Annu. Rev. Genet. 2004. 38:615–43
doi: 10.1146/annurev.genet.38.072902.092831

First published online as a Review in Advance on July 26, 2004

DUPLICATION AND DIVERGENCE: The Evolution of New Genes and Old Ideas

John S. Taylor[1] and Jeroen Raes[2]
[1]*Department of Biology, University of Victoria, Victoria, British Columbia, V8W 3N5, Canada; email: taylorjs@uvic.ca*
[2]*Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, B-9052 Gent, Belgium; email: jeroen.raes@psb.ugent.be*

Key Words gene, genome, duplication, evolution, functional divergence, subfunctionalization, neofunctionalization, polyploidy

> *We have formerly seen that parts many times repeated are eminently liable to vary in number and structure; consequently it is quite probable that natural selection, during the long-continued course of modification, should have seized on a certain number of the primordially similar elements, many times repeated, and have adapted them to the most diverse purposes.*
>
> Charles Darwin, 1859 (23)

■ **Abstract** Over 35 years ago, Susumu Ohno stated that gene duplication was the single most important factor in evolution (97). He reiterated this point a few years later in proposing that without duplicated genes the creation of metazoans, vertebrates, and mammals from unicellular organisms would have been impossible. Such big leaps in evolution, he argued, required the creation of new gene loci with previously nonexistent functions (98). Bold statements such as these, combined with his proposal that at least one whole-genome duplication event facilitated the evolution of vertebrates, have made Ohno an icon in the literature on genome evolution. However, discussion on the occurrence and consequences of gene and genome duplication events has a much longer, and often neglected, history. Here we review literature dealing with the occurence and consequences of gene duplication, begining in 1911. We document conceptual and technological advances in gene duplication research from this early research in comparative cytology up to recent research on whole genomes, "transcriptomes," and "interactomes."

CONTENTS

EVOLUTIONARY LINKS BETWEEN CHROMOSOME NUMBER, CHROMOSOME MORPHOLOGY, SPECIATION, AND MORPHOLOGICAL DIVERSIFICATION

In 1911, Kuwada concluded that maize (*Zea mays*) was an ancient tetraploid after recognizing two sets of paralogous chromosomes in its karyotype and he proposed that the production of many different varieties ("innumerable races") of *Z. mays* might be related to this chromosome duplication event. At about the same time, Tischler (126) noticed morphological differences that were correlated with variation in chromosome number in several closely related plants. The connection between chromosome and morphological variation, although more obvious today, was novel at that time. Tischler commented, "even recently, a drawing together of these two disciplines would have been considered absurd." In 1918, Calvin Bridges wrote that the main interest in duplications lay in their offering a method for evolutionary increase in lengths of chromosomes with identical genes, which could subsequently mutate separately and diversify their effects (13). Hermann Muller, like Bridges, studied in the *Drosophila* genetics lab of Thomas Morgan. He produced fruit flies with a small fragment of their X chromosome duplicated and inserted into chromosome two. The viability of these mutants led Muller to propose that such duplications might occur in nature and might be a way of increasing gene

number without the typically negative consequences of aneuploidy (the gain or loss of one or more whole chromosomes). He further proposed (92) that the redundent loci produced by the duplication of chromosomes parts could experience divergent mutations and eventually be regarded as nonhomologous genes.

In 1932, Haldane suggested that duplication events might be favorable because they produce genes that could be altered without disadvantage to the organism. He also proposed that organisms with multiple copies of genes would be less prone to harmful mutations (47). This second suggestion was based upon Stadler's (117) demonstration that polyploid species in *Avena* (oats) and *Triticum* (barley) were less susceptible than their diploid congeners to irradiation. The hypothesis that duplicate genes buffer organisms from harmful mutations continues to be tested (45, 134). Gu et al. (45) found that yeast with mutations in single-copy genes scored lower in growth tests than did strains with mutations in one copy of a duplicate gene. This apparent ability of duplicated genes to cover for one another was negatively correlated with their ages.

In 1938, Serebrovsky proposed that selection could be relaxed in genes that occur in duplicate (112). He also considered the possibility that both copies might be modified. Having characterized the roles of *achaete* and *scute*, closely linked genes on the *Drosophila* X chromosome that influence bristle morphology, Serebrovsky concluded that a single gene might influence multiple aspects of the phenotype and that after a duplication event, these multiple functions might be distributed between the duplicates: "This principle of loss of duplicate functions by one of the homologues [paralogues] in the process of genic evolution is considered by us as an important (though not the single) explication of a great number of phenomena discovered by genetics." "It should result in a specialization of genes, when each then fulfills only one function which is strictly limited and important for the life of the organism." Here, Serebrovsky clearly outlines what is now thought of as gene sharing or subfunctionalization, the only difference being the explicit connection between division of labor and mutations in gene regulatory elements that characterizes most current ideas concerning subfunctionalization (30, 139).

During the 1930s, several authors also expanded upon previous observations linking polyploidy to phenotypic variation. Müntzing (93) observed that morphology was strongly correlated with the chromosome number; cell size, and plant size often increased with chromosome number. Müntzing also reviewed ecological differences between races with different karyotypes. He cited Jenkin (61) who concluded that the consequence of polyploidy in *Festuca* (a genus of about 100 grass species), as in other genera, appears to be the production of types that are capable of extending their ecological range. Tischler (127) reported a correlation between polyploidy and latitude. Fifty five percent of species in Iceland were found to be polyploid compared with only 31% of the species in Sicily. Other studies also turned up data consistent with this observation of improved cold-hardiness in polyploids. Especially interesting is the conclusion by Nishiyama (96) that cold-hardiness in tetraploids is a consequence of chromosome number per se and is not due to the duplication of specific genes. Müntzing also cited multiple studies that

showed tetraploid plants grow more slowly than their diploid counterparts and that tetraploid tomatoes and apples contain more vitamin C than diploids.

In that same period, debate ranged over the relative contributions of autotetraploidy (intraspecific genome duplication) and allotetraploidy (genome duplication associated with a hybridization event) to the creation of new species (93). Many authors recognized that allotetraploidy provided a single species with greater genetic diversity than its diploid progenitors, but only a few appear to have considered autotetraploidy as a significant evolutionary event. Among those, Jørgensen (65) suggested that species diversity might be derived from the differentiation of duplicate genes in autotetraploids. Müntzing (93) considered the widespread occurrence of chromosome races, with the different races being members of a "typical (chromosome) series" and possessing distinct morphologies, as evidence that genome duplication was common and that it leads to speciation.

In 1933, Blakeslee studied the correlation between variation in karyotype and morphology in jimson weeds (*Datura stramonium*). Diverse aneuploid races were observed in nature and created in the lab. By comparing morphology among lab strains, Blakeslee associated morphological change with the duplication of specific chromosome parts. For example, duplication of a region of the largest chromosome caused the plant capsule to be small and the leaves to be narrow, whereas duplication of a different part of the same chromosome leads to a large capsule and relatively broad leaves (9). Like Muller, Blakeslee made pure-breeding lab strains with duplicated chromsome segments. He considered these to be artificial new species. "Whether nature has used such methods, we do not yet know," remarked Blakeslee, "but it should be remembered that often, when man has devised a method which he has thought unique, nature has been found to have had the priority in the use of the same method."

Bridges (14) also linked morphological variation and gene duplication and proposed that duplication could lead to speciation. The observation that certain sections of normal chromosomes have been built up in blocks through duplication goes far, he wrote, toward explaining species initiation. "For the duplication of sections of genes is known in *Drosophila* to cause many slight, poorly-defined differences in all parts of the duplicant type." Soon thereafter, Bridges combined his indirect observations of fly genes in polytene chromosomes with data from crossing experiments to conclude that the Bar-eye and Bar-double phenotypes (both have reduced eyes) were a consequence of rare small-scale tandem duplication events (14). These phenotypes had, in fact, been attributed to duplication events much earlier by Sturtevant (122), but at that time, resolution was poor because giant (polytene) fly chromosomes had not been discovered.

The first explicit links were made between organismal complexity and gene duplication in the 1940s. Goldschmidt (41), who focused on the role of chromosome repatterning in macroevolution, did not believe that human and amoeba were connected by mutations of the same genes. In 1944, Gulick (46) argued, "The whole history of many-celled organisms must undoubtedly have called for frequent increases in gene count as the organism advanced toward great complexity."

Metz (88) reiterated Goldschmidt's scepticism, stating that evolution cannot be explained upon the basis of loss or simple alteration of materials already present in the germplasm. "New elements must be added. Otherwise we would have to assume that the primordial amoeba was endowed with all the germinal components now present throughout the wide range of its descendents, from protozoa to man." Using a style that Ohno appears to have adopted, Metz (88) argued that the duplication of chromosome parts, together with its consequences, has probably been one of the most important factors in evolution, if not the most important. By contrast, Huxley (58) argued that, in spite of the frequency of these larger types of mutation, gene mutation (allele-level mutations), together with the "pseudomutation" due to position effects (rearrangements), was the most important source of evolutionary change.

Twenty years before Ohno, but clearly expanding upon ideas developed 20 years earlier, Stephens (119) questioned whether evolution took place by the slow accumulation of allelic mutations. He recognized that mutations were likely to impair original gene function, and he proposed that the only way of achieving "evolutionary progress" (the evolution of new species, genera, and higher categories) would be by increasing the number of genetic loci, either by the synthesis of new loci from nongenic material or by the duplication and subsequent differentiation of existing loci via genome duplication or unequal recombination.

Also in 1951, Lewis concluded that numerous traits thought to be based upon allelic variation were, in fact, under the control of so-called pseudoalleles, i.e., linked duplicates. Interestingly, Bailey et al. (8) recently concluded that many single nucleotide polymorphisms (SNPs) in human genomes are variants at paralogous loci, that is, pseudoalleles. Lewis (75) proposed a model for the creation of multistep biochemical reactions by gene duplication. This model included a form of subfunctionalization as an intermediate step during the evolution of complex biochemical processes.

This first section of our review shows that between 1911 and the 1950s, cytology and then cytogenetics, with enormous contributions from *Drosophila* research, produced substantial evidence for gene (and whole-genome) duplication. This brief review of the early literature shows that many hypotheses concerning the evolutionary significance of gene duplication (e.g., in speciation and morphological complexity) have a long history, as have ideas about the mechanisms by which gene duplication might contribute to these phenomena (e.g., subfunctionalization, neofunctionalization).

GENE DUPLICATION AND DNA CONTENT

Technological advances, including measurements of total DNA content, isozyme electrophoresis, and DNA sequencing, have supplemented but not supplanted microscopy-based research into gene duplication. Mirsky & Ris (90) were among the first to report that DNA content for haploid cells (sperm and eggs) was constant

within species but variable among species. By counting either sperm or erythrocytes and then dividing DNA content by cell number, they tested the hypothesis that DNA content was higher in "the more highly developed animals." Among invertebrates, the most primitive animals, sponges and coelenterates, have lower DNA content values than echinoderms, crustaceans, and molluscs. However, Mirsky & Ris found that in vertebrates, DNA content does not increase in the more highly developed animals. Furthermore, these authors were among the first to suggest that genome size and gene content might not be correlated. They remarked that some salamanders contain about 70 times more DNA than chickens, but that it was unlikely that they contain 70 times more genes.

Britten & Davidson (15) summarized DNA content from widely diverse species (e.g., viruses, bacteria, plants, vertebrates), with a plot showing that a great increase in DNA is "a necessary concomitant to increased complexity of organization." They also knew that DNA content might not be correlated with gene number and argued that the likelihood of utilization of new DNA for regulation is far greater than the likelihood of invention of a new and useful amino acid sequence.

Interestingly, Ohno also based his theories on vertebrate evolution on differences in haploid DNA content: He noted a twofold to threefold increase in genome size between *Ciona intestinalis* and placental mammals and concluded that this difference (together with the evolution of the mammalian body plan) was due to extensive gene duplication, possibly polyploidization (98).

GENE DUPLICATION AND ISOZYME ELECTROPHORESIS

In the late 1960s, variation at enzyme-encoding loci could be detected using starch gel electrophoresis. Although only a small fraction of the total DNA sequence variation among genes can be resolved by studying the isozymes they encode, the amount of variation uncovered was much higher than expected (49, 76).

Duplicated isozyme loci produced more bands (regions of enzymatic activity) than single-copy genes, and many isozyme electrophoresis studies uncovered gene duplicates in polyploids and also in species where no cytological data had predicted their occurrence. In several instances, isozyme studies were designed specifically to test gene and genome duplication hypotheses. For example, the discovery that some duplicated isozymes show parallel linkage in maize (121) supported Kuwada's (73) hypothesis that *Z. mays* evolved from a tetraploid ancestor (see also 37, 107). Avise & Kitto (7) uncovered duplicated PGI loci in a diversity of ray-finned fish. They concluded that this locus was duplicated after the divergence of gar and bowfin from the ancestor of teleosts. Hoegg et al. (50) recently suggested that a whole-genome duplication event maps to this point in the ray-finned fish phylogeny.

Data from isozyme studies not only uncovered new examples of gene duplication, it also shed light on the consequences of gene duplication. Avise & Kitto (7), for example, discovered that duplicated PGI genes were expressed in different tissues. By surveying representatives from a diversity of fish species, they

demonstrated that this expression-level divergence occurred shortly after the genes were duplicated.

Ferris & Whitt (29) compared expression patterns of duplicated isozymes among species in the tetraploid fish family Catostomidae. An average of 8 duplicated enzymes was studied in 15 species and for each species in 10 different tissues. They concluded that the rates of changes in regulatory genes (which might be interpreted as including regulatory elements) and structural genes were uncoupled; values for expression divergence were not correlated with subunit molecular weights or with heterozygosity (which was a proxy for sequence divergence of the coding region). By mapping isozyme expression domains onto a phylogeny of catastomid fishes, these authors (29) concluded that differential gene expression between some duplicates evolved soon after the 50 million-year-old genome duplication event whereas other differences in expression domain probably arose recently. Divergent expression was usually unidirectional in Ferris & Whitt's study. That is, one of the two duplicates typically had stronger staining across all tissues surveyed. Likewise, Wagner (136) recently found that the divergence of duplicated genes is often asymmetric in yeast, i.e., one gene loses more functions than the other. Ferris &Whitt (29) also found variation among tissues with respect to the degree of divergent expression: For eight pairs of duplicates, expression patterns were most similar in the brain and least similar in the liver.

In a survey of data from 100 human isozyme loci, Hopkinson et al. (53) observed that 20 loci occurred in duplicate. For these 20 loci, the proteins encoded by paralogous genes were very similar with respect to subunit size and subunit number. In contrast to this structural similarity, there were several instances of divergent expression among these sets of duplicated human enzymes. Interestingly, Wagner (134), using sequence and microarray data for 124 duplicate pairs of yeast genes, has recently shown almost no correlation between divergence in expression pattern and evolutionary distance based upon protein-coding sequences.

In 1972, Koch proposed a multistep model of enzyme evolution that involved gene duplication (68). According to this model, the evolutionary improvement of enzymatic function by one-step-at-a-time substitutions reaches a plateau and then only very rare multiple simultaneous mutations or locus duplication can improve enzyme function. Koch's model posits that if evolution takes the duplication route, at some point in the future the advantage of having two genes for one enzyme will diminish. Then, one copy is free to experience nonselective or, as Koch wrote, "non-Darwinian" mutational changes. Koch's model involves a second round of growth limitation (selection) but it included "reversion" as a potential solution. Reversion occurs when the degenerated duplicated gene is revived, i.e., codes for a better enzyme, which will take over the population by the selective virtue of its superior maximum activity. Thus, this model involves a race between degeneration and neofunctionalization, but only after selection established both paralogs in the population.

Jensen (63) also proposed a model for enzymatic evolution via gene duplication. Jensen's model involved an ancestral enzyme with very broad specificity being duplicated and the descendants of this molecule specializing on a subset of ancestral

functions. This model is similar to that proposed many years earlier by Serebrovsky for the evolution of *achaete* and *scute*. It is also a prelude to the gene-sharing model of Wistow & Piatigorsky (139) and subfunctionalization (30).

Reciprocal silencing is another idea derived from isozyme studies (138). It occurs when different allopatric populations lose different copies of a duplicated locus. Lynch & Force's updated version of this model (Divergent Resolution) considered not only the loss of different gene duplicates but also expression divergence between duplicates, and the implications of such events occurring following whole-genome duplication (82). Reciprocal silencing, or divergent resolution, creates a genetic environment in which alleles promoting within-population mating would be favored. Thus, it is a model in which gene duplication leads to speciation as a result of population-level variation in postduplication mutations. Several studies have shown variation among populations in the retention of duplicated loci (reviewed in 124), but none has uncovered the pattern of gene loss predicted by the model.

DNA SEQUENCES AND THE PREVALENCE AND CONSEQUENCES OF GENE DUPLICATION

Although protein electrophoresis led to the detection of gene and whole-genome duplication events in species where there was no cytological evidence for duplication, isozyme research has some drawbacks for the study of gene duplication. Most allozyme surveys are restricted to a set of approximately 30 enzyme loci. Also, duplicated isozyme-coding genes can be identified only if both copies are expressed in the tissue that is surveyed and if they produce proteins with different eletrophoretic mobilities. With DNA sequencing, duplicate genes can often be uncovered by using degenerate-primer PCR or DNA sequence-based probes to survey genomic DNA (e.g., BAC clones). DNA sequencing has also spawned a diversity of additional technologies that have contributed to our understanding of gene duplication including in situ hybridization, EST- and cDNA-based microarrays, and promoter-reporter gene transgenics, which all provide much more information than isozyme electrophoresis about the expression patterns of duplicated genes.

Space limitations prevent a review of all of the DNA sequence-based studies that have turned up evidence for gene duplication and that have described its consequences in terms of sequence divergence and gene expression. Also, new contributions to this field are published on a monthly basis, so no such review could be complete. Here we limit our disussion to the contributions of nine gene families. We also briefly review two papers that provided evidence for genome-wide duplication events by compiling scattered sequence data from individual gene families.

Crystallin Genes and "Gene Sharing"

Crystallins are enzymatic proteins that, in addition to their enzymatic roles, contribute up to 60% of the protein in the lenses of vertebrate and squid eyes (139).

There are α-, β-, γ-, and ε-crystallins. Alpha-crystallin belongs to a superfamily of heatshock proteins, β- and γ-crystallins belong to a family of calcium binding proteins, and ε-crystallin is a functional lactate dehydrogenase. Wistow & Piatigorsky (139) proposed that these crystallins might have been recruited as lens proteins because of their especially stable structure. Their model for α-, β-, and γ-crystallin evolution involved a single enzymatic protein that gained a structural role as a result of the acquisition of new gene promoter elements, followed by duplication, divergence, and specialization. For ε-crystallin, one gene continues to carry out dual roles, apparently awaiting duplication and divergence. Because crystalline enzymes appear to have evolved their structural role prior to duplication, this phenomenon was called gene sharing. Further evidence from crystallin genes for the gene-sharing model of gene evolution was reported in 1991 (103). Chickens and ducks have two δ-crystallin genes (δ1 and δ2), and a lens-preferred enhancer is present in both genes in both species. δ2-crystallin codes for argininosuccinate lyase (ASL) and δ1 appears to play a role only in lens structure. Both δ-crystallin genes are expressed in the lens of ducks, thus ASL activity is high in duck lenses. However, in chicken lenses, 95% of the δ-crystallin is of δ1 type, that is, the subdivision of roles has proceeded further in chickens.

Hughes (56) proposed a similar model from his research on *Xenopus*. Hughes proposed that gene duplication leading to the production of two genes encoding functionally distinct proteins is usually preceded by a period of gene sharing, that is, a period in which a single generalist protein performs two distinct functions. Once gene duplication occurs, he argued, it becomes possible for the products of the two duplicate genes to specialize so that each performs only one of the functions performed by the ancestral gene. Following the gene-sharing model, specialization can be achieved by a change in the regulation of expression of one or both daughter genes. Thus, in a multicellular organism, each daughter gene might be expressed in a more restricted set of tissues than was the ancestral gene. Hughes added the hypothesis that natural selection may act rapidly to favor certain amino acid replacements that better suit each daughter gene to its specific function.

Zebrafish *engrailed* Genes and the Subfunctionalization Model of Post-Duplication Gene Evolution

In 1999, Force and coworkers introduced the duplication-degeneration-complementation (DDC) model, in which complementary degenerative mutations in regulatory elements controlling the expression of duplicated genes allowed the partitioning of ancestral gene functions. They called this process subfunctionalization. Zebrafish engrailed genes, *eng1a* and *eng1b*, which were formed very early during the evolution of teleost fishes, were one of the examples of subfunctionalization more fully described by Force et al. (30). In situ hybridzation using *eng1a* and *eng1b* probes and 28.5-h zebrafish embryos showed *eng1b* expression in a specific set of hindbrain and spinal neurons, whereas *eng1a* was expressed in the pectoral appendage bud. *Eng1* in mouse and chicken is expressed in both domains, as would

be predicted if it reflects the ancestral single-copy fish sequence and if *enga1* and *engb1* had followed the subfunctionalization model of gene evolution.

Also in 1999, Stoltzfus described a conceptually similar model (120). The retention of duplicates was considered by Stoltzfus to be an example of constructive neutral evolution. As was the case with previous models, Stoltzfus suggested that duplication leads to redundancy, he called it "excess capacity," and that mutations in one copy that reduced function would not be opposed by purifying selection (i.e., such mutations would be neutral, but would prevent the subsequent loss of the second copy). Among those mutations that reduced function, some could lead to novel functions, the reason why Stoltzfus described the process as "constructive" neutral evolution.

Numerous other studies describing expressional differences between duplicated genes have been published, ranging from detailed studies of single gene duplicates (e.g., 6, 54, 80) to large-scale analyses of expression of duplicated genes in polyploids or by mining available expression data (1, 33, 43, 66, 136). Adams et al. (1), for example, observed several cases of differential expression patterns when analyzing organ-specific expression patterns in polyploid cotton. Whether these observations are due to the complementary loss of regulatory elements remains to be determined.

Subfunctionalization and *Drosophila* Bristles

The molecular characterization of the achaete-scute region of the *Drosophila* X chromosome was reviewed by Modolell and Campuzano (91). The achaete-scute complex, initially thought to include two genes (see first section), involves four related coding sequences (*ac*, *sc*, *l'sc*, and *ase*), but only *ac* and *sc* are indispensable for bristle development in *Drosophila*. The divergence of duplicated enhancers in these two genes appears to be responsible for the highly localized accumulation of *ac*- and *sc*-encoded proteins. This conclusion would not have surprised Serebrovsky. From his observation that these genes were closely linked and very similar in function, yet controlling the development of different sets of bristles, he postulated that they arose by duplication from an ancestral gene that controlled the development of all bristles. Serebrovsky explicitly proposed that mutations in one of the two genes would knock out its ability to control a particular bristle, while this loss was compensated by the other, a process that would lead to two genes having partitioned the ancestral function into the control of two specific bristle sets.

Gene Evolution, Genome Evolution, and *Hox* Genes

Hox genes, which encode DNA-binding proteins that regulate the expression of a diversity of genes, have contributed considerably to our knowledge of the prevalence and consequences of gene duplication. They typically occur in one or more clusters of up to 13 genes (38, 86). Phylogenetic analyses have been used to reconstruct the evolution of these clusters. Such analyses suggest that tandem duplication

of a protoHox gene produced a four-gene cluster and that this entire cluster was duplicated, producing a four-gene Hox cluster and, on a different chromosome, a four-gene ParaHox cluster (16). Following the nomenclature of Kourakis & Martindale (70), the four-gene Hox cluster included *anterior Hox*, *Hox3*, *central Hox*, and *posterior Hox*. *Hox1* and *Hox2* are derived from *anterior Hox* and the duplication of *central Hox* has led to the production of *Hox4* through *Hox8*. The remaining Hox genes are descendents of *posterior Hox*. Ferrier et al. (28) reported a fourteenth Hox gene in amphioxus, and there is evidence for 14-gene Hox clusters in sharks and in the coelacanth (104).

The observation that *Amphioxus*, (a cephalochordate) possesses one Hox gene cluster (34), whereas sarcopterygians (lobe-finned fish including coelacanths and lungfishes, amphibians, reptiles, birds, and mammals) have four, is consistent with the hypothesis that two whole-genome duplication events occurred early in vertebrate evolution (51, 52, 109, 116; see below).

Inspired by these Hox cluster observations, Pébusque et al. (101) combined phylogenetics and gene map data to identify additional large paralogous regions of the human genome. The results supported the hypothesis that large-scale gene duplication occurred before the evolution of bony vertebrates but after the Protostomia/Deuterostoma split. Pébusque et al. (101) did not explicitly link gene duplication to the evolution of vertebrates, but did remark that their study was part of the search for events that have molded animal evolution.

Irvine et al. (59) and Force et al. (31) sequenced Hox genes from the sea lamprey (*Petromyzon marinus*). By counting Hox clusters in this species, they hoped to provide a clearer picture of when *Hox* genes were duplicated and thereby determine whether Hox cluster duplication was correlated with the evolution of vertebrates. Irvine et al. (59) and Force et al. (31) concluded that lamprey have four Hox gene clusters, which suggests that cluster duplication long preceded increases in axial (body plan) complexity.

Amores and coauthors (3) uncovered seven Hox clusters in zebrafish (*Danio rerio*). Extrapolating from Hox clusters to whole genomes, they proposed that a fish-specific genome duplication might have been responsible for this gene cluster amplification (see also 89). More than four Hox clusters have also been described for medaka (*Oryzias latipes*) (94), an African cichlid fish *Oreochromis niloticus* (84), the spotted pufferfish *Takifugu rubripes* (5), and *Spheroides nephelus*, the southern pufferfish (4). Thus, Hox clusters, and perhaps the whole genome, duplicated in ray-finned fishes before the divergence of zebrafish, medaka, and pufferfishes.

Post-duplication Hox gene loss in fishes has been substantial. Zebrafish have only nine more Hox genes than human and mouse have (48 versus 39). Interestingly, and in stark contrast to the pattern in mammals where human and mouse possess an identical Hox gene complement, the pattern of gene loss differs among fish species. Medaka, the spotted pufferfish, and the southern pufferfish have two Hoxa, Hoxb, and Hoxd clusters, whereas the zebrafish has two Hoxa, Hoxb, and Hoxc clusters. Furthermore, Hox gene complement differs among pufferfish species; the

hoxb7a gene is absent from the spotted puffer, but present in the southern puffer. Both copies of *hoxb7* must have been retained from the time of duplication (circa 350 Mya) to the divergence of these two pufferfish species (circa 5–35 Mya). Therefore, silencing of duplicates appears possible long after their duplication.

Hox gene expression studies have uncovered evidence for neofunctionalization, "function shuffling," and subfunctionalization. McClintock et al. (85) studied expression of Hox1 genes, the so-called paralogy group (PG) 1. Zebrafish have four PG1 genes, *hoxa1a*, *hoxb1a*, *hoxb1b*, and *hoxc1a*, whereas mice have three PG1 genes, *Hoxa1*, *Hoxb1*, and *Hoxd1*. Using in situ hybridization, McClintock et al. (85) discovered that zebrafish *hoxa1a* expression is very different from mouse *Hoxa1* expression: In zebrafish, *hoxa1a* expression in small bilateral cell clusters in the mid- and hindbrain appears to be an example of neofunctionalization. Zebrafish *hoxb1b* appears to perform the same role in zebrafish development as *Hoxa1* performs in mouse. This pattern of nonorthologous genes fulfilling equivalent roles was called function shuffling (85).

Duplicated Hoxd4 genes in *Spheroides nephelus* (the southern pufferfish) have overlapping but not identical expression domains (4). The anterior limit of *hoxd4a* expression is rhombomere seven (r7), as has been observed in zebrafish and mouse. However, *hoxd4b* expression stops at r8. *Spheroides hoxd4b* is expressed more strongly than its paralog in the hindbrain and neural crest (see figure 5 in 4). These *hoxd4a* and *hoxd4b* data appear to be an example of subfunctionalization; however, the mutations responsible for these postduplication changes in gene expression have yet to be characterized.

Amores et al. (4) also characterized the expression domains of duplicated Hoxa2 genes and uncovered what might be an especially interesting example of neofunctionalization. In pufferfish and in zebrafish, *hoxa2b* is expressed in hindbrain rhombomeres r2–r5. The same pattern has been observed for *HoxA2* in mouse. Pufferfish have a *hoxa2a* gene, which is expressed in r1. Amores et al. (4) speculated that *hoxa2a* expression in pufferfish r1 might play a role in the pufferfish-specific invention of the buccal pump, which puffs the stomach with water.

Subfunctionalization at the Protein Coding Level

Dermitzakis & Clark (24) introduced an approach for detecting what might be considered coding-level subfunctionalization in duplicated genes, using a method based upon local evolutionary rate differences between paralogs. Protein-coding subfunctionalization also appears to have occurred in fish Microphthalamia-associated transcription factor (*mitf*) and Synapsin (*syn*) genes. The duplicates each code for proteins that correspond to isoforms generated by alternative splicing in their human single-copy orthologs (2, 79, 143). In the *Takifugu synA* and s*ynB* genes, Yu et al. (143) showed that divergence was the result of complementary degenerate mutations disabling alternative splicing in each duplicate gene and allowing only one of the transcripts, orthologous to one of the two respective mammalian isoforms, to be expressed. Duplicated DGCR6 genes appear to be

another example of coding-level subfunctionalization. Humans have two DGCR6 genes. Comparative FISH (fluorescence in situ hybridization) and genomic sequence analyses suggested that this gene was duplicated in the primate lineage approximately 35 Mya. Although the function of these genes is not known, Edelmann et al. (26) reported EST-based evidence to indicate that the retention of both copies in human might be a consequence of asymmetric mutations that decrease the efficacy of each gene and lead to selection for genomes to retain both.

RNAse: Duplication and Divergence by Positive Selection

Another example of neofunctionalization is the duplication of the ribonuclease (*RNAse1*) gene in a leaf-eating colobine monkey. While most monkeys eat fruit and insects, colobine monkeys eat primary leaves. Leaves are fermented in their foregut by symbiotic bacteria that, when digested, serve as a source of nutrition for the monkeys. Colobine monkeys have two RNAse1 genes. *RNAse1a* digests double-stranded RNA, whereas, *RNAse1b* has undergone several radical amino acid substitutions that appear to allow it to digest bacterial RNA in the acidic foregut. Thus, this example of duplication and divergence represents an adaptation by colobine monkeys to a new nutritional niche (144). The analysis of synonymous versus nonsynonymous mutations in the two genes showed that the driving force behind the acquisition of this new function was positive Darwinian selection, which occurs when the number of nonsynonymous substitutions per nonsynonymous site exceeds that of synonymous substitutions per synonymous site.

Opsins and Evolutionary Innovation by Gene Duplication

The number of examples of the evolution of new, potentially adaptive functions in duplicated genes is growing but still is small. Another striking example is the evolution of trichromatic color vision in primates. Old World primates have trichromatic vision due to the presence of three different visual pigments: retinal bound to one of three different opsin proteins. These pigments have different spectral properties, depending on whether the protein moiety is encoded by the short-wave (SW; autosomal), the middle-wave (MW; X-linked), or long-wave (LW; X-linked) opsin gene. The MW and LW proteins have arisen through a recent gene duplication of an ancestral MW/LW gene. New World monkeys, with only an SW and a MW/LW gene, have dichromatic vision (25). Monkeys possess up to three allelic forms of the MW/LW gene (each with specific spectral properties), which allows some heterozygous females to develop trichromatic vision from the expression of different alleles in different cone photoreceptors (i.e., due to a mosaic of cell-specific X inactivation in the retina). Howler monkeys, a group of New World primates, independently reinvented trichromatic vision through a recent gene duplication of the MW/LW gene (25, 57, 62), unambiguously linking duplication of these genes to this evolutionary innovation.

Guppies (*Poecilia reticulata*) and closely related species in the subgenus *Lebistes* have at least 25 different LW-sensitive (orange and red) opsin alleles (F. Breden,

A. Lindholm, D. Weigel & M. Wade, unpublished information). Eight alleles were sequenced in a single individual of *Poecilia parae*, suggesting that there are at least four loci in this species. Males of *P. reticulata* and *P. parae* are remarkably variable in coloration (78), and this variation plays an important role in sexual selection. The evolution of sexually selected color variation, combined with preliminary evidence for opsin gene family expansion, suggests that in this taxon, variation in color genes and in color perception genes (opsins) might be correlated in a novel form of run-away sexual selection.

Antifreeze Glycoproteins and Post-Duplication Remodeling

Duplicated RNAse1 and opsin genes have evolved new functions as a consequence of formerly forbidden amino acid substitutions. By contrast, the origin of the antifreeze glycoprotein (AFGP) of Antarctic notothenioid fishes is an example of extensive post-duplication protein remodeling. In this case, a trypsinogen-like protease was duplicated. In one copy, a small Thr-Ala-Ala-encoding region expanded through iterative (microsatellite-like) internal duplications. The expanded region codes for the AGFP polyprotein, which is posttranslationally cleaved to form the mature AGFPs that bind to growing ice crystals and prevent the fish from freezing. Later, the obsolete exons coding for protease-specific sequences were lost, giving rise to the AGFP gene in its current form (17, 18).

Olfactory Receptors

Olfactory receptor (OR) gene duplication, their genomic organization, and hypotheses concerning the regulation of OR expression have been recently reviewed (71). Briefly, the proteins coded by these genes are expressed in sensory neurons of the vertebrate nose. There are approximately 100 OR genes in zebrafish and about 1000 in mice and human, but in contrast to both zebrafish and mice, a large proportion of human OR genes are pseudogenes. Each neuron in the olfactory epithelium expresses only a single allele of a single OR gene locus; thus the sensitivity of a given olfactory neuron is limited by the range of ordorants to which a particular OR can bind (71).

Similar to Hox genes, ORs occur in clusters. In zebrafish there are two clusters and the most closely related ORs are adjacent to one another, indicating that tandem duplication and/or uneven crossing-over is the major mode by which this family has expanded in this species. In humans and mice, OR clusters are distributed among many chromosomes and members of OR subfamilies are dispersed among clusters and among chromosomes. Thus tandem and duplicative transposition events (or postduplication gene rearrangements) appear to have played roles in the expansion of this family in mammals (71).

Gilad et al. (40), in a recent study of human and chimpanzee OR genes, found a higher rate of gene loss (pseudogene formation) in humans than in chimps; 50% versus 30%. A reduced need for chemoreception in humans was offered as one explanation, but the observation that intact OR genes appear to have experienced

positive selection in humans immediately following the divergence of these two species indicated that the story was not so simple. Gilad et al. (40) proposed that the human-specific habit of cooking food might explain both the loss of duplicated OR genes and the occurrence of positive selection in a subset of these genes. Whatever the explanation, this study, like the isozyme study in catastomids (29) and fish *Hox* genes (4), shows that the modification and loss of duplicated genes continue long after the events that produced the duplicates.

Large-Scale Duplication Events: Tetralogs, Pro-Orthologs, Semiorthologs and Co-Orthologs

Jürg Spring pointed out that in addition to the Hox gene clusters, at least two other gene families occur once in invertebrates and in four unlinked clusters in vertebrates (the *syndecan* and *myc* gene families). To determine how many other genes occurred in a ratio of 1:4 in fly (*Drosophila*) and human, Spring (116) surveyed FLYBASE and the human genome database, then far from complete. He uncovered 53 sets of orthologous genes with a single fly gene and 2, 3, 4, 5, or 6 vertebrate paralogs. Spring (116) referred to the vertebrate genes as tetralogs. The terms co-ortholog (36) and semiortholog (113) have also been used to describe the relationships among duplicated genes and their single-copy orthologs, or pro-orthologs (113). Genome quadruplication by two rounds of hybridization (allotetraploidy) was put forward by Spring as an explanation for the data. This hypothesis is now best known as the 2R hypothesis.

Taylor et al. (123) used Spring's paper, as not only a methodological model but also a source of human gene query sequences in their study of gene duplication in zebrafish. They used BLASTp to search the NCBI database for zebrafish genes similar to the human genes in Spring's list. Large drops in BLAST e-values were used to identify sets of candidate orthologs, and then phylogenies were reconstructed from these sets. A large number of human genes occurred twice in zebrafish. The ages of the duplicates, estimated from substitutions at third-codon positions, and the discovery that many duplicates occurred on the same pairs of apparently paralogous chromosomes led Taylor et al. (123) to conclude that they were formed during a fish-specific whole-genome duplication event (see also 3, 125). In a follow-up study, the same authors remarked that transcription factors appeared to have been preferentially retained in duplicate, but also recognized that the zebrafish sequences available at the time might not have been an unbiased representation of the genome (128).

EVIDENCE FROM WHOLE-GENOME SEQUENCES

Whole-genome sequences make it theoretically possible to characterize the "paranome" (32), that is, the entire set of duplicated genes in a species. An all-against-all BLAST (or FASTA) search can be used to produce a list of homologous genes

(100). Orthologs and paralogs can then be identified using phylogenetic analyses and gene map (or location) data. Together, these methods reveal the extent to which gene duplication contributes to genome structure (see 114, 131, 141 for reviews). By estimating the ages of a large set of duplicates, it is also possible to estimate gene birth and death rate. Gene function databases, such as the Gene Ontology project or GO (http://www.geneontology.org/GO.doc.html), can be used to further characterize retained duplicates. Now that whole-genome sequences are available from sets of closely related species (two flies, two nematode worms, three mammals, two fishes), comparative genomics can be used to more precisely characterize gene family expansion, however, some methodological problems remain. For example, genome annotation is imprecise (e.g., paralogs can be confused for alleles). Also, BLAST cannot be relied upon to recover all members of a given gene family, and phylogenetic analyses are difficult to automate. However, success in finding gene duplicates in species with sequenced and assembled genomes is not limited by degenerate PCR primer design, oligonucleotide probe design, or incomplete genomic libraries.

Whole-genome sequencing has also allowed the scaling-up of various functional genomics technologies. For example, genome-wide microarrays, and/or large-scale studies of promoter plus reporter-gene transgenics allow gene expression to be studied at a whole-genome scale. Furthermore, large-scale cDNA cloning into expression vectors, combined with yeast two-hybrid experiments, facilitates the characterization of in vitro protein interactions (77, 132). Thus, for a few species it is now possible to identify all duplicated genes and to characterize their expression domains and the proteins they interact with in order to investigate the functional consequences of gene duplication.

How Many Duplicates? The Contribution of Duplication to Genome Structure

Genome sequencing projects show that large-scale gene duplication and complete genome duplication events have contributed to gene family expansion and to genome evolution in a great diversity of species. In *Mycoplasma pneumoniae*, for example, more than 28% of the genome appears to have been produced by lineage-specific duplication events involving about four genes at a time (64). In *Mycobacterium tuberculosis*, more than 33% of the genome is comprised of recently duplicated genes, but in this species some large clusters of between 20 and 90 genes are also involved (64). Gevers et al. (39) analyzed 106 bacterial genomes in a study designed to determine the extent to which gene duplication contributes to genome structure and to expose strain-specific gene family expansions. Paralogous genes were defined by within-genome all-against-all BLAST surveys. From 7% to 41% of the genes in the genomes surveyed had intragenomic BLAST hits, and the size of a species' or strain's genome was strongly correlated with the number of paralogous genes it contained.

In the parasite *Plasmodium falciparum*, the causal agent of human malaria, duplication events have produced paralogous sets or units of ribosomal RNA genes (35). S-type rRNA genes are expressed when the parasite is in the mosquito vector, and A-type rRNA genes are expressed when the parasite occurs in the human host (137). Thus, one consequence of gene duplication for *P. falciparum* is that it has different ribosomes for different environments.

The analysis of whole-genome sequences has turned up evidence for ancient genome duplication in yeast (*Saccharomyces cerevisiae*) (67, 110, 140, 142), *Arabidopsis thaliania* (10–12, 27, 106, 115, 133), and fish (125, 129), aneuploidy in rice (*Oryza sativa* spp. *japonica*) (130), and two large-scale duplication events early in the vertebrate lineage (43, 74, 87, 129 and references therein).

Analyses of the human genome show that recent gene duplication events have also played a large role in the shaping of our genome. During the sequencing of the human genome, each nucleotide was sequenced approximately five times (Celera Genomics database). Thus, a fragment of DNA that occurs once in human, when used as a BLAST query, would be expected to find, on average, five identical matches in the whole-genome shotgun reads. Recently duplicated sequences would be expected to yield an increase in the number of hits, but with a small decrease in average sequence identity among these hits. Bailey et al. (8) surveyed raw data from the human genome sequencing project (27.3 million reads) using 32,610 clones as queries. The cut-off for duplicates was set at >94% sequence identity over 5000 base pairs. This study identified 8595 duplicated regions and concluded that 130.5 megabases of the human genome can be considered to have been recently duplicated. This result has important implications for genome assembly and for human evolution. A very large number of positions that have been considered to be variable, i.e., sites of single nulceotide polymorphisms (SNPs), in fact represent different (paralogous) positions/loci. Also, recently duplicated genes are sites of nonhomologous recombination, which can lead to microdeletion, microduplication, and inversion. These events can, in turn, lead to disease. Indeed, 24 of the regions that Bailey et al. identified are associated with disease.

Birth and Death of Duplicates: Duplicate Gene Life History

Data from the human, mouse, fly, worm, rice, *Arabidopsis*, and yeast genome sequencing projects were used by Lynch & Conery (81) to address questions regarding the evolutionary impact of gene duplication. The most obvious and, perhaps most surprising, result of their study was that genes are duplicated approximately as frequently as individual nucleotides are substituted; new genes per genome per generation is in the hundreds. These authors (81) also found that humans and worms make new genes faster than do *Drosophila*, *Arabidopsis*, and yeast. They concluded, from the pattern of nucleotide substitution and from the frequency distribution of gene ages (estimated from mutations at silent sites), that duplicates experience a brief period of relaxed selection and that most become nonfunctional very quickly, i.e., by the time silent sites have diverged by only a

few percent. This second conclusion seemed to be inconsistent with data from tetraploid species, which retain a large proportion of duplicates. Lynch & Conery suggested that selection might preferentially retain duplicates produced during whole-genome duplication events in order to maintain relative gene dosage. Regarding the evolutionary implications of their results, they argued that the high rate of gene duplication has the potential to generate substantial molecular substrate for the origin of evolutionary novelties, but also that the window of time for such evolutionary "exploration" by gene duplicates is narrow. They proposed that the most significant consequence of gene duplication might be speciation caused by postreproductive isolation due to the loss of different duplicates in different populations (i.e., reciprocal silencing or divergent resolution).

Vandepoele et al. (129) used a BLASTp-plus-phylogeny reconstruction approach to survey the human and pufferfish genomes for duplicated genes and to estimate the ages of the duplicates they uncovered. Genes that occur once in *Ciona* and/or *Drosophila*, at least once in pufferfish, and from two and ten times in humans were identified using BLASTp. Phylogenetic trees were reconstructed and the ages of 447 duplication events were estimated. Vandepoele et al. concluded that a large number of human duplicates (360) were formed prior to the divergence of actinopterygians and sarcopterygians. The remaining 87 nodes (174 human paralogs) were formed relatively recently, i.e., during the past 50 million years. A large proportion of these young human duplicates were found to be linked and, therefore, probably formed by tandem duplication events.

Nembaware et al. (95) noticed that recently duplicated human genes (those with five or fewer synonymous substitutions per 100 synonymous positions) tended to be shorter than older paralogous pairs (those with 34 to 74 synonymous substitutions per 100 synonymous sites). They proposed that the probability of a whole gene being duplicated is correlated to its length and they suggested that gene length data might be useful for settling debates on gene versus genome duplication.

Identifying the Duplicates that Stand the Test of Time

In Gevers et al.'s (39) survey of 106 bacterial genomes, a large proportion of retained duplicates were ABC-type transporters, transcription factors, and dehydrogenases. Genomes with unique gene expansion patterns included *Borrelia burgdorferi* with an excess, relative to other species surveyed, of motility and chemotaxis genes. *Bacteroides thetaiotaomicron* has the largest gene family among the species surveyed. Gevers et al. (39) suggested that this species' 77 outer membrane proteins are involved in nutrient binding.

Conant & Wagner (19) also investigated functional biases in retained duplicates in the fully sequenced genomes of *S. cerevisiae*, *Schizosaccharomyces pombe*, *D. melanogaster*, *C. elegans*, and *E. coli*. After assigning genes to functional categories using the Gene Ontology or GO database, they used a BLASTP-based approach to determine whether each gene was a single-copy gene, a gene with one paralog, or a gene with multiple paralogs. Next, they asked whether genes with one

or genes with more than one paralog were over- or underrepresented in each functional class. The Ribosomal Protein Gene category had few genes with multiple paralogs in all species except *S. pombe*, suggesting selection against the retention of duplicated ribosomal proteins. For both yeast species, ribosomal protein genes often had one paralog. This observation was thought to be a consequence of polyploidy. That is, when ribosomal proteins are duplicated as part of a whole-genome duplication event, both copies are retained. Earlier, Seoighe & Wolfe (110) had also noted this overrepresentation of ribosomal proteins in their analysis of the genes retained after the genome duplication in *S. cerevisiae*. In addition, they reported an excess of cyclins and protein classes linked to signal transduction. Several other biases in the proportion of duplicated genes in functional categories were detected in Conant & Wagner's (19) study, including an overrepresentation of protein kinases (protein metabolism) in *D. melanogaster* and histone genes (cell cycle/DNA processing) in *C. elegans*. Another interesting observation reported in the Conant & Wagner (19) study was that in the yeasts, members of large gene families had a higher proportion of amino acid substitutions than genes had in smaller families. This observation is consistent with one made almost 70 years earlier (48) that gene duplication buffers against the effect of harmful mutations.

Stein et al. (118) sequenced and annotated the *C. briggsae* genome and compared it with the *C. elegans* genome. They used TRIBE-MCL to identify gene clusters with more genes from one species than the other. There were 718 putative chemoreceptor proteins in *C. elegans*, but only 429 in *C. briggsae*. Another gene family with more representatives in *C. elegans* than *C. briggsae* was cyclin-like F-box genes (243 in *C. elegans* and 98 in *C. briggsae*).

The duplicated regions of the human genome uncovered by Bailey et al. (8) (see above) were defined by size not by gene content. When they looked at the genes within the recently duplicated blocks, they found that some genes were more likely to be duplicated than others. For example, genes associated with immunity and defense, membrane surface interactions, drug detoxification, and growth and development were particularly common within the recently duplicated segments (8).

Evolutionary Rate Variation in Duplicates

Davis & Petrov (23a) compared evolutionary rates at nonsynonymous sites in genes that had been duplicated in *S. cerevisiae* and *C. elegans* with genes that occurred only once in these species. The goal of this study was to determine whether fast- or slow-evolving genes were better at producing duplicates, not to determine how duplication itself influences evolutionary rates at nonsynonymous sites. Each gene from yeast and worm was placed into either the duplicate category or the singleton category according to copy number, and then the rates of nonsynonymous substitutions for orthologs of these two sets of genes were measured in the two fly species, *Drosophila melanogaster* and *Anopheles gambiae*. Fly orthologs of yeast and worm duplicates have much slower rates of evolution than fly orthologs of yeast

and worm singletons. The most plausible explanation involved gene expression levels. In yeast, the slowly evolving genes that have been preferentially retained in duplicate appear to be most highly expressed. However, in this study, codon bias was used as a proxy for levels of gene expression. Codon bias was not observed in *C. elegans*, so there may be another explanation for the link between evolutionary rate (at nonsynonymous positions) and duplicate gene retention in *C. elegans*.

Several studies have investigated rate variation between duplicates and their single-copy orthologs (or pro-orthologs) in other species. According to many of the models for postduplication gene evolution proposed since the 1930s, gene duplication is followed by a period of relaxed selection, at least for one of the duplicates. Therefore, the analysis of changes in evolutionary rates after duplication should provide a tool for studying functional divergence of duplicated genes.

In several small-scale studies, rate differences between duplicates have been investigated that yield no (22, 55), little (108), or considerable (128) evidence for evolutionary rate increase following duplication. The first genome-scale analyses tend to support a "very little" hypothesis (69, 145), although more recent analyses yield larger rate differences. A study of closely related gene triplets in four completely sequenced genomes shows that 20%–30% of duplicated genes show a significant difference in evolutionary rate among each other (20). A comparison of duplicated genes in *S. cerevisiae* with their orthologs in *Kluyveromyces waltii* showed that in 17% of the cases, one or both duplicates had undergone accelerated evolution since the duplication event (67).

Nembaware et al. (95) exploited the genome sequence data available for human and mouse to investigate the effect of a paralogous gene in human on the divergence between its duplicate and the ortholog (or pro-ortholog) in mouse. They tested the hypothesis that a given human gene diverges from its mouse ortholog faster when it has a paralog. This study (95) showed that genes with distantly related paralogs evolve faster, whereas genes with closely related paralogs appear not to. This paralog-induced increased evolutionary rate was most prominant at nonsynonymous positions. Thus, a distantly related paralog appears to promote amino acid sequence divergence in its sister gene. If the duplicates have divided the expression domains of their single-copy ortholog in mouse, then selection might have been relaxed in the protein-coding portion of the gene. Thus, Nembaware et al. (95), in explaining their observations, reiterate the proposition of Force et al. (30) that changes in duplicate gene expression might facilitate the divergence of protein-coding sequences.

Gribaldo et al. (42) studied duplication and divergence in hemoglobin genes. They tested the hypothesis that site-specific changes in evolutionary rates in a member of a duplicate pair are correlated with functional divergence. Phillipe et al. (102) also investigated the correlation between the evolution of new gene function in paralogs and the rate of DNA sequence evolution. The premise for both studies was that new function is a consequence of a change in protein structure. Phillipe et al. (102) concluded that a significant increase in evolutionary rate is not an indicator of functional change; however, they did find a correlation between

constant but different (CBD) substitutions and functional divergence in proteins. CBDs occur when a typically constant amino acid residue changes once in the phylogenetic tree but not again. Creevy & McInerey (21) considered such mutations to be evidence for directional adaptive evolution, and they proposed that the discovery of CBDs, or "invariable replacement" (IR) substitutions, was an alternative to the traditional way of detecting positive selection, which involves comparing the rates of nonsynonymous and synonymous substitution. The idea behind the CBD/IR approach is that during an episode of positive directional selection, advantageous substitutions will occur at positions that then remain invariable at a rate significantly higher than expected from the neutral model.

Many other methods are used for investigating site- and branch-specific rate variation within gene families (i.e., after duplication), and for uncovering evidence of postduplication positive selection (reviewed by 105). Such methods have been used to identify mutations that have led to functional shift after gene duplication in a large amount of gene families.

Functional Divergence Between Duplicates

Over the past few years, functional genomics data have been increasingly used to study functional divergence of duplicated genes. In a pioneering study in yeast, Seoighe & Wolfe (111) detected a positive correlation between expression level and gene retention after duplication. This observation was later corroborated by Krylov and co-workers (72) who showed a negative correlation between expression level and propensity for gene loss. Wagner (134) analyzed the expression patterns of 124 duplicated pairs of yeast genes using a compilation of 79 microarray experiments. He showed that there was almost no correlation between the divergence in expression pattern and the evolutionary distance of the corresponding proteins, implying a decoupling of the rate of coding sequence evolution and that of expression divergence after duplication. Later, Gu and coworkers (44) showed, also in yeast, that these two rates are coupled, but only for a brief period after duplication. In addition, a significant correlation was found between expression divergence and the number of synonymous substitutions per site between duplicates, which shows that expression divergence increases with evolutionary time. This observation was also supported by a negative correlation between the number of conserved regulatory elements between duplicates and time (99).

A recent study investigating microarray-based expression data of human duplicated genes confirmed these observations, and showed that when the generation time of both species was taken into account, expression divergence was more rapid in humans than in yeast (83). In addition, proteins involved in the immune response appeared to show a more rapid divergence in expression after duplication. Wagner (135) showed, by analyzing large-scale interaction data in yeast, that duplicated genes rapidly diverge in the number of shared interaction partners. In a follow-up study, combining interaction and microarray data, the same author found that the divergence of duplicated genes through complementary degenerations of multiple

functions is often asymmetric, i.e., one gene loses more functions than the other (136).

SUMMARY AND CRYSTAL BALL

The large number of recent papers that begin with "ever since Ohno" statements, including those written by us, fail to acknowledge a long history of inquiry into the occurrence and evolutionary consequences of gene and genome duplication. Chromosome counts, studies of chromosome morphology, estimates of DNA content, and isozyme electrophoresis have made significant empirical and theoretical contributions to research on gene and genome duplication. These studies, combined with those dealing with gene and whole-genome sequences, show that gene duplication is a common occurrence, that the rate of duplication varies among species and among genes (shorter genes appear to be duplicated more often than long genes and slowly evolving genes produce duplicates more often than those with a rapid rate of evolution). Whether a duplicate is retained depends upon its function, its mode of duplication (i.e., whether it was duplicated during a whole-genome duplication event), the species it occurs in, and its expression rate. No consensus has been reached with respect to evolutionary rate variation among paralogs and their single-copy orthologs. In some cases, duplication leads to an increase in evolutionary rate, but not in others. Furthermore, as Philippe et al. (102) pointed out, a change in rate might not be correlated with functional divergence. Constant but different (CBD) or Invariable Replacement (IR) substitutions appear to occur more frequently following gene duplication. Finally, many studies have shown that expression divergence and gene loss are phenomena that can occur shortly after or a long time after gene duplication.

The future of this field clearly lies in the integration of results from diverse research programs. As mentioned above, it is now possible to describe the paranome, compare expression domains among paralogs and delimit, using promoter bashing in transgenic organisms, mutations responsible for expression variation. It is also possible to compare interaction partners among paralogs, and to correlate this information with ever-increasing knowledge of the pathways that genes act in. Thus, a full description of the occurrence and evolutionary consequences of gene duplication, at least for some species, is within our grasp. Tempering this optimism is the acknowledgment that the rigor of every component of this research program can be drastically improved. However, new technologies will likely continue to advance the field as they have for more than 100 years. To give just one example, a new way to delimit duplicated fragments (coding and noncoding) in fully sequenced species is high-resolution array comparative genomic hybridization (CGH) (60). So far, this high-resolution or "submegabase resolution tiling" array CGH has been used with success to look for variation in gene content only between normal and cancerous cell lines, but comparisons between normal cells/individuals or between different species will, no doubt, soon be reported. Lynch & Conery's (81)

estimation that new genes arise at a rate of hundreds per generation suggests that array CGH will turn up interesting among-individual variation in gene content.

ACKNOWLEDGMENTS

The authors thank Martine De Cock for help with the retrieval of old papers and Yves Van de Peer for comments on the manuscript. We also apologize to all whose work was not included. J.R. is a postdoctoral fellow of the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie. J.S.T. is supported by an NSERC (Canada) Discovery Grant.

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

LITERATURE CITED

1. Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific silencing. *Proc. Natl. Acad. Sci. USA* 100:4649–54
2. Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, et al. 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics* 161: 259–67
3. Amores A, Force A, Yan Y-L, Joly L, Amemiya C, et al. 1998. Zebrafish *hox* clusters and vertebrate genome evolution. *Science* 282:1711–14
4. Amores A, Suzuki T, Yan Y-L, Pomeroy J, Singer A, et al. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* 14:1–10
5. Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, et al. 1997. Organization of the *Fugu rubripes* Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. *Nat. Genet.* 16: 79–83
6. Averof M, Dawes R, Ferrier D. 1996. Diversification of arthropod *Hox* genes as a paradigm for the evolution of gene functions. *Cell Dev. Biol.* 7:539–51
7. Avise JC, Kitto GB. 1975. Phosphoglucose isomerase gene duplication in the bony fishes: an evolutionary history. *Biochem. Genet.* 8:113–32
8. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. 2002. Recent segmental duplications in the human genome. *Science* 297:1003–07
9. Blakeslee AF. 1933. New Jimson weeds from old chromosomes. *J. Hered.* 25:81–108
10. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell.* 12:1093–101
11. Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13:137–44
12. Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–38
13. Bridges CB. 1935. Salivary chromosome maps. *J. Hered.* 26:60–64
14. Bridges CB. 1936. The bar "gene" a duplication. *Science* 83:210–11
15. Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* 165:349–57
16. Brooke NM, Garcia-Fernadez J, Holland PW. 1998. The ParaHox gene cluster is an

evolutionary sister of the Hox gene cluster. *Nature* 392:920–22
17. Cheng C-HC, Chen L. 1999. Evolution of an antifreeze glycoprotein. *Nature* 401:443–44
18. Chen L, DeVries AL, Cheng CH. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci. USA* 94:3811–16
19. Conant GC, Wagner A. 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30:3378–86
20. Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* 13:2052–58
21. Creevy CJ, McInerey JO. 2002. An algorithm for detecting directional and nondirectional positive selection, neutrality and negative selection in protein coding DNA sequences. *Gene* 300:43–51
22. Cronn RC, Small RL, Wendel JF. 1999. Duplicated genes evolve independently after polyploid formation in cotton. *Proc. Natl. Acad. Sci. USA* 96:14406–11
23. Darwin C. 1859. *The Origin of Species*. London: Penguin Books. 477 pp.
23a. Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PloS Biol.* 2:E55
24. Dermitzakis ET, Clark AG. 2001. Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.* 18:557–62
25. Dulai KS, von Dornum M, Mollon JD, Hunt DM. 1999. The evolution of trichromatic colour vision by opsin gene duplication in New World and Old World primates. *Genome Res.* 9:629–38
26. Edelmann L, Stankiewicz P, Spiteri E, Pandita RK, Shaffer L, et al. 2001. Two functional copies of the DGCR6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus. *Genome Res.* 11:208–17
27. Ermolaeva MD, Wu M, Eisen JA, Salzberg SL. 2003. The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol. Biol.* 51:859–66
28. Ferrier DE, Minguillon C, Holland P, Garcia-Fernandez J. 2000. The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. *Evol. Dev.* 2:284–93
29. Ferris S, Whitt GS. 1979. Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* 12:267–17
30. Force A, Lynch M, Pickett FB, Amores A, Yan Y-I, Postlethwait J. 1999. The preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–45
31. Force A, Amores A, Postlethwait JH. 2002. Hox cluster organization in the jawless vertebrate *Petromyzon marinus*. *J. Exp. Zool.* 294:30–46
32. Friedman R, Hughes AL. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* 11:373–81
33. Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. 1999. Ploidy regulation of gene expression. *Science* 285:251–54
34. Garcia-Fernandez J, Holland PW. 1994. Archetypal organization of the amphioxus Hox gene cluster. *Nature* 370:563–66
35. Gardner MJ, Shallom SJ, Carlton JM, Saltzberg SL, Nene V, et al. 2002. Genome sequence of the human malaria parasite *Plasmodiumfalciparum. Nature* 419:498–11
36. Gates MA, Kim L, Egan ES, Cardozo T, Sirotkin HI, et al. 1999. A genetic linkage map for zebrafish: comparative analysis and localization of gene and expressed sequence tags. *Genome Res.* 9:334–47
37. Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA* 94:6809–14
38. Gehring WJ. 1998. *Master Control Genes in Development and Evolution: The Homeobox Story*. New Haven: Yale Univ. Press. 254 pp.

39. Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.* 12:148–54
40. Gilad Y, Bustamante CD, Lancet D, Pääbo S. 2003. Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am. J. Hum. Genet.* 73:489–501
41. Goldschmit R. 1940. *The Material Basis of Evolution*. New Jersey, USA: Yale Univ. Press. 436 pp.
42. Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional divergence prediction from evolutionary analysis: a cases study of vertebrate hemoglobin. *Mol. Biol. Evol.* 20:1754–59
43. Gu Z, Nicolae D, Lu HH-S, Li W-H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18:609–13
44. Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* 31:205–9
45. Gu X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet.* 19:354–56
46. Gulick A. 1944. The chemical formulation of gene structure and gene action. *Adv. Enzymol.* 4:1–39
47. Haldane JBS. 1932. *The Causes of Evolution*. Ithaca, NY: Cornell Univ. Press. 235 pp.
48. Haldane JBS. 1933. The part played by recurrent mutation in evolution. *Am. Nat.* 67:5–19
49. Harris H. 1971. Polymorphism and protein evolution. The neural mutation-random drift hypothesis. *J. Med. Genet.* 8:444–52
50. Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* In press
51. Holland PW. 1997. Vertebrate evolution: something fishy about Hox genes. *Curr. Biol.* 7:R570–72
52. Holland PW, Garcia-Fernandez J. 1996. Hox genes and chordate evolution. *Dev. Biol.* 173:382–95
53. Hopkinson DA, Edwards YH, Harris H 1976. The distribution of subunit numbers and subunit sizes of enzymes: a study of the products of 100 human gene loci. *Ann. Hum. Genet. London* 39:383–11
54. Hua LV, Hidaka K, Pesesse X, Barnes LD, Shears SB. 2003. Paralogous murine *Nudt10* and *Nudt11* genes have differential expression patterns but encode identical proteins that are physiologically competent diphosphoinositol polyphosphate phosphohydrolases. *Biochem. J.* 373:81–89
55. Hughes MK, Hughes AL. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* 10:1360–69
56. Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. London Ser. B* 256:119–24
57. Hunt DM, Dulai KS, Cowing JA, Julliot C, Mollon JD, et al. 1998. Molecular evolution of trichromacy in primates. *Vision Res.* 38:3299–306
58. Huxley J. 1942. *Evolution: The Modern Synthesis*. New York: Harper & Brothers. 645 pp.
59. Irvine S, Carr JL, Bailey WJ, Kawasaki K, Shimizu N, et al. 2002. Genomics analysis of Hox clusters in the sea lamprey *Petromyzon marinus*. *J. Exp. Zool.* 294: 47–62
60. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* 36:299–303
61. Jenkin TJ. 1933. Interspecific and intergeneric hybrids in herbage grasses. Initial Crosses. *J. Genet.* 28:205–64
62. Jacobs GH. 1996. Primate photopigments

and primate color vision. *Proc. Natl. Acad. Sci. USA* 93:577–81
63. Jensen RA. 1976. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* 30:409–25
64. Jorden IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11:555–65
65. Jørgensen CA. 1928. The experimental formation of heteroploid plants in the genus *Solanum*. *J. Genet.* 19:133–11
66. Kashkush K, Feldman M, Levy AA. 2002. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160:1651–59
67. Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617–24
68. Koch AL. 1972. Enzyme evolution: I. The importance of untranslatable intermediates. *Genetics* 72:297–16
69. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol* 3: RESEARCH0008
70. Kourakis MJ, Martindale MQ. 2000. Combined-method phylogenetic analysis of Hox and ParaHox genes of the metazoa. *J. Exp. Zool.* 288:175–91
71. Kratz E, Dugas JC, Ngai J. 2002. Odorant receptor gene regulation: implications from genomic organization. *Trends Genet.* 18:29–34
72. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13:2229–35
73. Kuwada Y. 1911. Meiosis in the pollen mother cells of *Zea Mays* L. *Bot. Mag.* 25:163
74. Larhammar D, Lundin LG, Hallbook F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res.* 12:1910–20
75. Lewis EB. 1951. Pseudoallelism and gene evolution. *Cold Spring Harbor Symp. Q. Biol.* 16:159–74
76. Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54:595–609
77. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* 303:540–43
78. Lindholm A, Brooks R, Breden F. 2004. Extreme polymorphism in a Y-linked, sexually selected trait. *Heredity* 92:156–62
79. Lister JA, Close J, Raible DW. 2001 Duplicate mitf genes in zebrafish: complementary expression and conservation of melanogenic potential. *Dev. Biol.* 237: 333–44
80. Locascio A, Manzanares M, Blanco MJ, Nieto MA. 2002. Modularity and reshuffling of Snail and Slug expression during vertebrate evolution. *Proc. Natl. Acad. Sci. USA* 99:16841–46
81. Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–55
82. Lynch M, Force AG. 2000. The origin of interspecific genomic incompatibility via gene duplication. *Am. Nat.* 156:590–605
83. Makova KD, Li W-H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13:1638–45
84. Malaga-Trillo E, Meyer A. 2001. Genome duplications and accelerated evolution of Hox genes and cluster architecture in teleost fishes. *Am. Zool.* 41:676–86
85. McClintock JM, Carlson R, Mann DM, Prince VE. 2001. Consequences of Hox gene duplication in the vertebrates: an investigation of the zebrafish Hox paralogue

group 1 genes. *Development* 128:2471–84

86. McGinnis W, Garber R, Wirz J, Kuroiwa A, Gehring WJ. 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* 37:403–8
87. McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31:200–4
88. Metz CW. 1947. Duplication of chromosome parts as a factor in evolution. *Am. Nat.* 81:81–103
89. Meyer A, Schartl M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* 11:699–704
90. Mirsky AE, Ris H. 1951. The Desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* 34:451–62
91. Modolell J, Campuzano S. 1998. The achaete-scute complex as an integrating device. *Int. J. Dev. Biol.* 42:275–82
92. Muller HJ. 1935. The origination of chromatin deficiences as minute deletions subject to insertion elsewhere. *Genetica* 17:237–52
93. Müntzing A. 1936. The evolutionary significance of autoployploidy. *Hereditas* 21:263–78
94. Naruse K, Fukamachi S, Hirochi M, Kondo M, Matsuoko T, et al. 2000. A detailed linkage map of Medaka, *Oryzias latipes*: comparative genomics and genome evolution. *Genetics* 154:1773–84
95. Nembaware V, Crum K, Kelso J, Seoighe C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res.* 12:1370–76
96. Nishiyama I. 1934. The genetics and cytology of certain cereals. Chromosome behaviour and its bearing on inheritance in triploid *Avena* hybrids. *Mem. Coll. Agric. Kyoto Imp. Univ.* 32:1–157
97. Ohno S. 1967. *Sex Chromosomes and Sex-linked Genes*. Berlin: Springler-Verlag. 192 pp.
98. Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer-Verlag. 150 pp.
99. Papp B, Pal C, Hurst LD. 2003. Evolution of *cis*-regulatory elements in duplicated genes of yeast. *Trends Genet.* 19:417–22
100. Pearson WR, Wood TC. 2001. Statistical significance in biological sequence comparison. In *Handbook of Statistical Genetics*, ed. DJ Balding, M Bishop, C Cannings, 2:39–65. Toronto: Wiley & Sons
101. Pébusque MJ, Coulier F, Birnbaum D, Pontarotti P. 1998. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* 15:1145–59
102. Phillipe H, Casane D, Gribaldo S, Lopez P, Meunier J. 2003. Heterotachy and functional shift in protein evolution. *IUBMB Life* 55:257–26
103. Piatigorsky J, Wistow G. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science* 252: 1078–79
104. Powers TP, Amemiya CT. 22004. Evidence for Hox-14 paralog group in vertebrates. *Curr. Biol.* 14:R183–84
105. Raes J, Van de Peer Y. 2003. Gene duplications, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl. Bioinformat.* 2:92–101
106. Raes J, Vandepoele K, Simillion C, Saeys Y, Van de Peer Y. 2003. Investigating ancient duplication events in the *Arabidopsis* genome. *J. Struct. Funct. Genomics* 3: 117–29
107. Rhoades MM. 1951. Duplicate genes in maize. *Am. Nat.* 85:105–49
108. Robinson-Rechavi M, Laudet V. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Mol. Biol. Evol.* 18: 681–83

109. Schughart K, Kappen C, Ruddle FH. 1989. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proc. Natl. Acad. Sci. USA* 86:7067–71
110. Seoighe C, Wolfe KH. 1999. Updated map of duplicated regions in the yeast genome. *Gene* 238:253–61
111. Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* 2:548–54
112. Serebrovsky AS. 1938. Genes *scute* and *achaete* in *Drosophila melanogaster* and a hypothesis of gene divergency. *C. R. Acad. Sci. URSS* 19:77–81
113. Sharman AC. 1999. Some new terms for duplicated genes. *Cell Dev. Biol.* 10:561–63
114. Simillion C, Vandepoele K, Van de Peer Y. 2004. Recent developments in computational approaches for uncovering genomic homology. *BioEssays*:In press
115. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 99:13627–32
116. Spring J. 1997. Vertebrate evolution by interspecific hybridization—are we polyploid? *FEBS Lett.* 400:2–8
117. Stadler LJ. 1929. Chromosome number and the mutation rate in *Avena* and *Triticum*. *Proc. Natl. Acad. Sci. USA* 15: 876–81
118. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLOS Biol.* 1:166–92
119. Stephens SG. 1951. Possible significance of duplications in evolution. *Adv. Genet.* 4:247–65
120. Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49:169–81
121. Stuber CW, Goodman MM. 1983. Inheritance, intracellular localization, and genetic variation of phosphoglucomutase isozymes in maize (*Zea mays* L.). *Biochem. Genet.* 21:667–89
122. Sturtevant AH. 1925. The effects of unequal crossing over at the bar locus in *Drosophila*. *Genetics* 10:117–47
123. Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc.* 356:1661–79
124. Taylor JS, Van de Peer Y, Meyer A. 2001. Genome duplication, divergent resolution and speciation. *Trends Genet.* 17:299–301
125. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* 13:382–90
126. Tischler G. 1915. Chromosomenzahl, Form und Individualität in Planzenreiche. *Progr. Rei Bot.* 5:164
127. Tischler G. 1935. Studien über *Festuca ovina* L. III. Weitere Beiträge zur Kenntnis der Chrosomensahlen vivparer Formen. *Hereditas* 15:13–16
128. Van de Peer Y, Taylor JS, Braasch I, Meyer A. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.* 53:436–46
129. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. USA* 101:1638–43
130. Vandepoele K, Simillion C, Van de Peer Y. 2003. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell.* 15:2192–202
131. Vandepoele K, Simillion C, Van de Peer Y. 2004. The quest for genomic homology. *Curr. Genomics.* In press
132. Vidalain P-O, Boxem M, Ge H, Li S, Vidal M. 2004. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* 32:363–70

133. Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290:2114–17
134. Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. USA* 97:6579–84
135. Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18:1283–92
136. Wagner A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Mol. Biol. Evol.* 19:1760–68
137. Waters AP. 1994. The ribosomal RNA genes of Plasmodium. *Adv. Parasitol.* 34: 33–79
138. Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploidy pteridophytes resulting from silencing of duplicate-gene expression. *Am. Nat.* 137:515–26
139. Wistow G, Piatigorsky J. 1987. Recruitment of enzymes as lens structural proteins. *Science* 236:1554–56
140. Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–13
141. Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2:333–41
142. Wong S, Butler G, Wolfe KH. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl. Acad. Sci. USA* 99:9272–77
143. Yu W-P, Brenner S, Venkatesh B. 2003. Duplication, degeneration and subfunctionalization of the nested *synapsin-Timp* genes in *Fugu*. *Trends Genet.* 19:180–83
144. Zhang J, Zhang Y-P, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* 30:411–15
145. Zhang L, Vision TJ, Gaut BS. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 19: 1464–73

Annual Review of Genetics
Volume 38, 2004

CONTENTS