CHAPTER 3

The Origin of New Genes

Chuanzhu Fan, J. J. Emerson, and Manyuan Long

Introduction

PARALLEL TO DARWIN'S CENTRAL QUESTION OF THE ORIGIN OF SPECIES is that of the origin of novel genetic elements. One of the most important roles of such new elements is to generate genes with new functions, which increases the biological diversity of organisms. Initial speculation claimed that gene origination must be accompanied by gene duplication (Ohno 1970), although recent studies provide evidence for other mechanisms (Long et al. 2003). In order to understand the molecular processes and mechanisms governing the evolution of novel genes and their functions, direct observation of newly originated gene copies is an indispensable approach. It is well established that many genes have persisted for long evolutionary times. Studying evolution over such timescales is difficult because the characteristic features that enable the elucidation of the evolutionary process erode with increasing time. Therefore, one productive strategy to learn more about gene origination is to investigate recently evolved genes. In this chapter, we will introduce the general features of new gene evolution, ranging from new genes that have been found in various organisms to molecular mechanisms and patterns of new gene origination. We will focus on the methods used to detect new genes and will describe a general genomic method adapted from microarray hybridization technology.

Recent Discoveries of New Genes

Experimental studies on new gene origination emerged in the early 1990s when a newly evolved gene, *Jingwei* (*Jgw*), was identified from two African species of *Drosophila* (Long and Langley 1993). Since then, by taking advantage of the availability of phylogenetic frameworks and rapidly expanding databases for the major model species, many examples of new genes

Table 3.1 Examples of new genes that have been fully described in mammals, flies, and plants

Organism	New gene	Parental gene(s)	Mechanism(s)	Age (my[a])
Primates[b]	*PGAM3*	Phosophoglycerate mutase (*Pgam*)	Retroposition	10
	RNASE1B	Pancreatic ribonuclease gene (*RNAse1*)	Gene duplication	4
	PMCHL1	Melanin-concentrating hormone (MCH)	Exon shuffling + retrotransposition	20
	PMCHL2	Melanin-concentrating hormone (MCH)	Segmental duplication	2.5–5
	Morpheus	*Morpheus*	Segmental duplication	12–25
	Tre2 (USP6)	*USP32 (NY-REN-60)* and *TBC1D3*	Segmental duplication	21–33
	CGβ	Luteinizing hormone β subunit gene (*LHbeta*)	Gene duplication	34–50
Rodents	*4.5Si RNA*	RNA gene	Gene duplication	25–55
	BC1 RNA	$tRNA^{Ala}$	Gene duplication	60–100
	Insulin 1(Ins1)	*Ins2*	Retroposition	10–15
Flies	*Jingwei*	*Alcohol dehydrogenase (Adh)+ yellow emperor*	Retroposition and fusion	2.5
	Adh-Finnegan	*Adh* + unknown sequence	Gene duplication and fusion	30
	Adh-Twain	*Adh* + CG9010	Retroposition and fusion	5
	Sdic	*AnnX* + *Cdic*	Gene fusion	<3
	Exuperantia2	*Exu1*	Ectopic recombination + transposition	25
	Sphinx	*ATP synthase chain F*	Retroposition	2.5

have been fully described in eukaryotes from protozoa to *Drosophila* to primates. Here, we will briefly summarize our knowledge of novel genes identified in various organisms and the mechanisms thought to govern their origination and subsequent evolution (for a detailed review, see Long et al. 2003).

Drosophila species have served as the best model organisms for new gene research since the 1990s. Fourteen new genes, derived within the last 30 million years, have been fully described in several species of *Drosophila* (Table 3.1). For example, *Adh-Twain* was created by the fusion of a retroposed *Adh* sequence with a target gene in the common ancestor of *D. subobscura*, *D. madeirensis*, and *D. guanche* (Jones et al. 2005). *Siren*, in the *D. bipectinata* complex, is another chimeric gene involving *Adh*, also created by retroposition (Nozawa et al. 2005). New *Monkey King* genes were formed by duplication, followed by partial degeneration in complementary parts of the parent and copy gene sequences, and final fusion of these two adjacent genes in the

Table 3.1 *Continued*

Organism	New gene	Parental gene(s)	Mechanism(s)	Age (my[a])
Flies	*Dnth-2r*	*Nuclear transport factor*	Retroposition	5
	Monkey King	CG7163	Gene fusion and fission	1–2
	K81	CG14251	Retroposition	15
	siren	*Nanos*+CG11779	Retroposition	20
	Ifc-2h	*Infertile crescent* (*Ifc*)	Retroposition	1–2
	Hun Hunaphu	*Baochen*	Illegitimate recombination	1–2
	Ifc-2h	*Ifc*	Retroposition	2
	Quijote (CG13732)	*Cervantes* (CG15645)	Retroposition	5
Fish	Arctic *AFGP*	Polyprotein	Gene duplication	2.5
	Antarctic *AFGP*	Pancreatic trypsinogen	Gene conversion and duplication	5–14
Plants	*Sanguinaria rps1*	*rps1*	Gene transfer from mitochondrion to nucleus	45
	Plantago ap1	*Bartsia ap1*	Gene transfer from host to parasite plant	?
	Cytochrome c1	*Cytochrome c1*	Exon-shuffling	>110
	Nuclear *Cox2*	Mitochondrial *cox2*	Gene transfer from mitochondrion to nucleus in legume and exon-shuffling	50
	At1g71920	*At5g10330* (histidinol phosphate aminotransferase-like gene)	Gene duplication	0.5
	At1g05090	*At4g20720* (unknown function)	Gene duplication	0.6

[a]Abbreviation: my, million years.

[b] See Marques et al. (2005) for more primate-specific new retrogenes.

common ancestor of *D. simulans*, *D. mauritiana*, and *D. sechellia* (Wang et al. 2004). *Hun Hunaphu* was identified as a young gene created by illegitimate recombination at some time after *D. simulans*, *D. mauritiana*, and *D. sechellia* diverged from *D. melanogaster* (Arguello et al. 2006). *Quijote* is a recognizable retroposed copy of CG15645 present in the branch leading to *D. melanogaster* and *D. simulans* (Bertran et al. 2006). Sequence divergence and polymorphism analyses showed, in all these examples, that directional selection (positive selection or a recent selective sweep) played a crucial functional role in the early stage of their evolution.

Among vertebrates, a number of new genes have been found in fish, rodents, and primates (see Table 3.1). For example, *RNAse1B* is a new duplicate under strong positive selection because it is involved in the unique digestive system of leaf-eating colobine monkeys (Zhang et al. 2002b). *Pgam3* is a primate lineage–specific retroposed gene with testis-biased expression (Betran et al. 2002). *Clorf37-dup* is a human-specific retroposed gene, driven

to evolve rapidly by positive Darwinian selection (Yu et al. 2006). *Ins1* is a rodent-specific insulin gene which was derived by retroposition 10–15 million years ago from *Ins2*, and is under positive Darwinian selection (M. S. Shiao, M. Long, and H. T. Yu, unpublished data). *Ubl4b* is a novel retroposed mouse ubiquitin-like protein with a testis-specific expression (Yang et al. 2007). In fish, antifreeze glycoproteins (*AFGPs*) help in adaptation to cold environments. Sequence comparisons indicate that two *AFGPs* arose independently in Arctic and Antarctic fishes through extensive gene duplication and gene conversion (Chen et al. 1997a,b). Finally, *TRIM5-CypA* in owl monkeys has been shown to be a novel chimeric gene, which evolved to resist HIV-1 virus infection (Nisole et al. 2004; Sayah et al. 2004).

Compared to other taxa, fewer young genes have been described in plants. The reason may be that whole genome duplications through hybridization are a common process contributing to plant species diversity (Otto and Whitton 2000). This may lessen the selective need for new gene origination via single gene duplication. Alternatively, retroposition may be relatively less common in plants because they lack active L1 retrotransposons. It is hard to evaluate these possibilities as the scarcity of new young genes may simply reflect the relative lack of attention this subject has received from plant biologists.

Nevertheless, several insights have been reached since genomic sequence data were completed for model plant species. First, horizontal gene transfer, which involves gene movements between species or between cytoplasm and nucleus, is far more frequent in the plant kingdom than in other organisms (see Table 3.1) (Bergthorsson et al. 2003; Park et al. 2007; Richardson and Palmer 2007). Second, a number of retroposed genes have been identified in the Arabidopsis genome by computational and experimental approaches, and some of them are newly derived genes that only occur within the *Arabidopsis* genus or are unique to *A. thaliana* (Zhang et al. 2005). For example, *At1g61410* and *At5g52090* are two retroposed genes derived from mRNA. Sequence analyses of these two genes indicate that they are only present in *Arabidopsis* species derived from Mediterranean Pleistocene refugia (Zhang et al. 2005). Third, a recent study using rice genome data demonstrates that extensive retroposition has resulted in thousands of functional retrogenes during grass genome evolution (Wang et al. 2006). A large proportion of these retrosequences are chimeric, having recruited new exons, introns, and coding regions from the sites in which they have inserted. Finally, Moore and Purugganan (2003) showed reduced nucleotide polymorphism in two new gene duplicates in *A. thaliana*, suggesting again a possible role for positive selection in the evolution of these genes.

Mechanisms to Generate New Genes

Molecular mechanisms involved in creating novel gene structures are now well understood and are described in detail elsewhere (Long et al. 2003).

Here, we will briefly describe the main forces at play. Many new genes have been created through a combination of two or more of the following mechanisms (see also Table 3.1):

- **Gene duplication** can be achieved at several cellular levels: whole genome duplication, segmental duplication, or single tandem gene duplication. Duplication allows the original functions to be maintained by one copy, while the second copy provides a substrate for other mechanisms leading to new gene origination (see below).
- **Exon shuffling** is achieved by illegitimate recombination of exons or retroposed exon insertions that create a new exon-intron gene structure. Such processes can lead to a new gene with novel function. The shuffling mechanism has been found to generate numerous chimeric proteins (e.g., Wang et al. 2005; Li et al. 2007). It can also create chimeric genes with previously unrelated regulatory sequences, for example, the *Siren* (Nozawa et al. 2005) and *Ste* genes in *D. melanogaster* (Usakin et al. 2005).
- **Retroposition** is a mechanism that generates new intronless gene copies (retrogenes) by reverse transcription of mRNA derived from parental genes. Three hallmarks can be used to identify retrogenes: (1) one member of the pair is intronless while the other contains introns in the coding regions; (2) the new (intronless) copy contains a poly(A) tail; and (3) the new copy may still have short duplicate flanking sequences. Experimental and computational genomics studies have both found large numbers of retroposed genes in eukaryotes, including yeasts, plants, and animals.
- **Mobile elements** can pick up host sequences and integrate them into new genomic positions. If the site of integration is near or within existing coding sequences, a new, chimeric gene structure can be generated. Many cases of new genes created by mobile elements have been described. For example, Pack-MULEs can recruit small chromosome fragments and combine with other genomic regions when they transpose to form chimeric gene structures (Jiang et al. 2004). Helitrons, which are helicase-bearing transposable elements, are likewise capable of shuffling genomic regions (Bennetzen 2005).
- **Horizontal gene transfer** is the movement of genes from one species to another or between organelles and the nucleus. This event occurs frequently in bacteria and yeasts. There is also some evidence for it occurring in plants (see Chapter 4; Bergthorsson et al. 2003).
- **Gene fusion/fission** occurs when two adjacent genes fuse together to form a single gene, or when a single gene splits into two genes that then evolve different functions. One example of gene fusion/fission is the formation of the *Monkey King* gene described above.

- **De novo origination** is a final possible source of new genes which cannot be ruled out, even though no clear evidence has been reported for such an origin of a complete protein-coding gene. There is some evidence that frame-shift mutations in a number of duplicated vertebrate genes created quasi-random sequences from which new genes subsequently evolved (Raes and Van de Peer 2005). This supports the possibility of de novo origins of protein-coding genes. More recently, Begun and colleagues (2007) found a significant number of X-linked testis-biased de novo noncoding RNA genes in the *D. yakuba/D. erecta* clade.

Evolutionary Forces for New Gene Retention

Positive Darwinian selection is likely to be the most important force acting for the retention and evolution of novel genes (Long et al. 2003). Particularly, most new genes that originated through exon shuffling and gene duplication have undergone significantly accelerated rates of evolution compared to their parental copies. For example, *Jingwei* has a significantly higher rate of substitution in its protein sequences and gene structure, and sequence divergence analysis suggests a high rate of protein adaptive evolution (Long and Langley 1993).

Two methods have been used to test whether positive selection acted on novel genes during their evolution. The first is to estimate the K_a/K_s ratio in new gene lineages (K_a = the nonsynonymous substitution rate, K_s = the synonymous substitution rate). For example, *RNASE1B* is a new, duplicate ribonuclease gene that arose 4 million years ago in the leaf-eating colobine monkey. The K_a/K_s ratio (4.03) of *RNASE1B* is significantly higher than unity. In contrast, its paralog, *RNASE1*, has accumulated no amino acid substitutions over the same time period (Zhang et al. 2002b).

The second method to test for positive selection is to compare sequence divergence between species and sequence polymorphism within species, formalized as the McDonald–Kreitman (1991) test of neutral molecular evolution. Positive selection is indicated when there is an excess of amino acid replacement substitutions between species compared to the neutral prediction that the variation of replacement substitutions and synonymous substitutions should be positively correlated. For example, McDonald–Kreitman tests of *Hun Hunaphu*, a recently evolved (within the last 2–3 million years) chimeric gene found in *D. simulans*, *D. sechellia*, and *D. mauritiana*, reveal it to have been subject to positive selection in the *D. simulans* branch (Arguello et al. 2006).

However, we cannot rely on analyses of individual cases to determine whether positive selection has a general role in driving new gene evolution and retention. An approach that uses genomic data derived from different chromosome regions might be a feasible way to detect general forces that drive the evolution of duplicated genes. Thornton and Long (2002)

compared the K_a/K_s ratios of more than 100 paralogous gene pairs on the X chromosome with 1743 paralogs on the autosomes in *D. melanogaster*. They found that X-linked duplicates have higher K_a/K_s ratios than autosomal duplicates. They further estimated the K_a/K_s ratios of single-copy genes and found no accelerated rate of amino acid substitutions of X-linked genes. Such an inconsistency suggests that different forces might act on single-copy and newly duplicated genes, which likely acquire new functions under positive selection.

The Location and Movement of New Genes

A new gene can be located adjacent to (tandem duplication) or far away from its parental copy. A random distribution of new gene movement might be expected. However, studies of the pattern of gene movements show a surprising asymmetry. Betran and colleagues (2002) computationally screened the genome sequence data of *D. melanogaster* to check the location of retroposed genes and their parental copies. They found that there was a significant excess of retrogenes originating from the X chromosome and retroposed to autosomes, and relatively few new genes retroposed in the opposite direction. This result was further supported by a recent study that investigated retrogene movement between and within chromosomes in the *D. melanogaster* genome (Dai et al. 2006). Emerson and coworkers (2004), extending this approach to human and mouse, showed that the mammalian X chromosome also generated a significantly higher number of functional retroposed genes than autosomes. In contrast to *D. melanogaster*, mammalian X chromosomes also recruited an excess of new retrogenes. Further experiments in *D. melanogaster* demonstrated that most new autosomal retroposed copies exhibited testis-biased expression, unlike the parental X-linked genes (Betran et al. 2004; Emerson et al. 2004). These observations provide strong evidence that genome position also plays a very important role in the recruitment of new gene copies (Betran et al. 2004).

Positional effects of new gene locations have also been observed in plants. In a study of retroposed genes in the rice genome, Wang and colleagues (2006) observed that functional retrogenes tend to "avoid" centromeric regions and prefer to insert into the middle of chromosomal arms. Compared to the random distribution of processed pseudogenes, the biased distribution of functional retrogenes probably reflects natural selection.

Inferring the Functionality of New Genes

In principle, the functionality of a new gene can be inferred in both direct and indirect ways. Direct experimental tests in various functional analyses provide explicit information about a new gene's functions, but these are costly and time-consuming. Indirect approaches, using bioinformatic techniques, are therefore valuable, as they are easily accomplished and pro-

vide candidates for further direct functional analysis. The simplest way is to detect evolutionary constraints that are associated with functional genes.

We can examine evolutionary constraints by calculating the K_a/K_s ratio in a new gene lineage. $K_a/K_s < 1$ indicates strong functional constraint under purifying selection. $K_a/K_s = 1$ indicates no functional constraints and neutral evolution—typical of pseudogenes. $K_a/K_s > 1$ indicates an accelerated amino acid evolution rate under positive selection. A more appropriate model may be to assume that the parental gene is subject to strong purifying selection with $K_a/K_s = 0$, whereas the new gene is a functionless pseudogene with $K_a/K_s = 1$. Therefore, the substitution rates calculated by comparing the parental copy and new pseudogene copy yield $K_a/K_s = 0.5$. Thus, a ratio of $K_a/K_s < 0.5$ suggests the new gene copy is likely to be functional, $K_a/K_s = 0.5$ suggests it may be a functionless pseudogene, and $K_a/K_s > 0.5$ suggests it may have experienced positive selection (Thornton and Long 2002). It should be noted that the criterion of $K_a/K_s = 0.5$ is very conservative, because it was derived based on the specific assumptions of evolutionary stagnation of the parental gene and equal synonymous substitution rates for the new and parental genes. The functional specificity of new genes can also be explored by analyzing their expression profiles. By comparing a new gene's transcription pattern (with respect to tissue or developmental stage) to that of its parental gene, we can tell whether the new gene has acquired new functions. For example, *Dnth-2r* was identified as a new retroposed gene that is only transcribed in testis, while its parental gene is ubiquitously expressed in both sexes of *D. melanogaster* (Betran et al. 2003).

There are two general direct approaches to detecting function in new genes. First, we can use biochemistry and immunological technology to obtain protein products, and then test their functions in vitro. For example, Jones and colleagues (2005) used western blotting to analyze the protein synthesized from a new chimeric retroposed fusion gene. Zhang and colleagues (2004) investigated the function of the new *Jingwei* gene in *Drosophila* by studying the enzymatic properties of JGW proteins collected from a microbial expression system. They found JWG was a novel dehydrogenase with alternative substrate specificity compared with the ancestral ADH protein. JGW protein also uniquely prefers to catalyze reactions involving the long-chain primary alcohols found in insect pheromone metabolism.

As a second direct approach, the functions of new genes can be tested by genetic silencing in vivo, for example by gene knockout or RNAi. In addition, transgenic lines carrying *gene::GFP* fusions can be produced to observe the location of a new gene's expression, which can provide insights into its biological functions (Loppin et al. 2005). Competition experiments between a strain that carries the new gene and a strain in which it is silenced can be used to measure the effect of the new gene on fitness. Observation of changes at the phenotypic, physiological, behavioral, and population genetic levels would provide further understanding of gene function, but as yet this approach has not been taken.

General Methods to Detect New Genes

Early findings

New genes were initially found by serendipity rather than intentional searches. For instance, the first novel gene, *Jingwei*, was identified in 1993 based on previous studies which had considered it to be a processed pseudogene in *D. yakuba* (Long and Langley 1993). Several more young *Drosophila* genes, for example *Adh-Finnegan*, *Sdic*, and *Exu2*, were accidentally found four to five years later. *Adh-Finnegan* was initially claimed to be an *Adh* pseudogene, but later analyses showed it was a functional gene recently descended from an *Adh* duplication (Begun 1997). The discovery of *Sdic* was built solely on an earlier observation that the genetic organization of the 19DE region on the *D. melanogaster* X chromosome differs from that of other species in the subgroup (Nurminsky et al. 1998). The gene *Exuperantia2* was detected because of a complete linkage disequilibrium between two single nucleotide polymorphisms in *D. pseudoobscura* (Yi and Charlesworth 2003).

Comparative molecular cytogenetic analyses

Phylogenetic comparison of genetic signals (e.g., fluorescence in situ hybridization [FISH] and genomic Southern blotting), has proved to be an efficient and reliable way of identifying young protein-coding genes in *Drosophila* and mammals. Wang and colleagues (2002 and 2004) used FISH to systematically search for new genes in *Drosophila* species, taking advantage of publicly available cDNA collections (Berkeley Drosophila Genome Project, www.fruitfly.org). They amplified and labeled the cDNA inserts and then hybridized them to polytene chromosomes of each member of the *D. melanogaster* subgroup. By counting hybridization signals on the polytene chromosomes of theses species, new homologs translocated to different cytological loci were detected. This approach detected about 100 new duplicates across eight species of the *D. melanogaster* subgroup, and three of these have been fully described: *Sphinx*, *synathase chain F*, and *Monkey King* (see Table 3.1).

Despite this success, the limitations of FISH screening are also obvious. First, FISH (or genomic Southern blotting) is technically demanding and laborious. Second, many organisms do not have polytene chromosomes. Third, even in the case of the *Drosophila* genus, which has polytene chromosomes, FISH cannot be used to detect new genes in heterochromatic regions, because polytene chromosomes do not include heterochromatin. Finally, FISH cannot resolve tandem duplications, where the duplicates are adjacent.

Computational genomic analysis

Extensive comparative sequencing and expression studies, coupled with evolutionary analyses and simulations, have been applied in several organisms to identify gene duplication events at the whole genome level. Completed

genome sequences in model organisms provide opportunities to search for duplicated genes and further examine the pattern of gene origination.

Betran and coworkers (2002) surveyed the whole *D. melanogaster* genome to search for new retroposed genes. They inferred parental and derived copies by examining potential retroposed genes for hallmarks of the retroposition process (see previous discussion). At a threshold of more than 70 percent protein sequence identity, they identified 24 retroposition events, all within the last 30 million years. They further reported a new gene in the *D. melanogaster* subgroup, *Drosophila nuclear transport factor-2-related* (*Dntf-2r*). Its sequence and phylogenetic distribution indicate that *Dntf-2r* is a functional retroposed gene that originated in the common ancestor of *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana* within the past three to five million years and is under positive Darwinian selection.

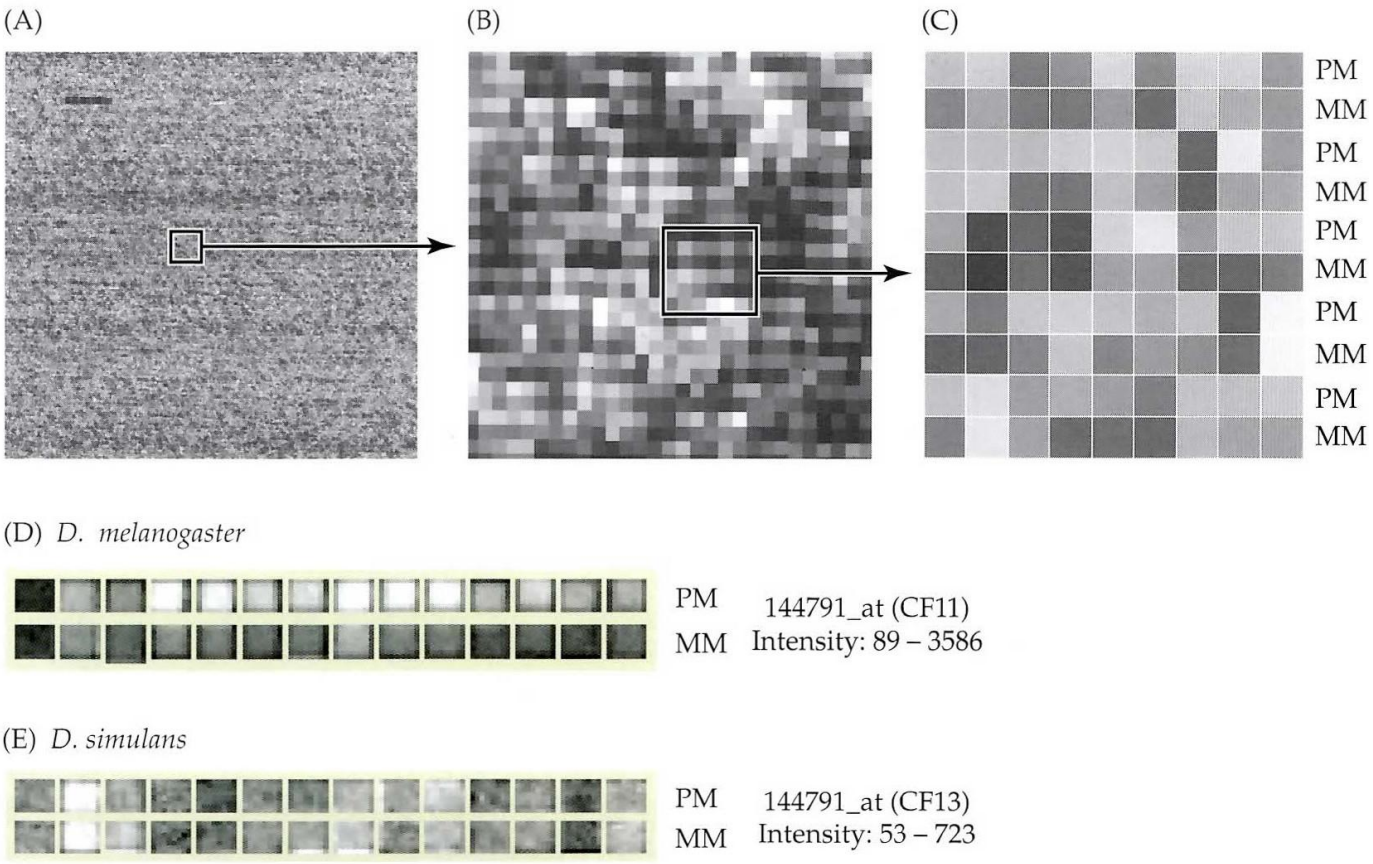
Marques and colleagues (2005) systematically screened the human genome for retrogenes by comparing the genome sequences of human, chimpanzee, and mouse. They identified 57 retrogenes in the human genome, estimating that one retrogene per million years has emerged on the primate lineage leading to humans. Comparative sequence and gene expression analyses suggest that a significant proportion of recent retrocopies represent human-specific genes. They concluded that retroposition significantly contributed to the formation of recent human genes and that most new retrogenes were progressively recruited during primate evolution by natural and/or sexual selection to enhance male germ line function.

Zhang and coworkers (2005) identified 69 retroposons in the *Arabidopsis thaliana* genome. Most of them were derivatives of mature mRNAs. Of them, 22 are processed pseudogenes and 52 genes are likely to be actively transcribed, especially in tissues from roots and flower apical meristems. This study estimated the rate of new gene creation by retroposition as 0.6 genes per million years. Forty-five of the parental genes were highly expressed in the germ line cells, which presumably predisposes them to be templates for retroposition.

Genomic computational searching also has noticeable restrictions. First, this method is limited to model organisms whose genomes have been sequenced. Second, although retrogenes can be found using this method, it is less useful for finding other types of gene duplication, such as exon shuffling and gene fusion. Finally, such a screening method will often miss new duplicates in genomes sequenced by the whole genome shotgun approach (e.g., International Chicken Genome Sequencing Consortium 2004). The time of gene duplication events can be estimated by sequence divergence analyses (e.g., sequence identity or the synonymous substitution rate K_s) or from the phylogenetic distribution of the duplication.

Comparative Genomic Hybridization to Detect New Genes

Microarrays are a developing technology used to study gene expression at the whole genome level. Single-stranded DNA (ssDNA), referred to as *probe*,



Signal intensity comparison between *D. melanogaster* and *D. simulans*

Figure 3.1 An example of images from an Affymetrix GeneChip hybridization experiment with *D. melanogaster* genomic DNA. The intensity of hybridization signal is characterized by the ratio of black and white: the darker the color, the more intense the hybridization signal (the more labeled DNA fragment was bound). (A–C) Images at three different resolutions of a GeneChip after hybridization with labeled genomic DNA. (D) Hybridization intensity of 14 probe pairs from feature 144791 hybridized with *D. melanogaster* genomic DNA. (E) Hybridization intensity of 14 probe pairs from feature 144791 hybridized with *D. simulans* genomic DNA. Abbreviations: PM, perfect match; MM, mismatch. (Fan and Long, unpublished data.)

is printed in a regular grid-like pattern. The target RNA or DNA from a particular biological sample is fluorescently labeled and allowed to hybridize to the array. Depending on the specific experimental design, the intensity of each spot or the average intensity difference between matches and mismatches can be related to variation in gene expression (mRNA abundance), DNA polymorphisms, or mutations caused by changes in copy number in whole genome samples (Figure 3.1; Pinkel et al. 1998; Barrett et al. 2004; Greshock et al. 2004; Toruner et al. 2007). In an effort to develop a more generally useful method to detect new gene candidates, we have adapted this technology to detect the variation in duplicate copies (gene gain or loss) in closely related species by hybridization.

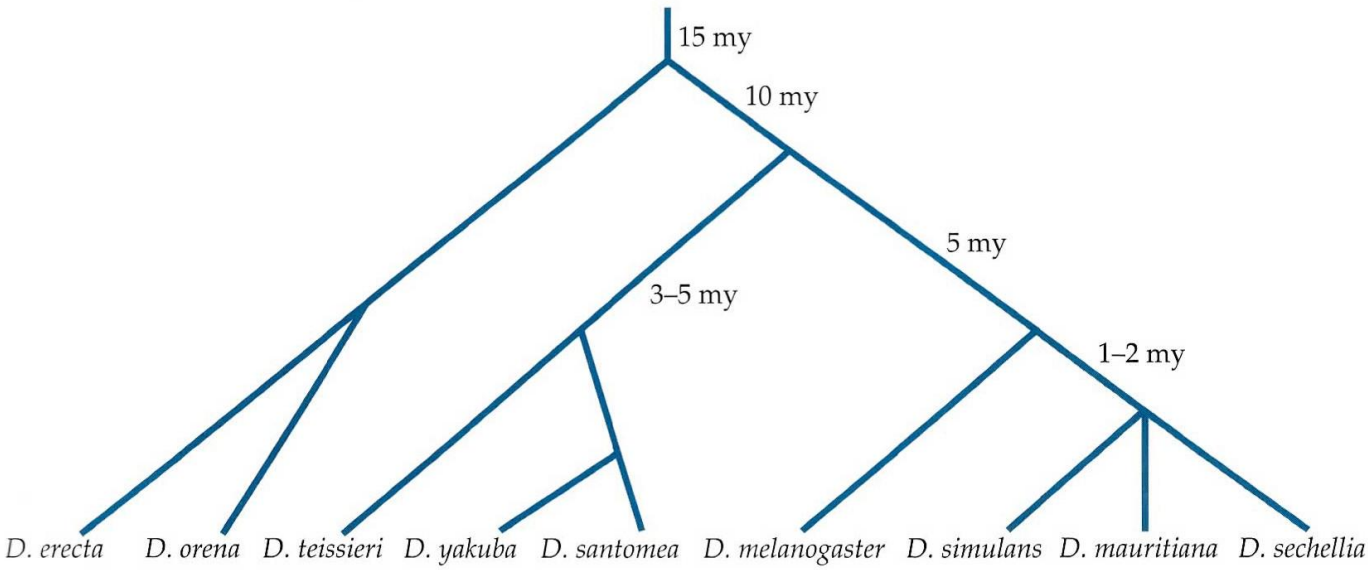


Figure 3.2 Phylogeny of the *D. melanogaster* subgroup with estimates of divergence times.

The availability of genomic sequences and GeneChip arrays for *D. melanogaster* has allowed us to systematically search for new genes using array-based comparative genomic hybridization (CGH). The subgroup of *D. melanogaster* includes nine species with a divergence time of less than 15 million years (Figure 3.2; Lachaise et al. 1988; Lachaise et al. 2000). The relatively small genome size of *D. melanogaster* and its well-defined phylogeny provide a convenient context to find young genes. Because the GeneChip arrays were made using genomic sequence data from *D. melanogaster*, the initial data analyses were conducted using *D. melanogaster* as a baseline to calculate the ratio of hybridization intensity for each gene between *D. melanogaster* and the other species. Considering the sequence divergence between different species, we took a ratio of 1.5 or higher as the threshold for gene duplication. This was based on an initial calibration using a known duplicated region (with 31 genes) in *D. melanogaster*, which yielded a ratio distribution for duplicates of about 1.3–1.5.

Examination of the microarray intensity ratios of pairwise comparisons allowed us to identify candidate duplicates with ratios of 1.5 or higher in these species. Next, we applied genomic Southern hybridization and BLAST searching of genomic sequence data of *D. simulans* and *D. yakuba* to confirm the duplicate copies and survey their phylogenetic distribution. Examples of candidate duplicates identified by CHG are given for *D. simulans* and related sibling species in Table 3.2.

A specific example of a new gene identified by array-based CHG is *infertile crescent-2h* (*Ifc-2h*), derived from its parental gene *infertile crescent* (*Ifc*). The hybridization ratios of *D. simulans* (1.74), *D. mauritiana* (1.73), and *D. sechellia* (1.56) to *D. melanogaster* were higher than those of other species (1.15–1.49), suggesting that the former group of species had a new gene copy. This was supported by a BLAST comparison of the *D. melanogaster Ifc*

Table 3.2 Examples of new duplicates identified by array-based CGH in the three sibling species *D. simulans*, *D. mauritiana*, and *D. sechellia*

GeneChip ID	Parental copy	Location of parental copy[a]	Copy number in *D. simulans*	Copy number in *D. melanogaster* and *D. yakuba*	Location of new duplicate[a]
153965	CG12081	X	2	1	2h
143838	*Rpl9*	2L	2	1	2R
153842	*Ifc*	2L	2	1	2h
153010	*Chmp1*	3L	2	1	X
145353	CG7914	X	2	1	2L

[a]Chromosomal location is given for arms (L = left, R = right) and heterochromatin (h).

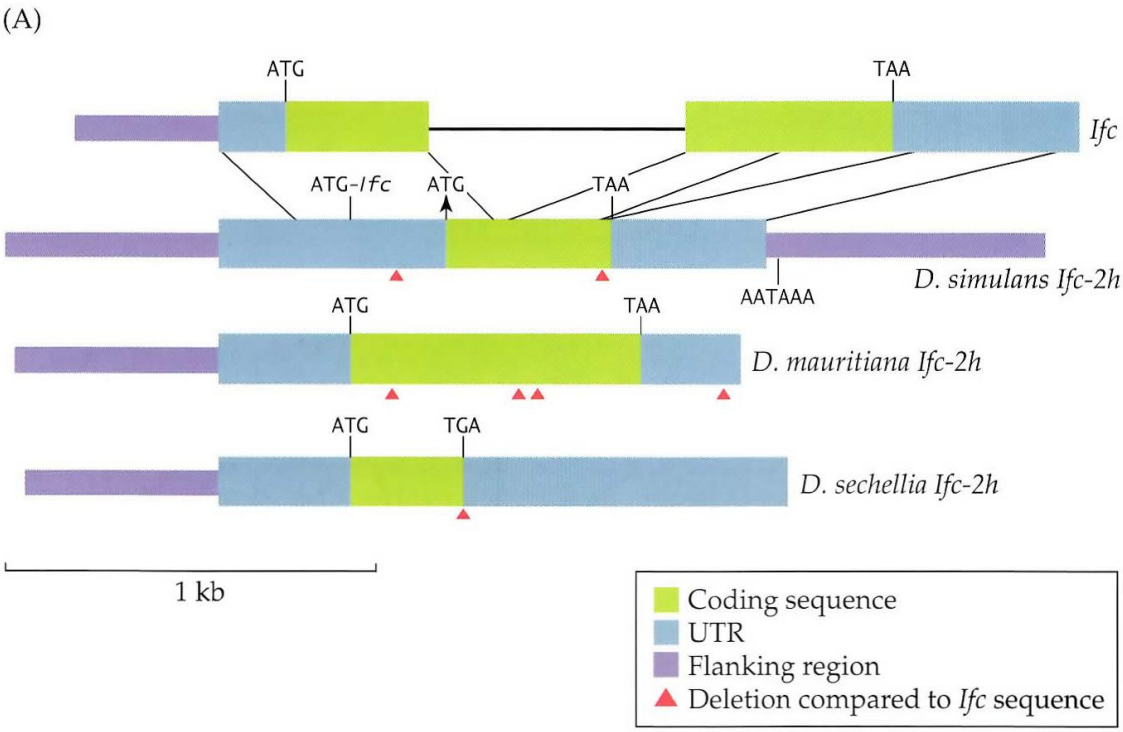
sequence to genomic data of *D. simulans*, which revealed two sequences homologous to *Ifc*. One copy had a single intron and is therefore the parental copy; the other lacked an intron and is likely to have been derived via retrotransposition (Figure 3.3A). A Southern hybridization using *Hin*dIII digested DNA confirmed the BLAST result. Only single bands exist in *D. melanogaster*, *D. teissieri*, *D. santomea*, and *D. erecta*. Two bands are found in *D. simulans*, *D. mauritiana*, and *D. sechellia*. Two bands are also found in *D. yakuba* but are generated by a unique *Hin*dIII digestion site in the intron of *Ifc*, so *D. yakuba* actually has only one copy (Figure 3.3B).

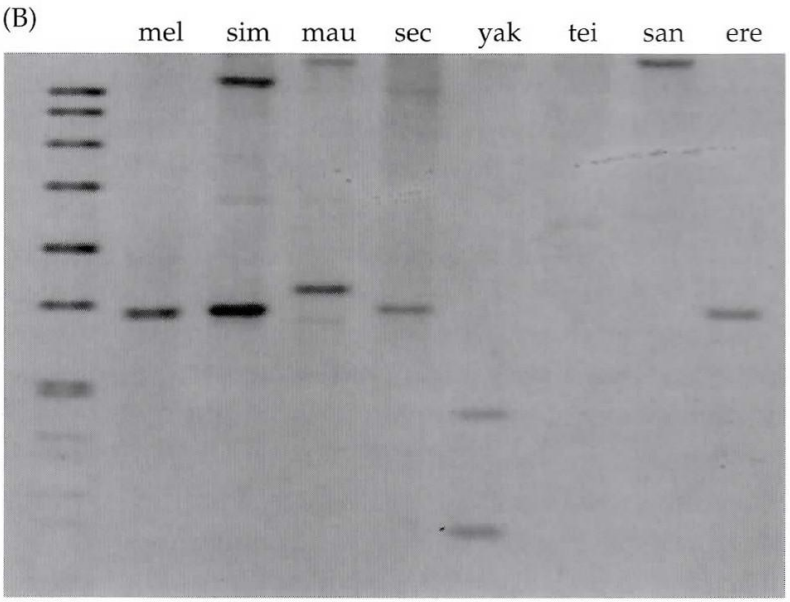
Sequence analysis of *Ifc-2h* shows 12 indels—four in coding and eight in noncoding flanking and UTR regions (Figure 3.4). The lengths of all four coding region indels are in multiples of three. This pattern is significantly different ($P = 0.0123$) from the random distribution that occurs in noncoding regions, revealing evolutionary constraint to maintain a nondisrupted reading frame in the coding region of the new gene copy. However, the *D. sechellia* copy is probably degenerating to a pseudogene, because a premature nonsense mutation in its coding region drastically shortens the reading frame.

The expression profile of *Ifc-2h* was examined by RT-PCR in *D. simulans* (Figure 3.3C). The results show that *Ifc-2h* is highly transcribed in eggs, second larvae, and adults, but the transcription in third larvae and pupae is relatively low. This expression pattern differs from the parental copy, *Ifc*, which is ubiquitously transcribed at all developmental stages. All the analyses described above indicate that *Ifc-2h* is a functional protein-coding gene (Fan and Long 2007).

Challenges in using array-based CGH in new gene studies

Array-based CGH is a potentially powerful tool to detect duplication events in different species. However, there are a number of challenges to this approach. In particular, sequence divergence between species limits hybridization of heterospecific DNA. We noticed that a one percent sequence


(A)
ATG
TAA
Ifc
ATG-Ifc
ATG
TAA
AATAAA
D. simulans Ifc-2h
ATG
TAA
D. mauritiana Ifc-2h
ATG
TGA
D. sechellia Ifc-2h
1 kb
Coding sequence
UTR
Flanking region
Deletion compared to Ifc sequence


(B)
mel
sim
mau
sec
yak
tei
san
ere

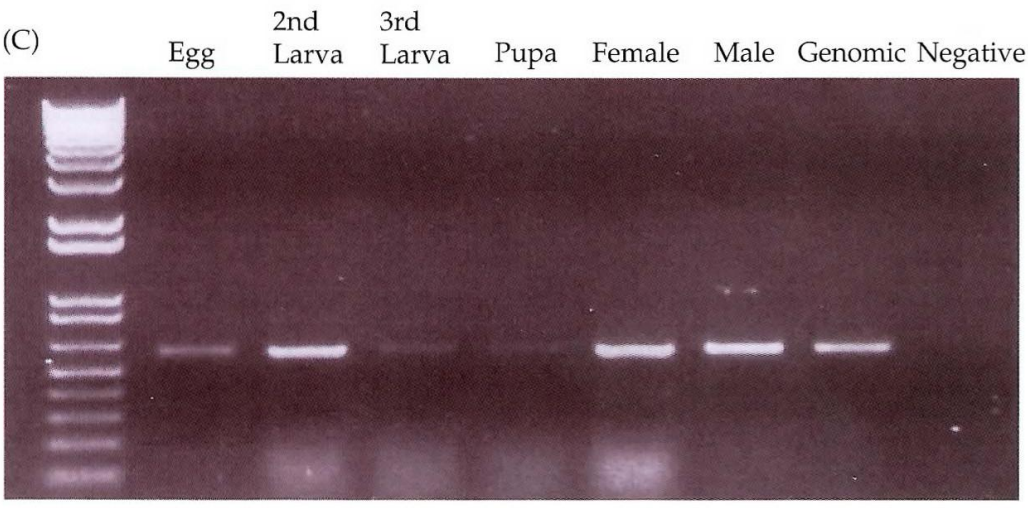

(C)
Egg
2nd Larva
3rd Larva
Pupa
Female
Male
Genomic
Negative

Figure 3.3 Schematic diagram of the gene structure of *Ifc* and *Ifc-2h*. (A) Start and stop codons and an adenylation signal are shown. (B) Genomic Southern blotting using a probe for *Ifc-2h* against *Hin*dIII digested DNA. Species names are shown at top of each lane (mel, *D. melanogaster*; sim, *D. simulans*; mau, *D. mauritiana*; sech, *D. sechellia*; yak, *D. yakuba*; tei, *D. teissieri*; san, *D. santomea*; ere, *D. erecta*). Note that two bands are seen in *D. simulans*, *D. sechellia*, *D. mauritiana* and *D. yakuba*. (C) RT-PCR for *Ifc-2h* transcripts in *D. simulans* at various developmental stages; similar patterns are seen in *D. sechellia* and *D. mauritiana*.

divergence usually accounts for a five to eight percent signal reduction in *Drosophila* genomic hybridization (Table 3.3). If the species is too distantly related (sequence divergence > ten percent), then many probes do not generate adequate signals from hybridization.

A related problem is sequence divergence between paralogs. Signals from duplicate copies are subject to a considerable range of variation. Furthermore, paralogous duplicates in the related species are associated with sequence divergence correlated both with the time of the duplication event and the degree of functional constraint on the new gene. New genes tend to evolve at an accelerated rate early after origination, thus lessening the signal intensity considerably and potentially limiting the ability to detect new genes.

Finally, we note that microarrays are a complementary technology to other tools in the arsenal of the molecular biologist. In the search for new genes, microarrays provide candidates that need to be verified by other molecular techniques (e.g., FISH and genomic Southern blotting, detailed sequences analyses, and expression profiles).

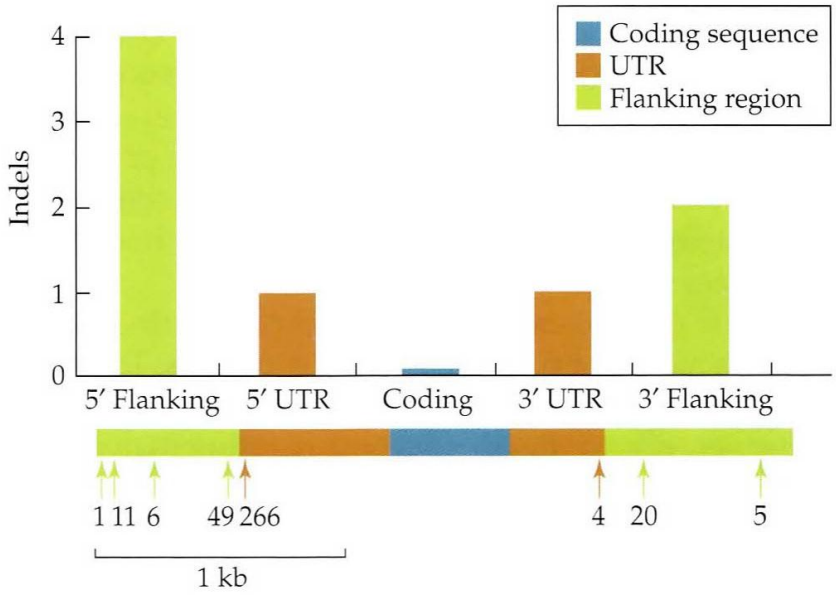


Figure 3.4 *Ifc-2h* polymorphic insertions and deletions (indels) found in *D. simulans* population sequences. Arrows indicate the relative positions of indels. The number below each arrow indicates indel size (bp).

Table 3.3 Sequence divergence between species and ability to detect related sequences by microarray hybridization

Species	Sequence divergence (%) from *D. melanogaster* (Canton-S)	Replicates	Detection by microarray (%)	Average intensity (three replicates)
D. melanogaster (Oregon-R)	0.5	3	96.6	522.4
D. simulans	4	3	77.9	532.2
D. mauritiana	4	3	79.0	536.0
D. sechellia	4	3	73.5	544.0
D. yakuba	7	3	47.0	562.0
D. teissieri	7	3	47.5	571.3
D. santomea	7	3	45.9	563.0
D. erecta	10	3	47.9	546.7

Outlooks and Perspectives

The power of searching for new gene candidates by comparative genomic screening is increasing with the expansion of genome sequence data from different species. For example, genome sequences have now been completed for twelve closely related *Drosophila* species; comparisons among these entire genomes will provide solid evidence of young gene candidates for each species and/or clade within this group. Such comparisons will also provide whole genome evidence for how often novel genes have arisen. We will also be able to see whether novel genes are associated with speciation and adaptation after new species evolved. Moreover, the role of selection in the retention of new gene copies can be fully characterized.

One of the major challenges for novel gene studies is to determine whether novel genes produce functional proteins. Fortunately, advances in global-scale analysis of proteins are expected to allow direct observation of protein function and regulation (Sauer et al. 2005). Endogenous proteins can be identified by two-dimensional gel electrophoresis and characterized using mass spectrometry. Further, we can examine the function of uncharacterized proteins using protein–protein interaction data produced by affinity-based proteomics (protein arrays) and other proteomic methods, such as yeast two-hybrid analyses (Y2H).

The survival value of new genes is determined by their networks of interactions with other genes, which give rise to biological processes. To understand how natural selection leads to the retention or loss of new genes requires placing the new genes' activities in this context, rather than simply inferring properties from the rates of gene sequence evolution (e.g., K_a/K_s ratios). Interesting questions to investigate include: (1) how gene–gene inter-

actions evolve in a new gene copy, and (2) how these interactions affect biological functions.

One of the ultimate goals of gene origination studies is to uncover general patterns for gene origination and evolution, and to further understand how this contributes to organismal evolution. Using applied, array-based CGH, combined with molecular biological tools and computational comparisons, we can identify young genes with known functions in sufficient numbers to learn about new gene evolution at the genomic level, and, further, to precisely measure the origination rate of new gene functions. This achievement will show how quickly organisms adapt by changes in gene diversity, and to what degree this is correlated with environmental change.