Volff J-N (ed): Gene and Protein Evolution.
Genome Dyn. Basel, Karger, 2007, vol 3, pp 131–146

Origination of Chimeric Genes through DNA-Level Recombination

J.R. Arguello[a], *C. Fan*[a], *W. Wang*[b], *M. Long*[a]

[a]The University of Chicago, Department of Ecology and Evolution, Zoology 301E, Chicago, Ill., USA; [b]CAS-Max Planck Junior Research Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China

Abstract

Comparative genomics is rapidly bringing to light the manifold differences that exist within and between species on the molecular level. Of fundamental interest are the absolute and relative amounts of the genome dedicated to protein coding regions. Results thus far have shown surprising variation on both the polymorphism and divergence levels. As a result, there has been an increase in efforts aimed to characterize the underlying genetic mechanisms and evolutionary forces that continue to alter genomic architecture. In this review we discuss the formation of chimeric genes generated at the DNA level. While the formation of chimeric genes has been shown to be an important way in which coding regions of the genome evolve, many of the detailed studies have been limited to chimeric genes formed through retroposition events (through an RNA intermediate step). Here we provide a short review of the reported mechanisms that have been identified for chimeric gene formations, excluding retroposition-related cases, and discuss several of the evolutionary analyses carried out on them. We emphasize the utility chimeric genes provide for the study of novel function. We also emphasize the importance of studying chimeric genes that are evolutionarily young.

Fundamental to an understanding of biological diversity are questions pertaining to the acquisition of new functions. As a result, the origin of evolutionary novelty has long captured the imagination of evolutionary biologists and naturalists. On the molecular scale, chimeric genes – by which we are referring to gene structures (both coding and noncoding) that have been derived from multiple parental loci – provide a useful system for studying the origin of new functions. The first reason for this is that chimeric genes are unlikely to have

redundant functions. This is in contrast to classical views of gene duplication in which theoretical models assume that an exact duplicate copy possessing redundant functions is initially produced [1, 2]. The second reason is that recent findings suggest chimeric proteins may arise at unexpectedly high rates [3–8], and, because of this, there is particular excitement over a growing number of reports on evolutionarily young cases. The identification of young cases represents an important advance for the field because it provides the opportunity to examine the early evolutionary steps involved in the acquisition of new functions. The third reason is that chimeric structures aid in the resolution of ancestor-offspring relationships. The ability to distinguish the derived copy from the ancestral copy becomes a substantial issue when trying to understand evolutionary forces that may differ between paralogs.

Considerable insight into the evolution of chimeric genes has recently been gained through the study of retroposed genes. Both genome-wide analyses and individual case studies have demonstrated that numerous new genes with novel functions have been generated as a result of initial retropositions [3, 4, 7, 9–14]. However, less is known about the origin of chimeric genes through recombination events at the DNA level. In this chapter our goal is to review data regarding this less understood phenomenon. In doing so we will omit instances of chimeric formations that have occurred through an RNA intermediate step, but for those interested we refer to several reviews [15–17]. This chapter is further narrowed by our treatment of sequence-level studies and not biochemical and molecular biology studies. Both are relevant, and have made major contributions to our understanding of the underlying genetic mechanisms, but they are beyond the scope of this chapter.

We have organized this review into two sections. We first introduce the genetic mechanisms that have been shown to be capable of generating chimeric proteins on the DNA level. Next, we focus on illustrious examples that provide evidence for each of these mechanisms. Nonhomologous recombination (NHR) will comprise the majority of this second section, and we have divided these data, and the corresponding experimental approaches, into ‘evolutionarily ancient’ and ‘evolutionarily recent’ examples. In the ‘evolutionarily recent’ section we present a chimeric protein, *Hun*, which our group recently reported on [18]. Throughout the chapter, we highlight the utility of young chimeric proteins as a system for studying novel functions.

Molecular Mechanisms Leading to Chimeric Genes

At the level of DNA, several molecular mechanisms have been observed to recombine different genic and nongenic regions to generate chimeric genes

structures. Nonhomologous recombination was the mechanism first proposed [19]. In this early model, known as exon-shuffling, NHR within introns leads to novel combinations of exonic regions. More recently, models for chimeric gene origination have become more numerous and more flexible as the number of recognized mechanisms giving rise to them continues to increase.

Nonhomologous Recombination

NHR, also called illegitimate recombination, refers to recombination events that occur without reliance on extensive stretches of identity, and instead occur between regions with little or no identity [20–22]. Both NHR and homologous recombination (HR) pathways occur efficiently, and likely in an overlapping manner, to repair double stranded breaks (DSB) [23–25]. HR usually results in accurate strand repair [23, 25, 26] while NHR results in imperfect repair, causing duplications, insertions, and deletions. The involvement of gene regions in any of these latter events could potentially result in novel chimeric genes.

Non-allelic Homologous Recombination

When strand repair occurs through HR, or if there is mispairing during synapsis, there is the potential for different low copy repeats (LCRs) to recombine [27, 28]. This is referred to as non-allelic homologous recombination (NAHR). LCRs are generally short blocks (1–400 kb) of duplicated DNA which share considerable sequence identity [28–30]. NAHR events result in numerous rearrangements including duplications, inversions, deletions and translocations (fig. 1). Similar to NHR, if the breakpoints of these events involve gene regions, chimeric genes may result.

Transposable Elements as 'Fragment Joiners'

Transposable elements (TEs), being a type of LCR, can be involved in the origin of chimeric genes through the facilitation of NAHR (above). However, they also have the capability of recombining short DNA sequences through their inherent biochemical processes. Currently, this has been recognized with two plant TEs, Pack-MULEs and *Helitrons*. Though the mechanism is still uncertain, Pack-MULEs can recruit small chromosome fragments and combine with other genomic regions through their own movements to form chimeric gene structures [31]. It is thought that ectopic gene-conversion across a nicked cruciform structure may play a role in this recruitment process [32] (fig. 2). *Helitrons*, which are helicase-bearing transposable elements, are likewise capable of shuffling genomic regions. Again, the mechanism has not been worked out for these TEs, but they are capable of transporting replicated elements to target sequence and replacing it with its own DNA. This has been referred to as a rolling circle transposition mechanism [33, 34].

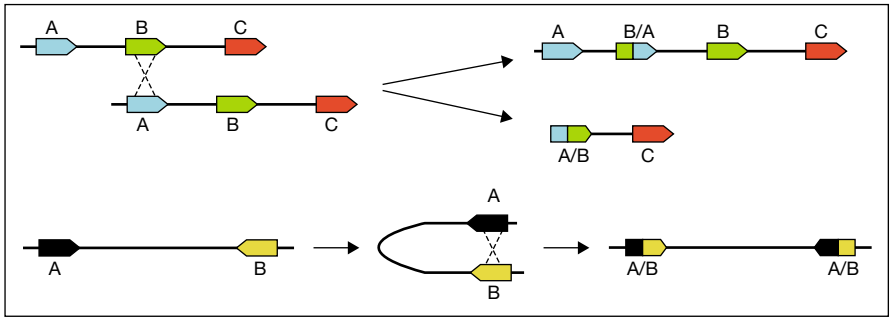

Fig. 1. Models representing non-allelic homologous recombination. The top panel depicts a recombination event between low-copy repeats (pointed boxes A–C). The products of this event are a duplicated (top right) and a deficient (bottom right) chromosome region. Note that the recombination event can occur between homologous or nonhomologous chromosomes. If the event occurs between nonhomologous chromosomes, a translocation would be produced. The bottom panel depicts a recombination event between two low-copy repeats that exist on the same chromosome, but results in an inverted configuration. Such non-allelic recombination events can occur through hairpin structures.

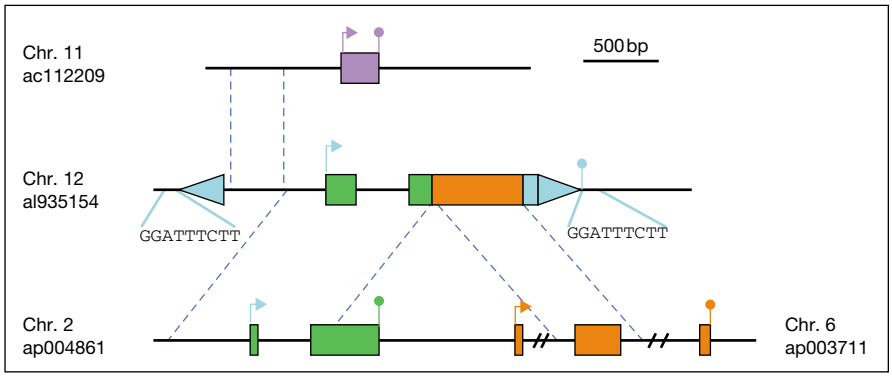

Fig. 2. Chimeric gene formation involving a Pack-MULE. The model provides an example of a novel chimeric gene (al935154 on chr. 12) that was created by a Pack-MULE containing gene fragments from three genomic loci, including both introns and exons (ac112209 at chr.11, ap004861 at chr. 2, and a putative bHLH transcription factor at chr. 6). The tandem inverted repeat is noted by the purple sequence and the start and stop codons are marked for each gene with an arrow and dot, respectively. Homologous fragments are indicated by dash lines (figure modified from [31]).

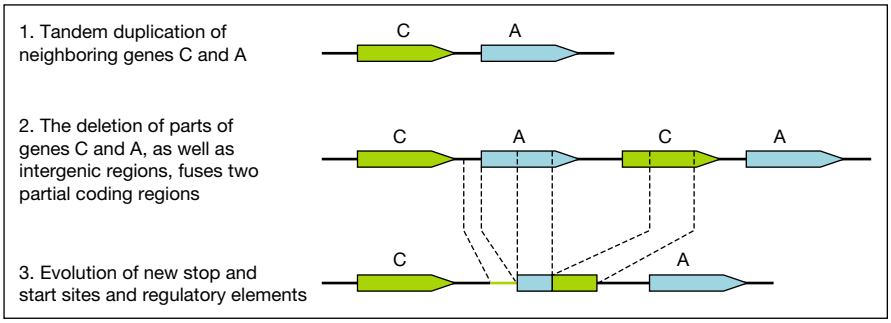

Fig. 3. Example of a chimeric gene formed by gene fusion. This model is a simplified version of that found in [35]. In it, a gene pair (C and A) is duplicated in tandem. The duplication is followed by deletions that combine the remaining exonic regions of the two middle genes (A and C). Later evolutionary events include the recruitment of regulatory elements and the establishment of new start and stop codons if they were deleted.

Gene Fusions

It has been shown that two distinct genes or gene regions in adjacent genomic positions can be fused to form a single chimeric gene. A requirement for generating a chimeric gene in this way is to delete or skip the stop codon in the upstream gene. In prokaryotic genomes, the stop codon can be eliminated through nucleotide insertions, as shown in a fusion experiment using *E. coli* tryptophan synthetase alpha and beta polypeptides [35]. In eukaryotic genomes, two naturally occurring examples of gene fusion events have been identified and have been shown to be the result of two different molecular processes. One instance resulted from multiple deletions in a region between two neighboring genes, while the other resulted from unusual alternative splicing across two genes ([36, 37]; fig. 3 and below).

Evidence for Chimeric Proteins

Nonhomologous Recombination: Ancient Chimeras

In detecting ancient chimeric genes, two general methods have been used. The first is protein sequence comparisons in which similarity searches are carried out between protein regions from different protein families. Second, statistical analyses have been used to detect signals left over from the shuffling processes, for example the phase of introns (the positions of introns within and between codons; see below).

The first documented chimeric genes created by exon shuffling were the human tissue plasminogen activator (TPA) [38, 39] and the low-density lipoprotein (LDL) receptor protein in humans [39, 40]. TPA, which is necessary for the conversion of plasminogen into its fibrin dissolving active form, was found to be composed of domains which share significant similarity with uroinase, epidermal growth factor, and fibronectin domains. The LDL receptor, which is a cell-surface protein that mediates the endocytosis of low density lipoprotein (LDL), has eighteen exons. Thirteen of the eighteen exons share significant similarity with other proteins such as the C9 component factor, the EGF precursor, and blood clotting factors (IX, X, and C). Following these studies, Patthy did intensive sequence comparisons which revealed more than 300 gene families that contain mosaic domain structures [41–44]. Included in Patthy's dataset are proteins found in the coagulation cascade of mammals and fish, indicating an ancient generation dating back approximately 450 million years ago [41, 42]. Also notable is the fact that many intracellular proteins involved in signaling pathways are present in humans, worm, and yeast which share a remote common ancestor (1.46 BYA) [42, 43].

While intriguing, the chimeric structures alone are not sufficient evidence for exon shuffling as originally proposed, in which introns play an essential role in the recombination process [19]. Additional support for the role of exon shuffling in the origination of these chimeric genes comes from further analyses demonstrating that the recruited domains are flanked by introns of identical phases (1, 1) (i.e. symmetric exons). This is a hallmark of exon shuffling because it maintains the reading frames of the new gene [44].

In both plants and animals, kinases provide canonical examples of exon shuffling, supported by both sequence identity and phase data [41]. These data indicate that exon shuffling events through DNA level recombination took place approximately 990 million years ago according to the estimate of divergence time for these organisms [45]. In plants, the receptor kinases possess functions equivalent to animal receptor tyrosine kinases [41], but it is unclear whether the low sequence similarity is due to independent originations or the long evolutionary time that separates animals and plants (1.58 billion years). Another insightful instance of exon shuffling is found in the origination of the cytochrome c1 precursor gene in potato. Cytochrome c1 is part of the mitochondrial respiratory chain and is found in most eukaryotes. In potato, it was found that the mitochondria-derived nuclear gene, cytochrom c1, recruited a target domain from GapC. This shuffling event resulted in a mitochondrial targeting function for the new protein [46].

Taking advantage of the available sequenced genomes, as well as recent methods used to define protein domains, several groups have taken a genome-wide approach to investigating the relationships between domains and introns. In a human genome study, Kaessman et al. [47] analyzed the effect of the position

of introns, within or outside domain boundaries, on the distribution of intron phases. They observed that the introns within the domain boundaries show significant intron phase correlations (for example the excess of (1,1) symmetric exons), however, the introns outside the boundaries (or within domains) do not show such correlations. These results reveal a role for exon shuffling in recombining protein domains to form chimeric genes. Similarly, in a study that included the genomes of human, mouse, rat, *Fugu*, zebrafish, *Drosophila*, mosquito, *C. briggsae*, and *C. elegans*, Liu and Grigoriev [48] found a significant correlation between exon borders and protein domain borders. Interestingly, the significance of this correlation increased as they moved from *C. briggsae* to humans. Moreover, they also found that most of the exon-correlating domains were bordered by symmetric introns, with 1-1 introns being the most frequent.

As a cautionary note we would like to point out that it has been shown that an intron-containing gene structure can also be transposed into a new genomic position by retroposition of the gene's antisense RNAs [5]. This can potentially recombine with preexisting exon-intron structures to form a new chimeric gene. Given recent findings that show a high proportion of genes in mammals and invertebrates have antisense RNAs [49, 50], an alternative retroposition model for some of the examples above cannot be discounted. To be able to exclude competing models of chimeric gene formation, recently evolved chimeric genes are often necessary.

Nonhomologous Recombination – Evolutionarily Recent Chimeras

Not all of the NHR events that result in chimeric gene structures fit within the exon-shuffling model. This has been most clearly demonstrated through studies of young cases in which, among other means, chimeric genes have formed through recombination events within exons as well as through the recruitment of previously nongenic DNA. Studies of recent NHR events leading to chimeric proteins primarily utilize breakpoint data obtained from case studies of naturally occurring NHR events [28–30, 51–54] or transfection-based experimental approaches [20–22, 24, 55–57]. Both methods present particular benefits and limitations. While transfection approaches are capable of generating considerable breakpoint data in a relatively short amount of time, the finite number of constructs used in transfections is a narrow representation of what occurs biologically. To the approach's credit, however, these experiments have been instrumental because studies of naturally occurring NHR events have been limited by the difficulty involved in identifying them. Perhaps unsurprisingly, prior to the availability of large genomic datasets, disease phenotypes led to the identification of many NHR events. As a result, most of our current knowledge regarding naturally occurring examples is disease-related. Though the disease cases are more interesting medically, their investigations provide

pertinent information for protein evolution; each additional reported rearrangement sheds light on the spectrum of possible mutational mechanisms. Fortunately, the large increase in whole genome sequence data is quickly lowering this identification hurdle. Whole genome comparisons, both within and between species, are enabling the identification of numerous rearrangements on which to carry out detailed sequence analyses and further experiments [58–60].

Through the combined efforts of transfection experiments and the study of disease-related NHR events, as well as molecular and biochemical approaches, a collection of motifs and elements that are commonly enriched at or near NHR breakpoints have been identified [30, 52, 56, 57, 61–69]. Several of these are known to be recognized by particular enzymes (Topo I and DNA polymerase, for example [56, 57, 62, 66, 69, 70]) or are thought to lead to DSB-prone structural changes in chromatin [30, 52, 66, 68]. The remaining motifs may also be prone to DSB or may possess enzymatic signals yet to be elucidated. A major goal of these studies is to connect sequence-based data with molecular genetic pathways so that a fuller mechanistic understanding of NHR events can be gained. These identified motifs provide useful sequences to search for when examining rearrangement breakpoints.

A Recent Example: Insights from Hun, a Young Gene Generated by NHR

We recently reported on a young chimeric gene, *Hun,* which was generated by NHR and fixed in the common ancestor of *D. simulans*, *D. mauritiana*, and *D. sechellia* [18]. Its finding provided the unique opportunity to observe the early stages of a new chimeric protein generated by DNA-level recombination. The characterization of its structure, expression, and evolutionary genetics has cast light on the role of nonhomologous recombination in the duplication of a sequence at an ectopic position.

In a large-scale effort to identify new genes, a combination of fluorescent in situ hybridizations (FISH), Southern hybridizations and BLAST techniques were applied across the *D. melanogaster* subgroup. *Hun* was identified as a partial duplicate of the *Bällchen* gene (1867 bp), which arose in the common ancestor of *D. simulans*, *D. sechellia*, and *D. mauritiana*, 2–3 mya. It was shown that the *Hun* duplication involved a t(3R;X) translocation. The total amount of DNA translocated to the X chromosome was ~1520 bp. The derived *Bällchen* coding region was truncated at the 3′ end by ~412 bp, and included only ~65 bp 5′ of the original start codon. Subsequently, ~99 bp were recruited into the 5′ coding region, with an additional ~167 bp to the polyadenylation site. In addition, ~400 bp 5′ of the paralogous *Bällchen* start site were recruited (Fig. 4a). Interestingly, this 5′ UTR contains a putative intron of 49 bp that appears to have evolved de novo.

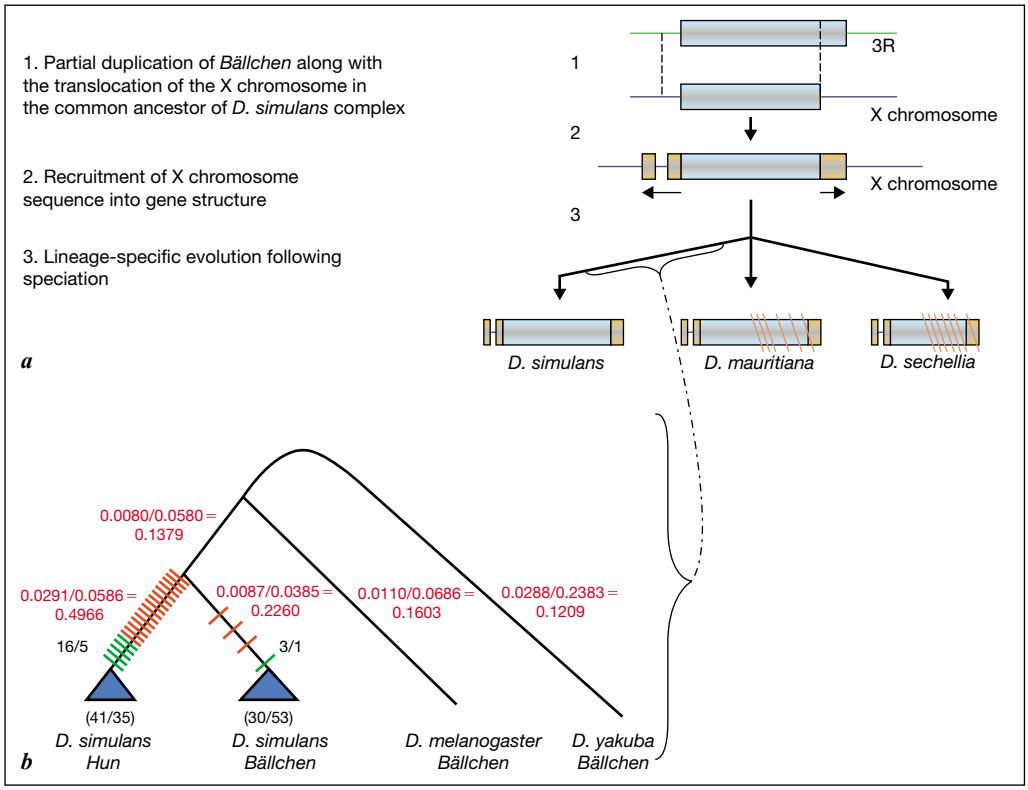

Fig. 4. Models depicting the evolution and population genetics of *Hun*. ***a*** A simplified 3-step model for the origin of *Hun*. Blue bars represent coding regions, brown bars represent UTR regions, and red dashes represent premature stop codons. ***b*** A gene tree for *D. simulans' Hun* and *D. simulans, D. melanogaster,* and *D. yakuba's Bällchen.* The tree displays measurements of divergence as measured by Ka/Ks (red ratios), nonsynonymous and synonymous fixations found along the *Hun* and *Bällchen* branches depicted by colored bars (red represents nonsynonymous changes and green represents synonymous changes, black ratio), and polymorphism found in the *D. simulans* population data (black ratios below triangles, nonsynonymous/synonymous). Taken from [18].

To investigate the mechanism that led to the duplication and translocation of *Hun,* its flanking sequence was examined. In particular, having ruled out the role of an RNA intermediate due to the maintained intron within the coding region, and lack of a poly-A tract, there was interest in the possibility that LCRs may have aided in the rearrangement. No evidence for transposable elements existed near the regions where identity is lost between *Hun* and *Bällchen*. In addition, no evidence for direct repeats was found. When *Bällchen* and

Hun's flanking regions are aligned, only short spurious stretches of identity exist. This lack of evidence for an intermediate RNA step, LCRs, or any other significant sequence identity led to the conclusion that *Hun* originated by an NHR event.

A translocation model proposed by Richardson et al. [71] and later used to explain several rearrangements in a human translocation dataset [58] may provide useful insight for understanding the duplication and translocation of *Hun*. This model accounts for interchromosomal recombination and duplication while avoiding crossovers. In it, recombination occurs between nonhomologous chromosomes through the NAHR of LCRs, such as *Alus* [58]. According to this model, a DSB occurs in one of the two chromosomes (the X chromosome for the *Hun* scenario) near the LCR, followed by strand invasion of homologous sequence belonging to the intact chromosome (chromosome 3R). Strand extension would carry on for some length before rejoining its own chromosome (the X chromosome) at either more distal regions of homology or nonhomology. *Hun's* scenario differs from the previous cases in that we suggest that the initial recombination event between chromosome 3R and the X occurred between regions without any LCRs.

Sequence analyses of *Hun* from *D. simulans*, *D. sechellia,* and *D. mauritiana* revealed that the gene structure has evolved differently in each species [18]. *D. simulans* maintains a single open reading frame, while both *D. sechellia* and *D. mauritiana* have sustained deletions leading to seven and six premature stop codons, respectively. In *D. sechellia,* three significant deletions have occurred in the center of the gene. In *D. mauritiana,* the frame shift was caused by a single base deletion (Fig. 4a). Somewhat surprisingly, screens for the deletions in additional *D. sechellia* and *D. mauritiana* lines suggest that they are fixed. Given the young age of *Hun*, these mutations fixed in a rather short time span.

Along with structural changes, *Hun* has experienced expression evolution. *Bällchen* was shown to be expressed in both sexes in all species within the *D. melanogaster* subgroup. *Hun's* expression, on the other hand, is limited to males in *D. sechellia*, *D. mauritiana*, and *D. simulans*. Tissue-specific RT-PCR revealed that the gene's expression is testes-specific for each of the three species.

Molecular evolution and population genetic analyses were carried out to examine the role that selection has played on *Hun*. Both divergence and polymorphism-based measurements indicated that *Hun* is currently under purifying selection. To infer the role of selection in *Hun's* past, *D. simulans* population data was used. Though standard tests for selection based polymorphism frequency spectrum of *D. simulans* (Fu and Li's D [72], Fu and Li's F [72], and Tajima's D [73]) were nonsignificant alone, the McDonald-Kreitman Test [74] revealed a significant excess of amino acid replacement substitutions along the *Hun* branch (Fig. 4b). These results, combined with the expression

data, were taken as evidence that positive selection for a novel sex-related function drove the fixation of these substitutions.

HR As a Way of Generating Chimeric Proteins – NAHR

So far the evolutionary consequences that can arise through recombination events between chromosomal regions that largely lack sequence identity have been discussed. We now move on to discuss the growing recognition for the role of HR in the formation of chimeric gene structures.

Similar to the approaches used to study NHR, research that has focused on the role that LCRs play in genome rearrangements have primarily been based on naturally occurring disease-related cases [29]. However, with the availability of high quality genomes such as the human genome [75], a more general understanding of the evolutionary role that LCRs play is coming to light [28, 59, 60]. For example, there is growing evidence in primates that LINEs and SINEs are important in mediating rearrangements. Importantly, these rearrangements are not necessarily disease-related but instead have likely produced non-deleterious chimeric proteins as well as making more general evolutionary contributions [58, 60, 76] to primate genome architecture.

A striking example that illustrates the importance of LCRs in producing chimeric gene structures is found in primate *Alu* elements. *Alu* elements are the most numerous members of the SINE family of transposable elements and have been tied to several well-characterized genomic disorders [54, 64, 76]. Until recently, little was known about *Alu*'s more general evolutionary role in shaping genomic architecture. The picture greatly expanded with Bailey et al.'s [59] first fine-scale chromosome-wide analysis of segmental duplications within human chromosome 22. This study resulted in the identification of a surprising number of recent duplication events that resulted in 11 putative chimeric transcripts. Upon further examination of chromosome 22, Babcock et al. [58] reported on numerous *Alu*-related rearrangements including transpositions and duplications, some of which were involved in known chimeric structures. The highly non-random association between *Alus* and rearrangement breakpoints strongly suggested an expansive role for them throughout the genome. Bailey et al. [60] provided strong support for this conjecture through a genome-wide analysis of segmental duplication junctions, 9,464 in total. Out of these duplications, 27% of them had a breakpoint contained within an *Alu*.

Transposable Elements as 'Fragment Joiners'

Current evidence suggests that an important mechanism behind the generation of novel chimeric genes at the DNA level in plants is through the activity of transposable elements. Surveys of the completed genomic sequences of several angiosperms have uncovered a high abundance of MULEs, along with a

subfamily of MULEs that carry gene fragments between terminal inverted repeats. This latter subfamily of MULEs has been named Pack-MULEs [31]. In plant species, Pack-MULEs have been identified in maize [77, 78], rice [79], and *Arabidopsis* [80]. A genome-wide search within rice has identified over 3,000 Pack-MULEs that contain gene fragments averaging 325 bp (with a range of 47–986 bp). Overall, these Pack-MULEs contain DNA fragments from more than 1,000 functional genes. Further, it is estimated that about one-fifth of these identified Pack-MULEs contain DNA fragments from multiple genomic sites and have created novel chimeric gene structures (see fig. 2 for example). At least 5% of these chimeras appear to be functional, with evidence coming from identical full-length cDNAs as well as sequence divergence analyses [31]. In *Arabidopsis*, 5 Pack-MULEs have been identified. The size of the acquired DNA fragments range from 94 to 570 bp and comprise most of the internal DNA of the corresponding elements. However, thus far no Pack-MULE-related chimeric genes in *Arabidopsis* have been identified [26].

Helitrons are a newly identified class of eukaryotic transposable elements that were discovered in the genomes of *A. thaliana*, rice, *Caenorhabditis elegans*, and subsequently in maize [34]. In maize, repeated amplifications and transpositions of *Helitrons* have created numerous gene fragment clusters. Such a system provides intriguing potential for the formation of chimeric gene structures. Illustrating this, Lal et al. [33] recently reported an event in which a *Helitron* had inserted into the maize *Sh2* gene. Though this insertion is believed to have rendered *Sh2* nonfunctional, the example demonstrates a capacity for *Helitron* elements to create chimeric genes.

Chimeric Genes Generated by Gene Fusions

In eukaryotes, two fusion processes have been found to be responsible for the formation of chimeric genes. The first example involved two adjacent genes whereby deletions of the 3′ portion of a 5′ gene and the 5′ portion of a 3′ gene created the novel gene named *Sdic* (S) [36]. The ancestrally adjacent genes were *Cdic* (C) and *AnnX* (A). It was found that the ‘CA gene pair’ tandemly duplicated forming a ‘CACA’ conformation. This was followed by several deletions that eliminated parts of the central ‘AC’ portion, so that a 3′ UTR region belonging to *AnnX* was combined with the 5′ ends of the *Cdic* gene, thus forming a chimeric gene structure. Subsequent evolution within this new gene structure turned intron 3 of *Cdic* into a new promoter region and a new start codon emerged (fig. 3).

The second example is more complex, involving alternative splicing across two adjacent human genes, *KUA* and *UEV* [37]. *KUA* is comprised of 6 exons while *UEV* has 4 exons. Read-through transcription created a large transcript that was alternatively spliced generating the chimerical protein KUA-UEV. The

resultant transcript skipped the sixth *KUA* exon that contained its stop codon, as well as the *UEV* exon that contained its start codon. This case study demonstrates a sophisticated strategy to get rid of stop codons between two tandem genes [16].

Conclusion

Numerous chimeric genes appear to have been generated through DNA level recombination events. While a majority of these genes are ancient, and as result provide limited insight into the mechanism that brought them about, the few young examples that have been collected up to this point demonstrate diverse genetic and evolutionary histories. A major goal for future research is to build upon these young examples and develop a greater understanding of the origination events on a genomic level. With the increasing number of high quality genome projects becoming available, as well as experimental advancements, fundamental estimates for underlying mutational events are becoming feasible both within and between species.

Acknowledgement

This work was supported by a CAS-Max Planck Society Fellowship, a CAS key project grant (No. KSCX2-SW-121), a NSFC award (No. 30325016), a CAS OOCS fund (2004-2-2) and a NSFC key grant (No. 30430400) to W.W.; a USA National Science Foundation CAREER award (MCB0238168) and USA National Institutes of Health R01 grants (R01GM065429-01A1 and 1R01GM078070-01A1) to M.L. at the University of Chicago; a GAANN genomics grant supports J.R.A.

References

1 Kimura M: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, 1983.
2 Ohno S: Evolution by Gene Duplication. Springer, Berlin, 1970.
3 Betran E, Thornton K, Long M: Retroposed new genes out of the X in *Drosophila*. Genome Res 2002;12:1854–1859.
4 Betran E, Long M: *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics 2003;164:977–988.
5 Courseaux A, Nahon JL: Birth of two chimeric genes in the Hominidae lineage. Science 2001;291: 1293–1297.
6 Emerson JJ, Kaessmann H, Betran E, Long M: Extensive gene traffic on the mammalian X chromosome. Science 2004;303:537–540.
7 Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H: Emergence of young human genes after a burst of retroposition in primates. PLoS Biol 2005;3:1970–1979.

8 Katju V, Lynch M: On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. Mol Biol Evol 2006;23:11056–11067.
9 Betran E, Emerson JJ, Kaessmann H, Long M: Sex chromosomes and male functions – Where do new genes go? Cell Cycle 2004;3:873–875.
10 Jones C, Custer AW, Begun DJ: Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics 2005;170:207–219.
11 Long MY, Langley CH: Natural-selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science 1993;260:91–95.
12 Loppin B, Lepetit D, Dorus S, Couble P, Karr TL: Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. Curr Biol 2005;15:87–93.
13 Wang W, Brunet FG, Nevo E, Long M: Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci USA 2002;99:4448–4453.
14 Wang W, Yu HJ, Long MY: Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet 2004;36:523–527.
15 Brosius J: The contribution of RNAs and retroposition to evolutionary novelties. Genetica 2003;118:99–116.
16 Long M: Evolution of novel genes. Curr Opin Genet Dev 2001;11:673–680.
17 Long M, Betran E, Thornton K, et al: The origin of new genes: Glimpses from the young and old. Nat Rev Genet 2003;4:865–875.
18 Arguello J, Chen Y, Yang S, Wang W, Long L: Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. PLoS Genet 2006;2:e77.
19 Gilbert W: Why genes in pieces? Nature 1978;271:44.
20 Roth D, Wilson J: Illegitimate recombination in mammalian cells; in Kucherlapati R, Simth GR (eds): Genetic Recombination. American Society for Microbiology, Washington DC, 1988, pp 621–653.
21 Roth DB, Porter TN, Wilson JH: Mechanisms of nonhomologous recombination in mammalian cells. Mol Cell Biol 1985;5:2599–2607.
22 Roth DB, Wilson JH: Nonhomologous recombination in mammalian cells – role for short sequence homologies in the joining reaction. Mol Cell Biol 1986;6:4295–4304.
23 Allen C, Halbrook J, Nickoloff JA: Interactive competition between homologous recombination and non-homologous end joining. Mol Cancer Res 2003;1:913–920.
24 Roth D, Wilson JH: Relative rates of homologous and nonhomologous recombination in transfected DNA. Proc Natl Acad Sci USA 1985;82:3355–3359.
25 Schwartz M, Zlotorynski E, Goldberg M, Ozeri E, Rahat A, et al: Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. Genes Dev 2005;19:2715–2726.
26 Yu V, Koehler M, Steinlein C, Schmid M, Hanakahi LA, et al: Gross chromosomal rearrangements and genetic exchange between nonhomologous chromosomes following *BRCA2* inactivation. Genes Dev 2000;14:1400–1406.
27 Alexander JRB, Schiestl RH: Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. Hum Mol Genet 2000;9:2427–2334.
28 Stankiewicz P, Lupski JR: Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 2002;12:312–319.
29 Shaw C, Lupski JR: Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet 2004;13:R57–R64.
30 Stankiewicz P, Shaw CJ, Dapper JD, Wakui K, Shaffer LG, et al: Genome architecture catalyzes nonrecurrent chromosomal rearrangements. Am J Hum Genet 2003;72:1101–1116.
31 Jiang N, Bao Z, Zhang X, Eddy S, Wessler S: Pack-MULE transposable elements mediate gene evolution in plants. Nature 2004;431:569–573.
32 Bennetzen J: Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 2005;15:621–627.
33 Lal SK, Giroux MJ, Brendel V, Vallejos E, Hannah C: The maize genome contains a Helitron insertion. Plant Cell 2003;15:381–391.
34 Kapitonov V, Jurka J: Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci USA 2001;98:8714–8719.

35 Burns D, Horn V, Paluh J, Yanofsky C: Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused *Escherichia coli* tryptophan synthetase alpha and beta chains. J Biol Chem 1990;265:2060–2069.
36 Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL: Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 1998;396:572–575.
37 Thomson T, Lozano JJ, Loukili N, Carrió R, Serras F, et al: Fusion of the human gene for the polyubiquitination coeffector UEV1 with *Kua*, a newly identified gene. Genome Res 2000;10: 1743–1756.
38 Banyai L, Varadi A, Patthy L: Common evolutionary origin of the fibrin-binding structures of fibronectin and tissue-type plasminogen activator. FEBS Lett 1983;163:37–41.
39 Li WH: Molecular Evolution. Sinauer, Sunderland, MA, 1997.
40 Sudhof T, Goldstein JL, Brown MS, Russell DW: The LDL receptor gene: a mosaic of exons shared with different proteins. Science 1985;22:815–822.
41 Patthy L: Modular assembly of genes and the evolution of new functions. Genetica 2003;118: 217–231.
42 Patthy L: Genome evolution and the evolution of exon-shuffling – a review. Gene 1999;238: 103–114.
43 Patthy L: Protein Evolution By Exon-Shuffling. Springer, New York, 1995.
44 Patthy L: Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 1987; 214:1–7.
45 Hedges S: The origin and evolution of model organisms. Nat Rev Genet 2002;3:838–849.
46 Long M, de Souza SJ, Rosenberg C, Gilbert W: Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. Proc Natl Acad Sci USA 1996;93: 7727–7731.
47 Kaessmann H, Zollner S, Nekrutenko A, Li WH: Signatures of domain shuffling in the human genome. Genome Res 2002;12:1642–1650.
48 Liu M, Grigoriev A: Protein domains correlate strongly with exons in multiple eukaryotic genomes–evidence of exon shuffling? Trends Genet 2004;20:399–403.
49 Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD: Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet 2005;21: 326–329.
50 Shendure J, Church GM: Computational discovery of sense-antisense transcription in the human and mouse genomes. Genome Biol 2002;3:research0044.1–0044.14.
51 Fu Y, Yu JC, Cheng TC, Lou MA, Hsu GC, et al: Breast cancer risk associated with genotypic polymorphism of the nonhomologous end-joining genes: a multigenic study on cancer susceptibility. Cancer Res 2003;63:2440–2446.
52 Nobile C, Toffolatti L, Rizzi F, Simionati B, Nigro V, et al: Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum Genet 2002;110:418–421.
53 Zucman-Rossi J, Legoix P, Victor JM, Lopez B, Thomas G: Chromosome translocations based on illegitimate recombination in human tumors. Proc Natl Acad Sci USA 1998;95:11786–11791.
54 Hu X, Worton RG: Partial gene duplication as a cause of human disease. Hum Mutat 1992;1:3–12.
55 Allgood ND, Silhavy TJ: Illegitimate recombination in bacteria; in Kucherlapati R, Simth GR (eds): Genetic Recombination. American Society for Microbiology, Washington, DC, 1988.
56 van Rijk A, Bloemendal H: Molecular mechanisms of exon shuffling: illegitimate recombination. Genetica 2003;118:245–249.
57 van Rijk AAF, de Jong WW, Bloemendal H: Exon shuffling mimicked in cell culture. Proc Natl Acad Sci USA 1999;96:8074–8079.
58 Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, et al: Shuffling of genes within low-copy repeats on 22qll (LCR22) by Alu-mediated recombination events during evolution. Genome Res 2003;13:2519–2532.
59 Bailey J, Yavor AM, Viggiano L, Misceo D, Horvath JE, et al: Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am J Hum Genet 2002;70: 83–100.
60 Bailey J, Liu G, Eichler EE: An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 2003;73:823–834.

61 Abeysinghe S, Chuzhanova N, Krawczak M, Ball EV, Cooper DN: Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. Hum Mutat 2003;22:229–244.
62 Been M, Burgess RR, Champoux JJ: Nucleotide sequence preference at rat liver and wheat germ type 1 DNA topoisomerase breakage sites in duplex SV40 DNA. Nucleic Acids Res 1984;12: 3097–3114.
63 Borgato L, Bonizzato A, Lunardi C, Dusi S, Andrioli G, et al: A 1.1-kb duplication in the p67-*phox* gene causes chronic granulomatous disease. Hum Genet 2001;108:504–510.
64 Chi C, Tsai CR, Chen LH, Lee HF, Mak BSC, et al: Maple syrup urine disease in the Austronesian aboriginal tribe Paiwan of Taiwan: a novel DBT (E2) gene 4.7 kb founder deletion caused by a nonhomologous recombination between LINE-1 and Alu and the carrier-frequency determination. Eur J Hum Genet 2003;11:931–936.
65 Chou C, Morrison SL: A common sequence motif near nonhomologous recombination breakpoints involving Ig sequences. J Immunol 1993;12:5350–5360.
66 Kumatori A, Faizunnessa NN, Suzuki S, Moriuchi T, Kurozumi H, Nakamura M: Nonhomologous recombination between the cytochrome b_{558} heavy chain gene (*CYBB*) and LINE-1 causes an X-linked chronic granulomatous disease. Genomics 1998;53:123–128.
67 Nikiforov Y, Koshoffer A, Nikiforova M, Stringer J, Fagin JA: Chromosomal breakpoint positions suggest a direct role for radiation in inducing illegitimate recombination between the *ELE1* and *RET* genes in radiation-induced thyroid carcinomas. Oncogene 1999;18:6330–6334.
68 Singh G, Kramer JA, Krawetz SA: Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Res 1997;25:1419–1425.
69 Zhu J, Schiestl RH: Topoisomerase I involvement in illegitimate recombination in *Saccharomyces cerevisiae*. Mol Cell Biol 1996;16:1805–1812.
70 Tanizawa A, Hohn KW, Pommier Y: Induction of cleavage in topoisomerase I c-DNA by topoisomerase I enzyme from calf thymus and wheat germ in the presence and absence of camptothecin. Nucleic Acids Res 1993;21:5157–5166.
71 Richardson C, Moynahan ME, Jasin M: Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. Genes Dev 1998;12:3831–3842.
72 Fu YX, Li WH: Statistical tests of neutrality of mutations. Genetics 1993;133:693–709.
73 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585–595.
74 McDonald JH, Kreitman M: Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 1991;351:652–654.
75 Venter J, Adams MD, Myers EW, Li PW, Mural RJ, et al: The sequence of the human genome. Science 2001;291:1304–1351.
76 Batzer M, Deininger PL: Alu repeats and human genomic diversity. Nat Rev Genet 2002;3: 370–379.
77 Bennetzen J, Springer PS: The generation of mutator transposable element subfamilies in maize. Theor Appl Genet 1994;87:657–667.
78 Talbert L, Chandler VL: Characterization of a highly conserved sequence related to mutator transposable elements in maize. Mol Biol Evol 1988,5:519–529.
79 Turcotte K, Srinivasan S, Bureau T: Survey of transposable elements from rice genomic sequences. Plant J 2001;25:169–179.
80 Yu Z, Wright SI, Bureau TE: Mutator-like elements in *Arabidopsis thaliana*. Structure, diversity and evolution. Genetics 2000;156:2019–2031.

Manyuan Long
Department of Ecology and Evolution, University of Chicago
1101 East 57th St., Zoology
301E, Chicago, IL 60637 (USA)
Tel. +1 773 702 0557, Fax +1 773 702 9740, E-Mail mlong@uchicago.edu