into a sense codon can also elongate the gene, which could be a part of recoding event (Singh and Pardasani, 2009). Similarly, either insertion of a foreign DNA segment into an exon or the occurrence of a mutation obliterating a splicing site will achieve same result. These types of molecular changes most probably disrupt the function of the elongated gene. In the vast majority of cases, such molecular changes have been found to be associated with pathological manifestations. By contrast, duplication of a structural domain is less likely to be problematic. Indeed, such a duplication can sometimes even enhance the function of the protein produced for example by increasing the number of active sites (a quantitative change), thus enabling the gene to perform its function more rapidly and efficiently or by having a synergistic effect yielding a new function (a qualitative change) (Nacher *et al.*, 2010).

Emergence of novel function can be derived from partial gene as divergence in the sequence may lead to different functions. Complete gene duplication produces two identical paralogous copies (Nacher *et al.*, 2010). Duplicated genes can be divided into two types – Variant and invariant repeats. Invariant repeats are identical or nearly identical in sequence to one another. Variant repeats are copies of a gene that, although similar to each other, differ in their sequences to a lesser or greater extent. All the genes that belong to a certain group of repeated sequences in a genome are referred to as a gene or a multigene family. Functional and nonfunctional members of a gene family may reside in close proximity to one another on the same chromosome or they may be located on different chromosomes. A member of a gene family that is located alone at a different genomic location than the other members of the family is called an orphan. The term *superfamily* was coined by Dayhoff in order to distinguish closely related proteins from distantly related ones (Dayhoff and Schwartz, 1978). Proteins that exhibit at least 50% similarity to each other at the amino acid level are considered as members of a superfamily, for example, α and β globins are classified into two separate families and together with myoglobin they form the globin superfamily (Li and Makova, 2001). An important feature associated with gene duplication is that as long as two or more copies of a gene exist in proximity to each other, the process of gene duplication can be greatly accelerated in this region, and numerous copies may be produced. There may be two reasons for the general positive correlation between genome size and number of copies of RNA specifying genes. Either a large genome requires large quantities of RNA, or the number of RNA-specifying genes is simply a passive consequence of genome enlargement by duplication. Highly repetitive genes, like rRNA genes, are generally very similar to each other (Li and Makova, 2001).

One factor responsible for homogeneity may be purifying selection. In addition to invariant repeats, the genomes of higher organisms contain numerous multigene families whose members have diverged to various extents such as genes coding for isozymes, such as lactate dehydrogenase, aldolase, creatine kinase, carbonic anhydrase, and pyruvate kinase. Isozymes are enzymes that catalyze the same biochemical reaction but may differ from one another in tissue specificity, developmental regulation, electrophoretic mobility, or biochemical properties. Isozymes are encoded by different loci, usually duplicated genes, as opposed to allozymes, which are distinct forms of the same enzyme encoded by different alleles at a single locus (Li and Makova, 2001; Vogel *et al.*, 2005).

Exon Shuffling

There are three types of exon shuffling: (1) Exon duplication, (2) exon insertion, and (3) exon deletion. Exon duplication refers to the duplication of one or more exons in a gene and so is a type of internal duplication (Patthy, 1999). Exon insertion is the process by which structural or functional domains are exchanged between proteins or inserted into a protein. Exon deletion results in the removal of a segment of amino acids from the protein. All types of shuffling have occurred in the evolutionary process of creating new genes (Kolkman and Stemmer, 2001; Patthy, 1999). Exon shuffling could be represented through various phases as shown in **Fig. 1**.

Mosaic or Chimeric Proteins

A mosaic or chimeric protein is a protein encoded by a gene that contains regions that are also found in other genes (Nicolson, 2015; Singer and Nicolson, 1972). The existence of such proteins indicates that exon shuffling has occurred during the evolutionary history of their genes. The first described mosaic protein was tissue plasminogen activator (Iñiguez and Hernández, 2017). For an exon to be inserted, deleted, or duplicated without causing a frameshift in the reading frame, certain phase limitations of the exonic structure of the gene must be respected. Mosaic proteins can be made when two adjacent genes are transcribed together and are therefore made into the same protein.

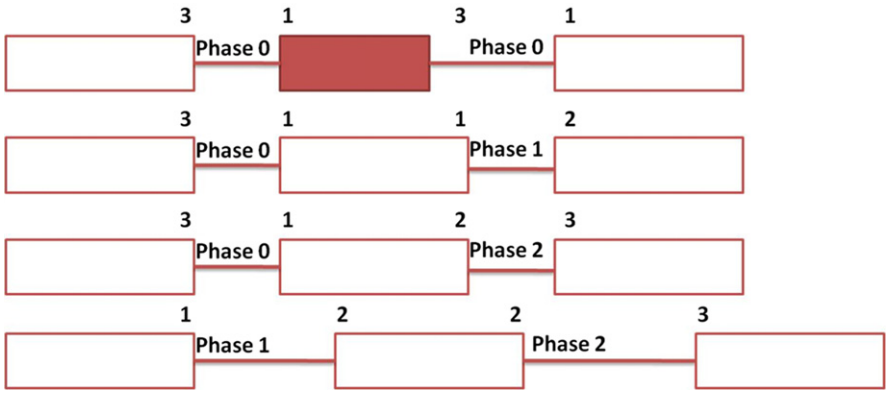

Fig. 1 Schematic representationof exon shuffling. Numbers above the box represents the position of nucleotide.

To understand this, we need to consider introns in terms of their possible positions relative to the coding regions (Nicolson, 2014). Introns residing between coding regions are classified into three types according to the way in which the coding region is interrupted (Jeon, 2004). An intron is of phase 0 if it lies between two codons, of phase 1 if it lies between the first and second nucleotide of a codon, and of phase 2 if it lies between the second and third nucleotides of a codon. Exons are grouped into classes according to the phases of their flanking introns. Here are four middle exons, said to be (0–0), (0–1), (0–2), and (1–2) exons. An exon that is flanked by introns of the same phase at both ends is called a symmetrical exon, otherwise it is asymmetrical. The first exon (middle) is a symmetrical exon, represented by black box. The length of a symmetrical exon is always a multiple of three nucleotides. Only symmetrical exons can be duplicated in tandem or deleted without affecting the reading frame.

Duplication or deletion of asymmetrical exons would disrupt the reading frame downstream. Similarly, only symmetrical exons can be inserted into introns, but with the restriction of, a 0–0 exon can only be inserted in phase-0-introns, a 1–1 exon is inserted into phase 1 introns, and 2–2 exons into phase-2 introns for avoiding frameshifts. All the exons coding for the modules of mosaic proteins are symmetrical. Since nonrandom intron phase usage is a necessary consequence of exon duplication or insertion, this property may be used as a diagnostic feature of gene assembly through exon shuffling. In terms of splicing, introns are classified into two categories, self-splicing and spliceosomal (Wang *et al.*, 2013). The vast majority of introns in eukaryotic nuclear genes are spliceosomal. Self-splicing introns play a vital role in their own removal, some regions of the introns are involved in self-complimentary interactions important for forming the 3-D structure possessing splicing activity. Exon shuffling probably did not play a role in the formation of genes in the early stages of evolution. Exon shuffling came to full bloom with the evolution of spliceosomal introns, which do not play a role in their self excision. These introns contain mainly nonessential parts and therefore could accommodate quantities of "foreign" DNA (Wang *et al.*, 2013).

Exonization and Pseudoexonization

Exonization is the process through which an intronic sequence becomes an exon. An exon created by exonization must abide by the same rules of exon insertion (Schmitz and Brosius, 2011). The opposite process is called pseudoexonization. It occurs when nonfunctionalization affects a single exon rather than the entire gene. The result is the creation of a pseudoexon, and the most obvious consequence of such a process is gene abridgement (as opposed to gene elongation). Pseudoexons are often created by the nonfunctionalization of internal gene duplications, for example, the aggrecan gene in rat contains 18 repeated exons and one pseudoexon. Some complex biological functions that require several enzymes may be specified by genes encoding different combinations of protein modules. In some species we may find single-module proteins, while in others we may find different combinations of multimodular proteins, for example, genes in a multistep process in the synthesis of fatty acids from Acetyl Co-A require seven enzymatic activities and an acyl–carrier protein. In fungi, these activities are distributed between two nonidentical polypeptides encoded by two unlinked intronless genes, FAS1 and FAS2. FAS1 encodes two and FAS2 encodes five enzymatic activities. In most bacteria, however, these functions are carried on by discrete monofunctional proteins (Schweizer and Hofmann, 2004).

In animals, key functions associated with fatty acid metabolism are controlled with a single polypeptide chain called fatty acid synthase. The fatty acid synthase gene in fungi and mammals are most probably mosaic proteins that have assembled from single-domain proteins like the ones found in bacteria (Stower, 2013). The fact that arrangement of domains is different in fungi from that in mammals indicates not only that the two lineages evolved multimodularity independently, but also that different strategies may be employed in the assembly of genes encoding multimodular proteins.

Nested and Overlapping Genes and Their Association With Speciation

In addition to gene duplication and exon shuffling, many other mechanisms for producing new genes or polypeptides are available. Few such entities are overlapping genes, pseudogenes, and nested genes.

Overlapping Genes

A DNA fragment (segment) can code for more than one gene product by using different reading frames or different initiation codons. This phenomenon of overlapping genes is widespread in DNA and RNA viruses, as well as in organelles, and bacteria, also known in nuclear eukaryotic genomes (Makalowska *et al.*, 2005). Overlapping genes can also arise by the use of the complementary strand of a gene; for example, genes specifying $tRNA^{ILE}$ and$tRNA^{GLN}$ in the human mitochondrial genome are located on different strands and there is a three-nucleotide overlap between these that reads 5′-CTA-3′ in the former and 5′-TAG-3′ in the latter (Makalowska *et al.*, 2005). The rate of evolution is expected to be slower in stenches of DNA encoding overlapping genes than in similar DNA sequences that only use one reading frame. The reason is that the proportion of nondegenerate sites is higher in overlapping genes than in nonoverlapping genes, thus vastly reducing the proportion of synonymous mutations out of total number of mutations (Johnson and Chisholm, 2004; Normark *et al.*, 1983). Since gene duplication is a widespread phenomenon for the maintenance of overlapping genes, it would require quite strong selective pressure (against increasing genome size). Studies on aminoacyl tRNA synthetases indicate that overlapping genes may have played a momentous role in the evolution of life (Johnson and Chisholm, 2004; Normark *et al.*, 1983).

Alternate Splicing

Alternative splicing of a primary RNA transcript results in the production of different mRNAs from the same DNA segment, which in turn may be translated into different polypeptides (Baralle and Giudice, 2017). There are two types of exons: Constitutive, that is, exons that are included within all the mRNAs transcribed from a gene, and facultative,that is, exons that are sometimes spliced in and sometimes spliced out (Kornblihtt *et al.*, 2013). There are different types of alternative splicing; the most trivial form is the intron retention (Lee and Rio, 2015). However, more commonly, intron retention results in the premature termination of translation due to frameshifts. Sometimes, alternative splicing involves the use of alternative internal donor or acceptor sites, that is, excisions of introns of different lengths with complementary variation in the size of neighboring exons. Such use of competing splice sites was found in several transcription units of adenoviruses, as well as in eukaryotic cells such as the transformer gene in *Drosophila melanogaster* (Lee and Rio, 2015). Some cases of alternative splicing involved the use of mutually exclusive exons, that is, two exons are never spliced out together, nor are both retained in the same mRNA for example, M1 and M2 from a single gene by mutually exclusive use of exons 9 and 10. A special case of mutual exclusivity is the cassette exon (Roy *et al.*, 2013). A cassette is either spliced in or spliced out in the alternative mRNA molecules. Alternative splicing has often been used as a means of developmental regulation (Wang *et al.*, 2015). A very intriguing situation is seen in several genes involved in the process of sex determination in *D. melanogaster*. At least three genes, doublesex (dsx), Sexlethal (sxl), and transformer (tra), are spliced differently in males and females (Wang *et al.*, 2015). There is rich literature available on alternate splicing and its distribution in almost all available lineages.

Intron-Encoded Proteins and Nested Genes

An intron may sometimes contain an ORF that encodes a protein or part of a protein that is completely different in function from the one encoded by the flanking exons (Kumar, 2009). In many cases, intron-encoded protein genes are located within type-I self-splicing introns. From a mechanistic point of view, an intron-encoded protein gene that is transcribed from the same strand as the neighboring exons may be regarded as special instance of alternative splicing (Lee and Chang, 2013; Yu *et al.*, 2005). When an intron-encoded protein gene is transcribed from the opposite strand of the other gene, it is referred to as a nested gene. A case of nested genes was found in *Drosophila*, where a pupal cuticle protein gene is encoded on the positive strand of an intron within the gene encoding the purine pathway enzyme glycinamide ribotide-transformylase (Kaer *et al.*, 2011).

Functional Convergence

Function of a protein is frequently determined by only a few of its amino acids; a protein performing one function may sometimes arise from a gene encoding a protein performing a markedly different function. If the new function is performed in other species by proteins of unrelated structure and descent, functional convergence may occur. The myoglobin of abalone *Sulculus* consists of 377 amino acids, which are 2.5 times larger than myoglobins belonging to the globin superfamily (Suzuki *et al.*, 1996). Functional convergence provides a robust parameter associated with the existence of a functional protein for a family or superfamily. This convergence is reflected in various levels of organisms at the family and superfamily level based upon rate of selection pressure.

RNA Editing

RNA editing is a molecular process by which protein-coding gene change its message. One of the most common types of RNA editing is C-to-U conversion. This conversion may occur partially or completely in some tissues but not in others, leading to differential gene expression. Occasionally, it can produce a new protein with a different function from the unedited transcript, foe example, apolipoprotein B gene, one of the lipid carriers in the blood (Cooper, 1999). There are two types – Apo B-100 and apo B-48. Despite differences in length, amino acid sequences of the gigantic protein apo B-100 (4536 amino acid) with that of

apo B-48 (2152 amino acid), the result of alignment for the alignable part is 100% identity. It was found that apo B-48 is translated from a very long mRNA that is identical to that of apo B-100 with the execution of an in-frame stop codon resulting from the RNA editing of codon 2153 from CAA (Gln) to UAA (stop). Thus, by using RNA editing, two quite different proteins are produced from the same gene (Cooper, 1999). (**Figs. 2,3,4,5**).

Gene Sharing

Gene sharing means that a gene acquires and maintains a second function without divergent duplication and without loss of the primary function. Gene sharing may, however, require a change in the regulation system of tissue specificity or developmental timing (Cvekl and Zheng, 2009; Patthy, 2007). In literature, the term "multifunctional protein" is frequently used instead of "gene sharing." Gene sharing was first discovered in crystallins, which are the major water-soluble proteins in the

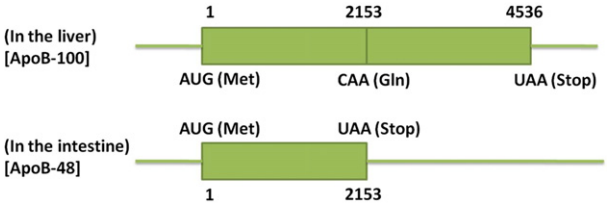

Fig. 2 Mechanism of RNA editing may leads to truncation.

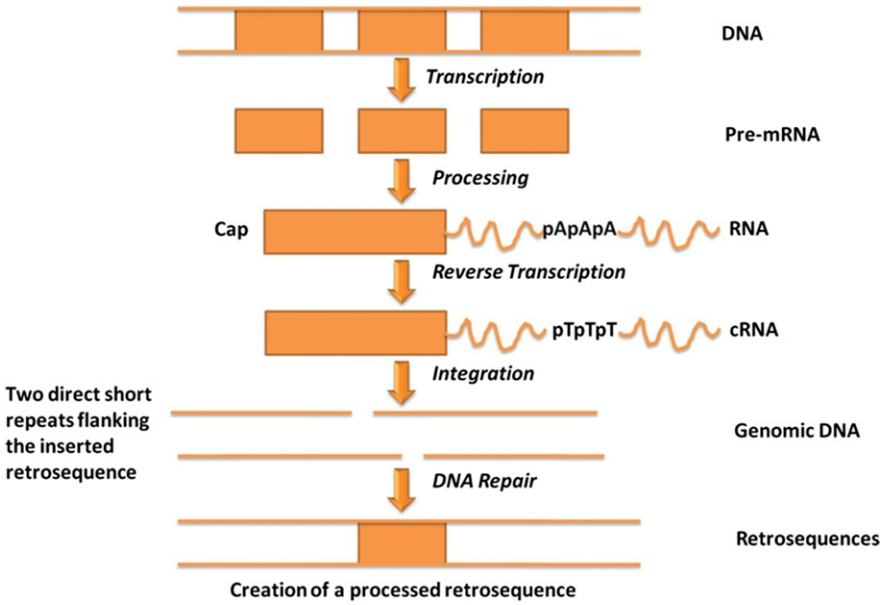

Fig. 3 Molecular mechanism explaining molecular repair.

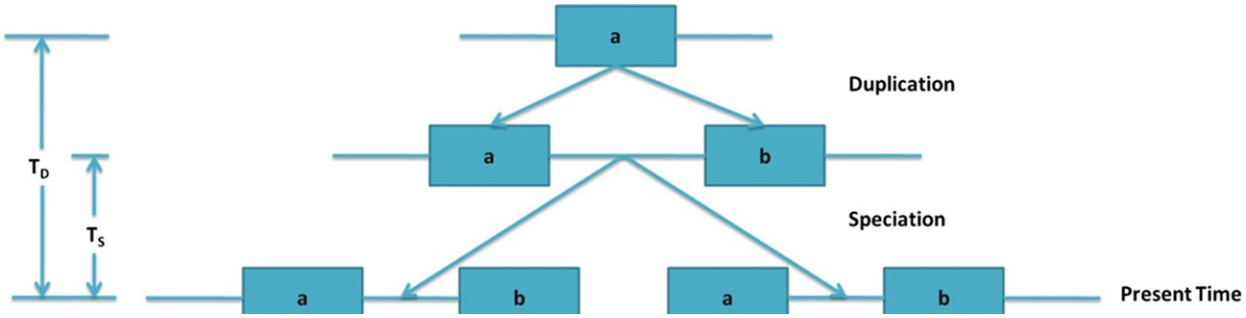

Fig. 4 Duplication and Speciation Events.

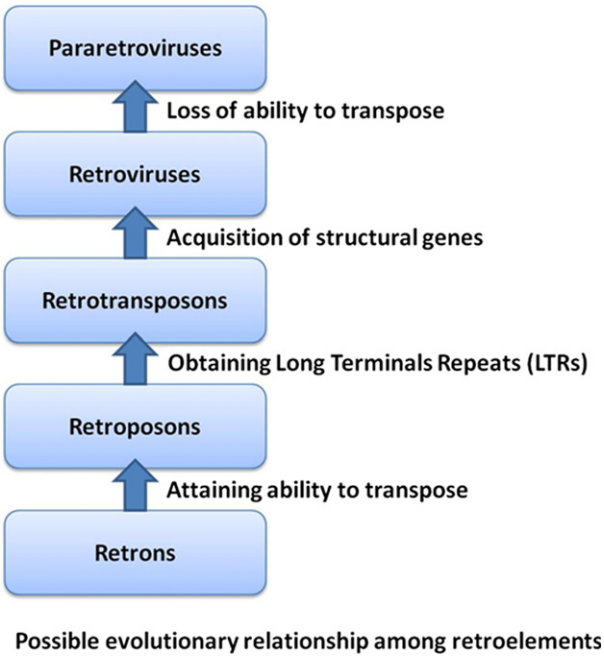

Fig. 5 Evolutionary relationship among retro elements.

eye lens, and whose function is to maintain lens transparency and proper light refraction (Cvekl and Zheng, 2009). Gene sharing might be a fairly common phenomenon, also suspected for several proteins in the cornea and other tissues. Gene sharing clearly adds to the compactness of the genome, even though compactness does not seem to have a high priority in eukaryotes (McKay, 2008).

Molecular Repair

The more we learn about the evolution of genes, the more we recognize that true innovations are only rarely produced during evolution. Many proteins that were originally considered to be relatively recent evolutionary additions turned out to be derived from ancient proteins (McKay, 2008). Besides, all the discussed mechanisms that facilitate tinkering at the molecular level are gene conversion and transposition. We may deduce that molecular tinkering is most probably the paradigm of molecular evolution, and it is reasonable to assume that tinkering also characterizes the evolution of morphological, anatomical, and physiological traits as well.

Gene Gain and Loss and Speciation Events

Gene Loss

More than 7000 diseases and disorders were documented in research papers and literature sources, which state that mutations have a crucial role in destroying and gaining function at the gene level (McKusick, 2007). A number of such mutations either get deleted at a very fast rate from population or sustained at very low frequency due to genetic drift. If there are many copies of genes present and functions normally then deleterious mutations which occur more often collate together compared to significant ones. Repeated duplicate genes usually become nonfunctional instead of being a new gene (Stone *et al.*, 1998).

Lateral Gene Transfer

Lateral gene transfer (LGT) is a process that makes complicated distribution of genes and dissimilar phylogenies with an rRNA tree. Still there is a debate on robust organismal phylogeny over extensive LGT (Kettler *et al.*, 2007). In case there is a presence of a core set of genes that are resistant to LGT, then there should be a reflection of vertical descent and ascent along with cell division. Moreover, various next generation sequencing techniques like metagenomics, genomics, and proteomics pave a path in understanding the core mechanism of LGT. In particular, it will be informative to know the complete genome diversity (Kettler *et al.*, 2007).

Dating Gene Duplication and other Computations

Dating Gene Duplications

Two genes are said to be paralogous if they are derived from a duplication event, but orthologous if they are derived from a speciation event.

Here, genes a and b were derived from the duplication of an ancestral gene and are paralogous, while gene a from species 1 and gene a from species 2 are orthologous, as are genes b from species 1 and gene b from species 2 (Chen *et al.*, 2000). We can estimate the date of duplication, T_D, from sequence data if we know the rate of substitution in genes a and b. The rate of substitution can be estimated from the number of substitutions between the orthologous genes in conjunction with knowledge of the time of divergences T_S, between species 1 and 2. For gene a, let K_a be the number of substitutions per site between the two species. Then the rate of substitution in gene a, γ_a is estimated by:

$$\gamma a = \frac{\mathrm{Ka}}{2\mathrm{TS}} \tag{1}$$

The rate of substitution in gene b, γ_b can also be obtained in a similar way. The average substitution rate for the two genes:

$$\gamma = \frac{\gamma a + \gamma b}{2} \tag{2}$$

To estimate T_D, we need to know the number of substitutions per site between gene a and b (Kab). This number can be estimated from four pairwise comparisons: (1) Gene a from species 1 and gene b from species 2, (2) (Robinson-Rechavi *et al.*, 2004). Gene a from species 2 and gene b from species 1. (3) Gene a and gene b from species 1. (4) Gene a and b from species 2. From these four estimates, we can compute the average value for K ab from which we can estimate T_D as:

$$\mathrm{TD} = \frac{\mathrm{Kab}}{2\gamma} \tag{3}$$

In case of protein coding genes, by using the number of synonymous and nonsynonymous substitutions separately, we can obtain two independently estimates of T_D. The average of these two may be used as the final estimate of T_D. Sometimes problems are due to concerted T_D estimation. Another method for dating gene duplication events is to consider the phylogenetic distribution of genes in conjunction with paleontological data pertinent to the divergence date of the species in question (Zhou *et al.*, 2010).

Unprocessed Pseudogenes

The silencing of a gene due to deletion in a nucleotide and causing deleterious mutation ultimately results in production of pseudogene. Such pseudogenes do not undergo RNA processing. Unprocessed pseudogenes may be the result of derivation from nonfunctionality of a duplicate functional gene. There is much less chance that functional genes come into existence without duplication. Unprocessed pseudogenes may result in various diseases and disorders like frameshift mutation, nonmature stop codon, and misorientation of sliced transcripts or regulatory elements; therefore, it is easy to identify a change in sequence in terms of mutations resulting directly in gene silencing (Tutar, 2012). It is possible in some cases to identify the mutation responsible for nonfunctionalization of a gene through a phylogenetic analysis, for example, human pseudogene $\psi\eta$ in the b-globin family contains numerous defects, each of which could have been sufficient to silence it (Pink *et al.*, 2011). The b-globin clusters in chimpanzee and gorilla were found to contain the same number of genes and pseudogenes as in humans, indicating that the pseudogenes were created and silenced before these three species diverged from one another.

Interestingly, mutations that cause nonfunctionalization are only rarely missense mutations, most probably because such mutations result in the production of defective proteins that may be incorporated into final biological products and thus may have deleterious effects. Because unprocessed pseudogenes are usually created by duplication, they usually found in the neighborhood of the homologous functional genes from which they have been derived.

Unitary Pseudogenes

The pseudogene has no functional correlation with the human genome, called unitary pseudogene. Guinea pigs and humans suffer from scurvy unless they consume L-ascorbic acid in their diet, because they lack a protein called L-gulono-y-lactone oxidase, an enzyme that catalyzes the terminal step in L-ascorbic acid synthesis (Pink *et al.*, 2011). In humans, L-gulono-y-lactone oxidase is a pseudogene that contains molecular defects as the deletion of at least two exons (out of 12), deletions and insertion of nucleotides in the reading frame, and obliterations of intron–exon boundaries (Pink *et al.*, 2011). It has been assumed that the guinea pig and human ancestors managed to survive on a naturally ascorbic acid-rich diet; hence, the loss of this enzyme did not reflect a disadvantage.

Case Studies

There are several case studies where bioinformatics has been implemented successfully to connect molecular evolution events with their functional consequences. Genomics studies paved the path to understand variation at the genomic level through phylogenetic approaches (Pradhan *et al.*, 2017; Wang *et al.*, 2018). Further network reconstruction approaches using co-expression network modules also helps to trace the gene duplication events (Feng *et al.*, 2016; Malviya *et al.*, 2016). For instance, phylogenetic investigation of human FGFR-bearing paralogons favors piecemeal duplication theory of vertebrate genome evolution (Ajmal *et al.*, 2014). All given studies indicate that duplication and speciation events are very much diverse in nature and take many years to evolve and confirm to contribute to significant changes.

Tools and Methods

S.No.	*Tool or method*	*Description*	*References*
1.	Control-FREEC	CNV detection: HTS analysis	A tool for detection of copy-number changes and allelic imbalances (including LOH) using deep-sequencing data. Control-FREEC automatically computes, normalizes, segments copy number and beta allele frequency profiles, then calls copy number alterations and LOH. The control (matched normal) sample is optional. The program can also use mappability data (files created by GEM).	Boeva *et al.* (2012, 2011)
2.	mrCaNaVaR	micro-read Copy number variant regions	A copy number caller that analyzes the whole-genome next-generation sequence mapping read depth to discover large segmental duplications and deletions. mrCaNaVaR also has the capability of predicting absolute copy numbers of genomic intervals.	Alkan *et al.* (2009)
3.	forestSV	Structural variant detection: HTS analysis	Integrates prior knowledge about the characteristics of SVs. forestSV is a statistical learning approach, based on Random Forests, that leads to improved discovery in high throughput sequencing (HTS) data. This application offers high sensitivity and specificity coupled with the flexibility of a data-driven approach. It is particularly well suited to the detection of rare variants because it is not reliant on finding variant support in multiple individuals.	Michaelson and Sebat (2012)
4.	CTDGFinder	Duplication detection: HTS analysis	Formalizes and automates the identification of clusters of tandemly duplicated genes (CTDGs) by examining the physical distribution of individual members of families of duplicated genes across chromosomes. Application of CTDGFinder accurately identified CTDGs for many well-known gene clusters (e.g., Hox and beta-globin gene clusters) in the human, mouse, and 20 other mammalian genomes. Examination of human genes showing tissue-specific enhancement of their expression by CTDGFinder identified members of several well-known gene clusters (e.g., cytochrome P450s and olfactory receptors) and revealed that they were unequally distributed across tissues. By formalizing and automating CTDG identification, CTDGFinder will facilitate understanding of CTDG evolutionary dynamics, their functional implications, and how they are associated with phenotypic diversity.	Ortiz and Rokas (2017)
5.	cn.MOPS	Copy number estimation by a Mixture Of PoissonS	A data processing pipeline for copy number variations and aberrations (CNVs and CNAs) from next generation sequencing (NGS) data. The package supplies functions to convert BAM files into read count matrices or genomic ranges objects, which are the input objects for cn.MOPS. It models the depths of coverage across samples at each genomic position. Therefore, it does not suffer from read count biases along chromosomes. Using a Bayesian approach, cn.MOPS decomposes read variations across samples into integer copy numbers and noise by its mixture components and Poisson distributions, respectively.	Klambauer *et al.* (2012)
6.	RDXplorer	CNV detection: HTS analysis	A computational tool for copy number variants (CNV) detection in whole human genome sequence data using read depth (RD) coverage. CNV detection is based on the event-wise testing (EWT) algorithm. The read depth coverage is estimated in nonoverlapping intervals (100 bp Windows) across an individual genome based on the pileup generated by SAMTools.	Yoon *et al.* (2009)
7.	cnvHiTSeq	CNV detection: HTS analysis	A set of Java-based command-line tools for detecting copy number variants (CNVs) using next-generation sequencing data.	Bellos *et al.* (2012)

Source: *This data is referred from OMICSTOOLS resource. There are various other tools which may be referred at https://omictools.com/duplication-detection-category.

Conclusion

Gene duplication and speciation are two prominent mechanisms for finding clues for evolution. Phylogenetic analysis helps researchers to understand the ancestral association of species or sequence of their interest. Gene duplication events, exon shuffling,

and speciation have a potent role in the process of evolution and to study convergence and divergence from ancestral data. This article provides descriptive information about basic concepts of phylogeny that may help students and researchers to become aware of terminologies used in gene duplication and speciation analysis. It is presented in a way where basic to advanced level information is being compiled on the diverse topics and it is estimated that this information will serve as comprehensive information to students, faculty members, and researchers working in this area. It reflects how bioinformatics can shape a computational pipeline for the analysis of biological data in an evolutionary scenario and will help the evolutionary biologists also to implement bioinformatics at some level for further exploration of basic principles.

References

Ajmal, W., Khan, H., Abbasi, A.A., 2014. Phylogenetic investigation of human FGFR-bearing paralogons favors piecemeal duplication theory of vertebrate genome evolution. Molecular Phylogenetics and Evolution 81, 49–60. Available at: https://doi.org/10.1016/j.ympev.2014.09.009.

Alkan, C., Kidd, J.M., Marques-Bonet, T., *et al.*, 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics 41, 1061–1067. Available at: https://doi.org/10.1038/ng.437.

Baralle, F.E., Giudice, J., 2017. Alternative splicing as a regulator of development and tissue identity. Nature Reviews Molecular Cell Biology 18, 437–451. Available at: https://doi.org/10.1038/nrm.2017.27.

Bellos, E., Johnson, M.R., M. Coin, L.J., 2012. cnvHiTSeq: Integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. Genome Biology 13, R120. Available at: https://doi.org/10.1186/gb-2012-13-12-r120.

Boeva, V., Popova, T., Bleakley, K., *et al.*, 2012. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics (Oxford, England) 28, 423–425. Available at: https://doi.org/10.1093/bioinformatics/btr670.

Boeva, V., Zinovyev, A., Bleakley, K., *et al.*, 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics (Oxford, England) 27, 268–269. Available at: https://doi.org/10.1093/bioinformatics/btq635.

Bridges, C.B., 1936. The bar "gene" a duplication. Science 83, 210–211. Available at: https://doi.org/10.1126/science.83.2148.210.

Chen, K., Durand, D., Farach-Colton, M., 2000. NOTUNG: A program for dating gene duplications and optimizing gene family trees. Journal of Computational Biology 7, 429–447. https://doi.org/10.1089/106652700750050871.

Conant, G.C., Wolfe, K.H., 2008. Turning a hobby into a job: How duplicated genes find new functions. Nature Reviews Genetics 9, nrg2482. Available at: https://doi.org/10.1038/nrg2482.

Cooper, D.N., 1999. Human Gene Evolution. Elsevier.

Cvekl, A., Zheng, D., 2009. Gene sharing and evolution. Human Genomics 4, 66–67. Available at: https://doi.org/10.1186/1479-7364-4-1-66.

Dayhoff, M.O., Schwartz, R.M., 1978. Chapter 22: A model of evolutionary change in proteins, in: In Atlas of Protein Sequence and Structure.

Feng, K., Liu, F., Zou, J., *et al.*, 2016. Genome-wide identification, evolution, and co-expression network analysis of mitogen-activated protein kinase kinase kinases in *Brachypodium Distachyon*. Frontiers in Plant Science. 7, 1400. Available at: https://doi.org/10.3389/fpls.2016.01400.

Haldane, J.B.S., John, B.S., 1932. The Causes of Evolution. London: Longmans, Green.

Hu, X.Y., Burghes, A.H., Ray, P.N., *et al.*, 1988. Partial gene duplication in Duchenne and Becker muscular dystrophies. Journal of Medical Genetics 25, 369–376.

Hu, X., Worton, R.G., 1992. Partial gene duplication as a cause of human disease. Human Mutation 1, 3–12. Available at: https://doi.org/10.1002/humu.1380010103.

Iñiguez, L.P., Hernández, G., 2017. The evolutionary relationship between alternative splicing and gene duplication. Frontiers in Genetics 8.Available at: https://doi.org/10.3389/fgene.2017.00014.

Innan, H., Kondrashov, F., 2010. The evolution of gene duplications: Classifying and distinguishing between models. Nature Reviews Genetics 11, nrg2689. Available at: https://doi.org/10.1038/nrg2689.

Jeon, K.W., 2004. International Review of Cytology: A Survey of Cell Biology. Academic Press.

Johnson, Z.I., Chisholm, S.W., 2004. Properties of overlapping genes are conserved across microbial genomes. Genome Research 14, 2268–2272. Available at: https://doi.org/10.1101/gr.2433104.

Kaer, K., Branovets, J., Hallikma, A., Nigumann, P., Speek, M., 2011. Intronic L1 retrotransposons and nested genes cause transcriptional interference by inducing intron retention, exonization and cryptic polyadenylation. PLOS ONE 6, e26099. Available at: https://doi.org/10.1371/journal.pone.0026099.

Kettler, G.C., Martiny, A.C., Huang, K., *et al.*, 2007. Patterns and Implications of gene gain and loss in the evolution of prochlorococcus. PLOS Genetics 3, e231. Available at: https://doi.org/10.1371/journal.pgen.0030231.

Klambauer, G., Schwarzbauer, K., Mayr, A., *et al.*, 2012. cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Research 40, e69. Available at: https://doi.org/10.1093/nar/gks003.

Kolkman, J.A., Stemmer, W.P.C., 2001. Directed evolution of proteins by exon shuffling. Nature Biotechnology 19.Available at: https://doi.org/10.1038/88084.

Kornblihtt, A.R., Schor, I.E., Alló, M., *et al.*, 2013. Alternative splicing: A pivotal step between eukaryotic transcription and translation. Nature Reviews Molecular Cell Biology 14, 153–165. Available at: https://doi.org/10.1038/nrm3525.

Kumar, A., 2009. An overview of nested genes in Eukaryotic genomes. Eukaryotic Cell 8, 1321–1329. Available at: https://doi.org/10.1128/EC.00143-09.

Lee, Y.C.G., Chang, H.-H., 2013. The evolution and functional significance of nested gene structures in Drosophila melanogaster. Genome Biology and Evolution 5, 1978–1985. Available at: https://doi.org/10.1093/gbe/evt149.

Lee, Y., Rio, D.C., 2015. Mechanisms and regulation of alternative pre-mRNA splicing. Annual Review of Biochemistry 84, 291–323. Available at: https://doi.org/10.1146/annurev-biochem-060614-034316.

Levasseur, A., Pontarotti, P., 2011. The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biology Direct 6, 11. Available at: https://doi.org/10.1186/1745-6150-6-11.

Liu, Z., Tavares, R., Forsythe, E.S., *et al.*, 2016. Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. Nature Communications 7. Available at: https://doi.org/10.1038/ncomms13026.

Li, W.-H., Makova, K.D., 2001. Domain Duplication and Gene Elongation. ELS. John Wiley & Sons, Ltd. Available at: https://doi.org/10.1038/npg.els.0005097.

Malviya, N., Jaiswal, P., Yadav, D., 2016. Genome- wide characterization of Nuclear Factor Y (NF-Y) gene family of sorghum [Sorghum bicolor (L.) Moench]: a bioinformatics approach. Physiol. Mol. Biol. Plants 22, 33–49. Available at: https://doi.org/10.1007/s12298-016-0349-z.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., Ravikesavan, R., 2013. Gene duplication as a major force in evolution. Journal of Genetics 92, 155–161.

Makalowska, I., Lin, C.-F., Makalowski, W., 2005. Overlapping genes in vertebrate genomes. Computational Biology and Chemistry 29, 1–12. Available at: https://doi.org/10.1016/j.compbiolchem.2004.12.006.

Ma, Q., Reeves, J.H., Liberles, D.A., *et al.*, 2014. A phylogenetic model for understanding the effect of gene duplication on cancer progression. Nucleic Acids Research 42, 2870–2878. Available at: https://doi.org/10.1093/nar/gkt1320.

McKay, S.A.B., 2008. Gene sharing and evolution: The diversity of protein functions. By Joram Piatigorsky. The Quarterly Review of Biology 83, 99. Available at: https://doi.org/10.1086/586921.
McKusick, V.A., 2007. Mendelian inheritance in man and its online version, OMIM. American Journal of Human Genetics 80, 588–604.
Mendivil Ramos, O., Ferrier, D.E.K., 2012. Mechanisms of gene duplication and translocation and progress towards understanding their relative contributions to animal genome evolution [WWW Document]. International Journal of Evolutionary Biology. Available at: https://doi.org/10.1155/2012/846421.
Michaelson, J.J., Sebat, J., 2012. forestSV: Structural variant discovery through statistical learning. Nature Methods 9, 819–821. Available at: https://doi.org/10.1038/nmeth.2085.
Nacher, J.C., Hayashida, M., Akutsu, T., 2010. The role of internal duplication in the evolution of multi-domain proteins. Biosystems 101, 127–135. Available at: https://doi.org/10.1016/j.biosystems.2010.05.005.
Nicolson, G.L., 2015. Cell membrane fluid-mosaic structure and cancer metastasis. Cancer Research 75, 1169–1176. Available at: https://doi.org/10.1158/0008-5472.CAN-14-3216.
Nicolson, G.L., 2014. The Fluid – Mosaic Model of Membrane Structure: Still relevant to understanding the structure, function and dynamics of biological membranes after more than 40 years. Biochimica et Biophysica Acta (BBA) - Biomembranes 1838, 1451–1466. Available at: https://doi.org/10.1016/j.bbamem.2013.10.019.
Normark, S., Bergstrom, S., Edlund, T., *et al*., 1983. Overlapping Genes. Annual Review of Genetics 17, 499–525. Available at: https://doi.org/10.1146/annurev.ge.17.120183.002435.
Ortiz, J.F., Rokas, A., 2017. CTDGFinder: A novel homology-based algorithm for identifying closely spaced clusters of tandemly duplicated genes. Molecular Biology and Evolution 34, 215–229. Available at: https://doi.org/10.1093/molbev/msw227.
Panchy, N., Lehti-Shiu, M., Shiu, S.-H., 2016. Evolution of gene duplication in plants1[OPEN]. Plant Physiology 171, 2294–2316. Available at: https://doi.org/10.1104/pp.16.00523.
Patthy, L., 2007. A general theory of gene sharing. Nature Genetics 39, ng0607–ng0701. Available at: https://doi.org/10.1038/ng0607-701.
Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling – A review. Gene 238, 103–114.
Pink, R.C., Wicks, K., Caley, D.P., *et al*., 2011. Pseudogenes: Pseudo-functional or key regulators in health and disease? RNA 17, 792–798. Available at: https://doi.org/10.1261/rna.2658311.
Pradhan, S., Kant, C., Verma, S., Bhatia, S., 2017. Genome-wide analysis of the CCCH zinc finger family identifies tissue specific and stress responsive candidates in chickpea (Cicer arietinum L.). PLOS ONE. 12. Available at: https://doi.org/10.1371/journal.pone.0180469.
Qian, W., Zhang, J., 2014. Genomic evidence for adaptation by gene duplication. Genome Research 24, 1356–1362. Available at: https://doi.org/10.1101/gr.172098.114.
Raj Singh, T., 2008. Mitochondrial gene rearrangements: New paradigm in the evolutionary biology and systematics. Bioinformation 3, 95–97.
Reams, A.B., Roth, J.R., 2015. Mechanisms of gene duplication and amplification. Cold Spring Harbor Perspectives in Biology. 7. Available at: https://doi.org/10.1101/cshperspect.a016592.
Robinson-Rechavi, M., Boussau, B., Laudet, V., 2004. Phylogenetic dating and characterization of gene duplications in vertebrates: The cartilaginous fish reference. Molecular Biology and Evolution 21, 580–586. Available at: https://doi.org/10.1093/molbev/msh046.
Rose, M.R., Oakley, T.H., 2007. The new biology: Beyond the modern synthesis. Biology Direct 2, 30. Available at: https://doi.org/10.1186/1745-6150-2-30.
Roy, B., Haupt, L.M., Griffiths, L.R., 2013. Review: Alternative Splicing (AS) of genes as an approach for generating protein complexity. Current Genomics 14, 182–194. Available at: https://doi.org/10.2174/1389202911314030004.
Schacherer, J., Tourrette, Y., Souciet, J.-L., Potier, S., de Montigny, J., 2004. Recovery of a function involving gene duplication by retroposition in saccharomyces cerevisiae. Genome Research 14, 1291–1297. Available at: https://doi.org/10.1101/gr.2363004.
Schmitz, J., Brosius, J., 2011. Exonization of transposed elements: A challenge and opportunity for evolution. Biochimie 93, 1928–1934. Available at: https://doi.org/10.1016/j.biochi.2011.07.014.
Schweizer, E., Hofmann, J., 2004. Microbial Type I Fatty Acid Synthases (FAS): Major players in a network of cellular FAS systems. Microbiology and Molecular Biology Reviews, 68. pp. 501–517. Available at: https://doi.org/10.1128/MMBR.68.3.501-517.2004.
Seehausen, O., Butlin, R.K., Keller, I., *et al*., 2014. Genomics and the origin of species. Nature Reviews Genetics 15, nrg3644. Available at: https://doi.org/10.1038/nrg3644.
Shapiro, B.J., Leducq, J.-B., Mallet, J., 2016. What is speciation? PLOS Genetics 12, e1005860. Available at: https://doi.org/10.1371/journal.pgen.1005860.
Simillion, C., Vandepoele, K., Montagu, M.C.E.V., Zabeau, M., Peer, Y.V. de, 2002. The hidden duplication past of Arabidopsis thaliana. Proceedings of the National Academy of Sciences 99, 13627–13632. Available at: https://doi.org/10.1073/pnas.212522399.
Singer, S.J., Nicolson, G.L., 1972. The fluid mosaic model of the structure of cell membranes. Science 175, 720–731.
Singh, T.R., Pardasani, K.R., 2009. Ambush hypothesis revisited: Evidences for phylogenetic trends. Computational Biology and Chemistry 33, 239–244. Available at: https://doi.org/10.1016/j.compbiolchem.2009.04.002.
Singh, T.R., Tsagkogeorga, G., Delsuc, F., *et al*., 2009. Tunicate mitogenomics and phylogenetics: Peculiarities of the Herdmania momus mitochondrial genome and support for the new chordate phylogeny. BMC Genomics 10, 534. Available at: https://doi.org/10.1186/1471-2164-10-534.
Stephens, S.G., 1951. Possible significance of duplication in evolution. In: Demerec, M. (Ed.), Advances in Genetics. Academic Press, pp. 247–265. Available at: https://doi.org/10.1016/S0065-2660(08)60237-0.
Stone, D.L., Agarwala, R., Schäffer, A.A., *et al*., 1998. Genetic and physical mapping of the McKusick-Kaufman syndrome. Human Molecular Genetics 7, 475–481. Available at: https://doi.org/10.1093/hmg/7.3.475.
Stower, H., 2013. Alternative splicing: Regulating *Alu* element "exonization". Nature Reviews Genetics 14, nrg3428. Available at: https://doi.org/10.1038/nrg3428.
Suzuki, T., Yuasa, H., Imai, K., 1996. Convergent evolution. The gene structure of Sulculus 41 kDa myoglobin is homologous with that of human indoleamine dioxygenase. Biochimica et Biophysica Acta 1308, 41–48.
Toll-Riera, M., Laurie, S., Radó-Trilla, N., Alba, M.M., 2011. Partial Gene Duplication and the Formation of Novel Genes. IntechOpen. Available at: https://doi.org/10.5772/21846.
Tutar, Y., 2012. Pseudogenes [WWW Document]. International Journal of Genomics 2012, 4. Available at:.https://doi.org/10.1155/2012/424526.
Vogel, C., Teichmann, S.A., Pereira-Leal, J., 2005. The relationship between domain duplication and recombination. Journal of Molecular Biology 346, 355–365. Available at: https://doi.org/10.1016/j.jmb.2004.11.050.
Wang, Y., Liu, J., Huang, B., *et al*., 2015. Mechanism of alternative splicing and its regulation (Review). Biomedical Reports 3, 152–158.
Wang, T.-T., Si, F.-L., He, Z.-B., Chen, B., 2018. Genome-wide identification, characterization and classification of ionotropic glutamate receptor genes (iGluRs) in the malaria vector Anopheles sinensis (Diptera: Culicidae). Parasites & Vectors 11 (1), 34. Available at: https://doi.org/10.1186/s13071-017-2610-x.
Wang, X., Wheeler, D., Avery, A., *et al*., 2013. Function and evolution of DNA methylation in nasonia vitripennis. PLOS Genetics 9, e1003872. Available at: https://doi.org/10.1371/journal.pgen.1003872.
White, M.J.D., 1977. Animal Cytology and Evolution. CUP Archive.
Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research 19, 1586–1592. Available at: https://doi.org/10.1101/gr.092981.109.
Yu, P., Ma, D., Xu, M., 2005. Nested genes in the human genome. Genomics 86, 414–422. Available at: https://doi.org/10.1016/j.ygeno.2005.06.008.
Zhou, X., Lin, Z., Ma, H., 2010. Phylogenetic detection of numerous gene duplications shared by animals, fungi and plants. Genome Biology 11, R38. Available at: https://doi.org/10.1186/gb-2010-11-4-r38.