Version 1.0 a0020756

a0001

Transposable Element-driven Duplications during Hominoid Genome Evolution

Advanced article

Article Contents
- Introduction
- Analysis of Segmental Duplications
- Implications of Segmental Duplications

Online posting date: 14th March 2008

Bud Mishra, *New York University, New York, USA*

Hominoid genomes exhibit a significant number of large segmental duplications; these and other similar duplications in mammalian genomes are hypothesized to be mediated by transposable elements such as Alus or L1s in hominoids and rodents, respectively. The true evolutionary mechanisms that sculpted these duplications are emerging to be much more subtle and complex.

s0001

Introduction

s0002

Genome structure and duplications

p0001

Akin to any large texts in a natural language, hominoid genomes appear as palimpsests of morphemes, lexemes and other lexical modules, each with its own structure, distribution and fluctuating copy numbers (Zhou and Mishra, 2004; Thomas *et al*., 2004). Duplication appears to be one of the main mechanisms in shaping and reshaping this genomic architecture. Duplication occurs at multiple scales, ranging at one end from small local tandem repeats that are the results of polymerase slippage or unequal crossing-over, to the duplication of a whole genome at the other extreme. Segmental duplication (SD) occurs at a scale between small repeats and the whole genome. Duplications provide the genome with additional and (initially) redundant copies of genes and their regulatory elements, thus initiating gene family expansion and offering the species adequate freedom to explore new functionality and develop more refined regulations (Armengol, 2005; Cheng, 2005). Duplication promotes a faster and more complex dynamics during genome evolution – the duplicated sequences can serve as the homologous region for further recombination events, which can lead to further duplication, deletion and other rearrangement events (Zhou and Mishra, 2004). As was found in the subtelomeric and pericentromeric regions of the mammalian genomes, SDs in such regions generate a complicated pattern of recursively nested events and creates a fertile breeding ground for novel functional elements. Years before the genomic era tools had pinpointed the duplicated motifs in genomes, in a surprisingly prescient thesis, S Ohno had already suggested a possible reciprocative roles between duplications and selections in the evolutionary dynamics, as he noted that 'Natural selection merely modified while redundancy created' (Ohno, 1970).

AU:2

AU:3

ELS subject area: Evolution and Diversity of Life

How to cite:
Mishra, Bud (March 2008) Transposable Element-driven Duplications during Hominoid Genome Evolution. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020756

s0003

Segmental duplications

p0002

Among the duplications discovered at many different scales, large SDs appear to be one of the most mysterious, both in their origin as well as function. SDs are defined as regions with multiple copies that are 1–20 kb in length with at least 90% sequence identity. These are also referred to as low-copy repeats (LCRs) (Shaikh *et al*., 2000). Incidence of surprisingly large amount of SDs in hominoid genomes (particularly, for humans and great apes including bonobos, chimpanzees, gorillas and orangutans) has been quantified and characterized to some degree, and has become a source of confusion and speculation (Bailey and Eichler, 2006). Most of these duplications appear to have occurred rather recently: namely, 30–60 million years ago (Mya), cover both coding and noncoding regions and include both intra- and inter-chromosomal events (Bailey *et al*., 2002, 2004; Cheng *et al*., 2005; Tuzun *et al*., 2004). SDs are distributed in the genome in a clustered manner, mostly around pericentromeric and subtelomeric regions, and are likely to have contributed considerably to the evolutionary dynamics of the mammalian genomes. Different studies have hypothesized and quantified the significant association between SDs and syntenic breakpoints (Armengol *et al*., 2003), indicating a role for SDs in large genomic rearrangement events. Additionally, many of the duplicated segments in the human genome have been found to be involved in further rearrangements, some leading to genetic diseases (Emanuel and Shaikh, 2001). The genic contents of the SDs suggest that they may also play a role in adaptive

evolution and a domain accretion process (Samonte and Eichler, 2002). Recently, it has been suggested that SDs in human and chimpanzee have contributed far more significantly to their inter-species genomic-level differences than any other mutations; for instance, SDs are estimated to contribute more than twice as much to the genomic differences than single nucleotide substitutions (Cheng *et al.*, 2005).

p0003 Estimates based on computational analysis of the draft genome sequences and FISH analysis of randomly chosen clones point to a higher incidences of SDs in human, chimpanzee and macaque (an old world monkey) genomes, but smaller in marmoset (a new world monkey) genomes (Bailey and Eichler, 2006). In comparison, other mammalian genomes exhibit somewhat lower SD incidences (estimates based on the analysis of rat, mouse and dog) (Bailey and Eichler, 2006). This subtle variation from genome to genome is now thought to be a function of many features of genome structure and complex dynamics (Cheng *et al.*, 2005; Zhou and Mishra, 2005; Zhou, 2005): namely, regions of thermodynamic instabilities, subterminal caps, composition of transposable elements and their rate of transposition, the other repeats at different scale, etc.

s0004 Transposable elements driven duplication

p0004 There have been several plausible hypotheses and confirmatory studies focusing on the molecular mechanisms of the duplication process. For instance, repeat elements, especially transposable elements, have been suggested to play an important role (Bailey *et al.*, 2003). A well-known example illustrates how, in an early ancestor of simian primates, repeat elements such as L1 long interspersed repetitive elements (LINEs) may have initiated the duplication of the γ-globin gene by unequal crossover (Fitch *et al.*, 1991). More recently, Alu, a short interspersed nucleotide element (SINE) in the primate genomes, has been hypothesized to be actively involved in various chromosomal rearrangements, including duplications, deletions and translocations, in the process creating recombination hotspots in both genetic diseases, such as tumour and normal genomic polymorphisms (Kolomietz *et al.*, 2002). Independently, detailed breakpoint flanking sequence analyses in the in-laboratory experiments that evolved *Escherichia coli* and *Saccharomyces cerevisiae* strains (Riehle *et al.*, 2001; Dunham *et al.*, 2002), also showed that the large genomic evolutionary events were mostly caused by the homologous recombination or transposition of the mobile elements (insertion sequences, or transposable elements and their relics). However, other studies have now suggested that duplications are also caused by repeat-independent mechanisms. For example, the presence of left-handed helical Z-DNA structure can induce recombination events by altering chromatin organization (Smith and Moss, 1994). Double strand breakage followed by nonhomologous end joining (NHEJ) may similarly lead to gene amplification (Difilippantonio *et al.*, 2002).

Analysis of Segmental Duplications s0005

Mapping segmental duplications s0006

SDs are relatively short and highly conserved, as defined p0005 operationally in terms of their length (1–20 kb) and degree to which their sequences are conserved (90–99.5%) (Samonte and Eichler, 2002). The limit on sequence identity is somewhat artificial, but necessary to avoid confusion with the false-positives in duplication mapping, which are mostly due to errors introduced by the shotgun sequence assembly heuristics. Recently, SDs in some mammalian genomes were mapped using both assembly-dependent and independent methods (Bailey *et al.*, 2003, 2004; Tuzun *et al.*, 2004; Zhang *et al.*, 2005). These duplication mappings, as characterized by their compositions, boundaries and flanking sequences, have been repeatedly 'mined' to derive clues about their origin and mechanisms that drive them.

Among the mechanisms that have been proposed, the p0006 most prominent ones are the followings: Alu-mediated transposition in pericentromeric regions (Bailey *et al.*, 2003; Cheng *et al.*, 2005, chromosomal instability (Mishra and Zhou, 2005), copy number expansion via NAHR (nonallelic homologous recombination) mediated by deoxyribonucleic acid (DNA) repeats, translocation followed by transmission of unbalanced chromosomal complements in human subtelomeric regions (Linardopoulou *et al.*, 2005), etc.

Flanking sequences and their properties s0007

Despite the extensive findings on the effect of SDs, little is p0007 known on how they rose. To look for hints of the molecular mechanisms responsible for SDs, the flanking sequences of the recent SDs in the mammalian genomes were carefully examined to detect important sequence motifs or signatures of molecular mechanisms (Zhou and Mishra, 2005; Zhou, 2005). Other genomic structures close to the SD flanking regions, such as synteny group structures and copy number polymorphism (CNP), were also examined for any association.

More generally, the 'signatures' of the duplication mech- p0008 anism have been sought through the mer-analysis (i.e. analysis of the frequency and statistical distribution of 'mers' or short words in a sequence) of the flanking sequences using computation of the mer-frequency distributions, focusing primarily upon 5- and 6-mer substrings (a *k*-mer is a substring of length *k*, **Figure 1**) (Paxia *et al.*, 2002). AU:4 These analyses have reported strong enrichment of A(T)-rich mers around the breakpoints as well as other positionally over-represented mers, which align with each other and yield the Alu consensus sequence. Although Alu contains A(T)-rich regions, after the removal of Alus, the A(T)-rich words remained significantly enriched at the breakpoints. The clustering of A(T)-rich words around the breakpoints and the computation of relative stability of the DNA duplex around the junctures of the duplicated regions suggest that the duplication breakpoints

preferentially reside in sequence regions that are more susceptible to strand dissociation. Therefore, there seems to be intertwined interactions among the transposons and genomic instabilities in the creation of SDs (Zhou and Mishra, 2005; Zhou, 2005).

s0008 ## Roles of transposable elements

p0009 Several statistical analyses of the genomic sequences, in the neighbourhoods of SDs, have concluded that a significantly higher proportion of the duplication pairs share common repeats in their flanking regions than what would be expected by random distribution. Based on these analyses, it was reasoned that Alus and LINEs, and possibly, even LTRs (long terminal repeats), e.g. mammalian apparent LTR-retrotransposons (MaLRs) and human endogenous retrovirus-like (HERV) LTR, must have played crucial roles at the molecular level leading to duplications by homologous recombination (Bailey *et al*., 2003). It was also discovered that only the repeats belonging to the relatively 'younger' subfamilies (those that amplified more recently) are significantly over-represented, suggesting that different repeats may have played different roles in this highly complex dynamic process (Bailey *et al*., 2003; Zhou and Mishra, 2005). To better understand the evolutionary function of these repeat families and the history of their involvement in mammalian SDs, Zhou and Mishra (2005) proposed a rigorous mathematical model. This stochastic Markov model, focusing on repeat-induced as well as alternative mechanisms, is capable of accurately describing the process of duplication, and the evolution of repeat distributions in the duplication pairs after duplication, along with independent contributions from other mechanisms of purely physical nature.

s0009 ## Zhou–Mishra model of segmental duplications

p0010 Zhou–Mishra (Zhou and Mishra, 2005; Zhou, 2005) model focuses on the hypothesis that recombination between homologous repeats from a family *X*, e.g. Alu or L1, contributes to the recent SD processes in mammalian genomes. The model reflects following intuitive observations: If some of the SDs were caused by repeat recombination, these duplications should contain compatible repeat configurations in its flanking regions right after the duplication events. In contrast, if the alternative hypothesis holds, then the configurations of repeats in the flanking regions would be statistically indistinguishable from any other randomly drawn genomic segments. The model, however, must and does take into account the mutational effects over time: namely, the possible gradual obliteration of the configurational signals that may have been originally presented by the causative repeats, or accidental introduction of bystander (i.e. noncausative) repeats that by happenstance align in the expected configuration. The model assumes that each of such genome-evolution process occurs in history-independent manner and can be encoded faithfully using a Markov process. After the passage of sufficiently long time and assuming stationarity in evolutionary rates, the repeat configurations in the flanking regions reach a stationary distribution over different duplication age groups. By carefully examining these stationary distributions, which vary depending on how frequently duplications were caused by repeat recombination, and comparing them against a null model derived from other randomly drawn genomic sequences, Zhou and Mishra (Zhou and Mishra, 2005; Zhou, 2005) were able to quantify how repeat-induced versus other possible mechanisms mediate SDs. However, as indicated earlier, the statistical inferences employed here must also appropriately account for both the highly active history of the over-represented repeats in the duplication flanking regions as well as (un)reliability of the genome assembly and duplication mapping data, as further described in Zhou's thesis (Zhou, 2006).

p0011 It should be noted that Zhou–Mishra model is currently the only mathematically rigorous model of SD that incorporates various evolutionary mechanisms simultaneously. However, many similar ideas also appear in other biological (some qualitative) models, for instance, the models describing duplication-dependent strand annealing in drosophilae (Fiston-Lavier *et al*., 2007), Alu transposition models in human (Zhang *et al*., 2005), models describing duplication-induced replication in yeast (Koszul *et al*., 2004) and models of telomere–telomere fusion in fungal pathogen/*Cryptococcus neoformans* (Fraser *et al*., 2005).

s0010 ## Quantifying mechanisms using Zhou–Mishra model

p0012 Using Zhou–Mishras' Markov model of the SD process in the human genomes, the authors (Zhou and Mishra, 2005; Zhou, 2005) discovered that approximately 12% of these recent SDs were caused by recombination mediated by the recent active interspersed repeats in the human genome. They also discovered that, in addition, the physical instabilities in the DNA sequence also affect the process to some extent by introducing 'fragile' sites in the genomes. Specifically, in human genomes, they detected significant activities of the repeats from the younger Alu subfamilies (AluY and AluS), but not a similarly significant role for the LINEs. While a similar picture is expected to hold for other hominoid genomes, the necessary computational analysis, using Zhou–Mishra model, is yet to be performed for chimpanzee (or other great apes) genome. In the mouse and rat genomes, Zhou–Mishra analysis did not find similar activities mediated by the SINEs (B1, B2, ID and B4), but only a role for the younger LINE1 (L1) subfamilies. In general, there is now an accepted view that the recombination mediated by high-homology repeats is a ubiquitous mechanism driving SDs in hominoid as well as all the mammalian genomes.

p0013 Furthermore, the results from Zhou–Mishra model also suggested that the SDs are likely to be caused by multiple mechanisms, and a large fraction (approximately 70%) of the duplications are caused by some unknown mechanism

independent of the interspersed repeat distributions, which is consistent with the conclusions of Zhang *et al*. (2005). Using other analyses, Zhou and Mishra also discovered an enrichment of DNA sequences that are physically unstable and occur predominantly around the duplication. Thus, they suggested that the variability in helix stability and the DNA flexibility might have also played a role in initiating or facilitating the SD process.

s0011 Implications of Segmental Duplications

s0012 Relation to gene duplications

p0014 In a modern genome, one detects duplications of both gene and nongenic regions, as they occur at different scales: namely, gene duplications, large SDs, chromosomal duplications resulting in polyploidy and whole-genome duplications. For instance, from the sequence of a related species, *Kluyveromyces waltii*, that diverged from *S. cerevisiae* before the duplication event and from the comparative study on the gene orders and copy numbers, scientists gathered the most convincing evidence for a whole-genome duplication in *S. cerevisiae* followed by a massive gene loss (Kellis *et al*., 2004).

p0015 Unlike such large-scale duplications, often, SDs contain only fragments of coding sequence, and do not necessarily cover a functional gene unit; in fact, on occasions, SDs may not carry any coding regions at all. But, by adopting a genome-scale view for the study of SDs, one could expand on the gene-centric view and recognize the subtle functional role of these duplications. In the most direct account, SDs are interpreted as the source of sites of new gene formation by domain shuffling, and may thus relate to the evolutionary implications of gene duplications. Although the idea of evolution by duplication (EBD) has appeared in the writing of JBS Haldane, the most unambiguous suggestion for it arrived in 1970s, when S Ohno proposed gene duplication as the primary driving force in evolution (Ohno, 1970). Ohno's theory of evolution by gene duplication became both verifiable as well as amenable to further generalizations, when large-scale sequencing and experimental efforts made available whole-genomic sequences of many organisms and open to various comparative genomics analyses.

p0016 However, taxonomy of paralogous genes in genomes (Lynch and Conery, 2000) and elucidation of mechanisms responsible for gene duplication through analysis of the age, scale and functional category of the duplicated pairs, had already been carried out long before the current interest in duplication processes in their full generality. For instance, the rates of gene duplication and deletion in different genomes had been quantified and found to be at a similar scale to the substitution rate (Lynch and Conery, 2000). For instance, these studies had resulted in some understanding of the fate of the duplicated genes: namely, it was hypothesized that after gene duplication, one of the duplicated copy preserves the original function while the selection pressure on the other copy is relaxed, allowing it to accumulate various mutations; the mutational copy eventually becomes a pseudogene by loss-of-function, or by chance gives rise to an advantageous gene with a new function, gained.

p0017 There have been several additional theories fleshing out the preceding skeletal scenario, of which two have acquired considerable prominence: mutation-during-non-functionalization (MDN) theory (Hughes, 1994), and its generalization in duplication–degeneration–complementation (DDC) theory (Force *et al*., 1999). In the MDN model of Hughes (1994), coupled with the population genetic theory, the model predicts that a duplicated gene is much more likely to experience loss-of-function in typical situations than gaining a new function, thus suggesting a low retention rate of the duplicated genes. However, observed negative and positive selection in the duplicated gene pairs, as evidenced by unusually high retention rate of duplicated genes in tetraploid fish lineages and *Xenopus laevis* (Van de Peer *et al*., 2001; Nadeau and Sankoff, 1997), has led to alternative theories, e.g. theory of gene sharing, in which the ancestral unduplicated genes first gain multiple functions, and after duplication, each daughter gene specializes one of the functions of the ancestral gene. In a more widely accepted theory due to Force and colleagues, centred around DDC model (Force *et al*., 1999), after duplication, the two gene copies acquire complementary loss-of-function mutations in independent subfunctions; both genes produce the full complement of functions of the single ancestral gene; and as population genetic theory (Walsh, 2003) would predict, duplicated genes are preserved by subfunctionalization. Thus, this model predicts significant extension of the period during which both genes are exposed to natural selection, thus improving the chance of gaining rare beneficial mutations to innovate novel functions under a greatly relieved selection pressure. Both models have found support from individual experimental data, such as the *Hox* genes and the *nodal* genes in zebrafish (Prince and Pickett, 2002), and the higher retention rate of the duplicated genes in tetraploid fish lineages and *X. laevis* (Van de Peer *et al*., 2001; Nadeau and Sankoff, 1997).

Speciation and segmental duplications s0013

p0018 Direct comparisons of the SDs between the mouse and rat genomes (Cheng *et al*., 2005) and between the human and chimp genomes (Armengol *et al*., 2005) have suggested that SDs contribute significantly to the genome-level inter-species differences. Specifically, there exists a significant enrichment of SDs in the regions where synteny is not preserved. For example, SDs contribute most significantly to the differences between human and chimpanzee genomes (more than a 2-fold contribution relative to that resulting from single nucleotide substitutions), only second to the contribution from deletion of ancestral duplications. More interestingly, the genes involved in SDs exhibit higher rates of molecular evolution than anywhere else in

the genome. Thus, SDs have come to be seen as the most significant contributor to the quantitative differences between closely related species and to the changes in their genomic structural landscapes.

s0014

Polymorphisms and segmental duplications

p0019 Once detailed studies of human genetic variation at the DNA level became a reality of the post-genomic era, they led to an intense focus on two major types of polymorphisms: namely, single nucleotide polymorphisms (SNPs) (Thorisson and Stein, 2003), and CNPs (gains and losses of certain genomic segments) (Iafrate *et al*., 2004; Sebat *et al*., 2004). Although our picture of human genetic diversity is still incomplete and biased by the limitations of the current genomic technologies, there have been many important strides: for instance, projects such as HAPMAP (International HapMap Consortium, 2005) have collected millions of SNPs and organized them into haplotype blocks; other collaborative efforts have resulted in similar combined catalogues of copy number variations, etc. Nonetheless, due to the limitation on the resolution and nature of the CNP analyses (Lucito *et al*., 2000; Mishra, 2002; Iafrate *et al*., 2004; Sebat *et al*., 2004) and the lack of more sophisticated computational analysis tools, our understanding of the relation among SDs, genomic instabilities and genomic rearrangements still remain preliminary.

p0020 Despite these limitations, rudimentary analyses have indicated that normal human genomes are able to tolerate a large number of loci of genomic imbalances that overlap with genes and frequently coincide with SDs in the genome. Since these local imbalances contribute directly to phenotypic variation and susceptibility to diseases, it becomes important to model how SDs and other associated higher-order architecture influence recombination, recurrent chromosomal rearrangements and copy number fluctuations. Various studies in this direction have led to following conclusions: Although there is a balance in the frequency of duplications and deletions, an observed 4-fold enrichment of CNPs within hotspot regions suggests that SDs very likely catalyse CNPs and other large-scale genomic variation (Locke *et al*., 2006; Redon *et al*., 2006). Conversely, SDs themselves were found to be significantly enriched ($>$4-fold) within regions with high density of CNPs. Furthermore, SDs define chromosomal rearrangement hotspots, thus mediating not only normal variation but also genomic diseases (Locke *et al*., 2006; Redon *et al*., 2006). Studies focused on heritability and linkage disequilibria (LD) of CNPs in duplication-rich regions of the genome have suggested that the CNPs in duplication-rich regions have strong LD with nearby SNPs and segregate on ancestral SNP haplotypes (Sharp *et al*., 2006; Locke *et al*., 2006). However, most of these results remain conjectural and need to be placed upon a firmer foundation through large, complete, high-density and high-quality genome-wide data and rigorous statistical analysis.

Diseases and segmental duplications

s0015

p0021 Genomic disorders, such as mental retardation and autism, have been suggested to co-occur with genomic rearrangements flanked by SDs (Sebat, 2007). In the emerging picture, many of the SDs (particularly, paired intrachromosomal duplications separated by 50 kb–10 Mb of intervening sequences) are postulated to act as substrates for NAHR, which results in pathogenic rearrangements exhibited as deletion, duplication or inversion of the intervening sequences. Some of these rearrangements (mapped as 'rearrangement hotspots' in the human genome) are found to be pathogenic and sources of genomic disorders (Sharp *et al*., 2006).

p0022 Consequently, there are conjectures that many SDs predispose certain neighbouring genomic regions to recurrent rearrangements (Sharp *et al*., 2006). For instance, in a study conducted by Sharp *et al*. (2006), 130 candidate regions were examined in the context of certain previously uncharacterized genomic disorders. In 290 individuals with mental retardation several pathogenic rearrangements were discovered through array-comparative genome hybridization (CGH) analysis, and with further higher resolution array analysis, six genes in chromosome 17 (17q21.31) were isolated as consistently deleted in a significant fraction of the affected individuals. Association studies such as these have now intensely focused on the relation among SDs, genes, CNPs, genomic rearrangements and genomic diseases.

References

Armengol L, Pujana MA, Cheung J, Scherer SW and Estivill X (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Human Molecular Genetics* **12**: 2201–2208. AU:7

Bailey JA, Church DM, Ventura M, Rocchi M and Eichler EE (2004) Analysis of segmental duplications and genome assembly in the mouse. *Genome Research* **14**: 789–801.

Bailey JA and Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews. Genetics* **7**(7): 552–564.

Bailey JA, Gu Z, Clark RA *et al*. (2002) Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Bailey JA, Liu G and Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics* **73**: 823–834.

Cheng Z, Ventura M, She X *et al*. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**(7055): 88–93.

Difilippantonio MJ, Petersen S, Chen HT *et al*. (2002) Evidence for replicative repair of dna double-strand breaks leading to oncogenic translocation and gene amplification. *The Journal of Experimental Medicine* **196**: 469–480.

Dunham MJ, Badrane H, Ferea T *et al*. (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the USA* **99**: 16144–16149.

Emanuel BS and Shaikh TH (2001) Segmental duplications: an 'expanding' role in genomic instability and disease. *Nature Reviews. Genetics* **2**: 791.

Fiston-Lavier AS, Anxolabehere D and Quesneville H (2007) A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Research* **17**: 1458–1470.

Fitch DH, Bailey WJ, Tagle DA *et al.* (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proceedings of the National Academy of Sciences of the USA* **88**: 7396–7400.

Force A, Lynch M, Pickett KB *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Fraser JA, Huang JC, Pukkila-Worley R *et al.* (2005) Chromosomal translocation and segmental duplication in *Cryptococcus neoformans*. *Eukaryotic Cell* **4**(2): 401–406.

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological Sciences* **256**: 119–124.

Iafrate AJ, Feuk L, Rivera MN *et al.* (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**(9): 949–951.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **27**: 1299–1320.

Kellis M, Birren BW and Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.

Kolomietz E, Meyn MS, Pandita A and Squire JA (2002) The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes, Chromosomes & Cancer* **35**: 97–112.

Koszul R, Caburet S, Dujon B and Fischer G (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *The EMBO Journal* **23**: 234–243.

Linardopoulou EV, Williams EM, Fan Y *et al.* (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**(7055): 94–100.

Locke DP, Sharp AJ, McCarroll SA *et al.* (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics* **79**(2): 275–290.

Lucito R, West J, Reiner A *et al.* (2000) Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Research* **10**(11): 1726–1736.

Lynch M and Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151.

AU:5 Mishra B (2002) Comparing genomes. *Computing in Science and Engineering* 42–49.

Nadeau JH and Sankoff D (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147**: 1259.

AU:6 Ohno S (1970) *Evolution by Gene Duplication*. Springer.

Paxia S, Rudra A, Zhou Y and Mishra B (2002) A random walk down the genomes: DNA evolution in VALIS. *Computer* **35**(7): 73–79.

Prince VE and Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews. Genetics* **3**: 827–837.

Redon R, Ishikawa S, Fitch KR *et al.* (2006) Global variation in copy number in the human genome. *Nature* **444**(7118): 444–454.

Riehle MM, Bennett AF and Long AD (2001) Genetic architecture of thermal adaptation in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the USA* **98**: 525–530.

Samonte RV and Eichler EE (2002) Segmental duplications and the evolution of the primate genome. *Nature Reviews* **3**: 65–72.

Sebat J (2007) Major changes in our DNA lead to major changes in our thinking. *Nature Genetics* **39**: S3.

Sebat J, Lakshmi B, Troge J *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**(5683): 525–528.

Shaikh TH, Kurahashi H, Saitta SC *et al.* (2000) Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Human Molecular Genetics* **9**(4): 489–501.

Sharp AJ, Hansen S, Selzer RR *et al.* (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics* **38**(9): 1038–1042.

Sharp AJ, Locke DP, McGrath SD *et al.* (2005) Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics* **77**(1): 78–88.

Smith PD and Moss SE (1994) Z-DNA-forming sequences at a putative duplication site in the human annexin VI-encoding gene. *Gene* **138**: 239–242.

Thomas EE, Srebro N, Sebat J *et al.* (2004) Distribution of short paired duplications in mammalian genomes. *Proceedings of the National Academy of Sciences of the USA* **101**(28): 10349–10354.

Thorisson GA and Stein LD (2003) The SNP consortium website: past, present and future. *Nucleic Acids Research* **31**(1): 124–127.

Tuzun E, Bailey JA and Eichler EE (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Research* **14**: 493–506.

Van de Peer Y, Taylor JS, Braasch I and Meyer A (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *Journal of Molecular Evolution* **53**: 436.

Walsh B (2003) Population-genetic models of the fates of duplicate genes. *Genetica* **118**: 279–294.

Zhang L, Lu HHS, Chung W-Y, Yang J and Li W-H (2005) Patterns of segmental duplication in the human genome. *Molecular Biology and Evolution* **22**: 135–141.

Zhou Y (2005) *Statistical Analyses and Markov Modeling of Duplication in Genome Evolution*. PhD thesis, Department of Biology, New York University.

Zhou Y and Mishra B (2004) Models of genome evolution. *Modeling in Molecular Biology*, *Natural Computing Series*, pp. 287–304, Lecture Notes in Computer Science.

Zhou Y and Mishra B (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proceedings of the National Academy of Sciences of the USA* **102**(11): 4051–4056.

Further Reading

Durand D and Hoberman R (2006) Diagnosing duplications – can it be done? *Trends in Genetics* **22**(3): 156–164.

Web Links

Chimpanzee Segmental Duplication Database, Genome Sciences, University of Washington; URL Link: http://chimpparalogy.gs.washington.edu.

Ensembl Genome Browsers; URL Link: http://www.ensembl.org.

Human Genome Segmental Duplication Database, Hospital for Sick Children, University of Toronto; URL Link: http://projects.tcag.ca/humandup.

Human Segmental Duplication Database, Genome Sciences, University of Washington; URL Link: http://humanparalogy.gs.washington.edu.

UCSC Genome Bioinformatics; URL Link: http://genome.ucsc.edu.

Article Title: Transposable Element-driven Duplications during Hominoid Genome Evolution

Article ID: A0020756

Article doi: 10.1002/9780470015902.a0020756

Version: 1.0

Previous version(s): None

Article type: Standard

Readership level: Advanced

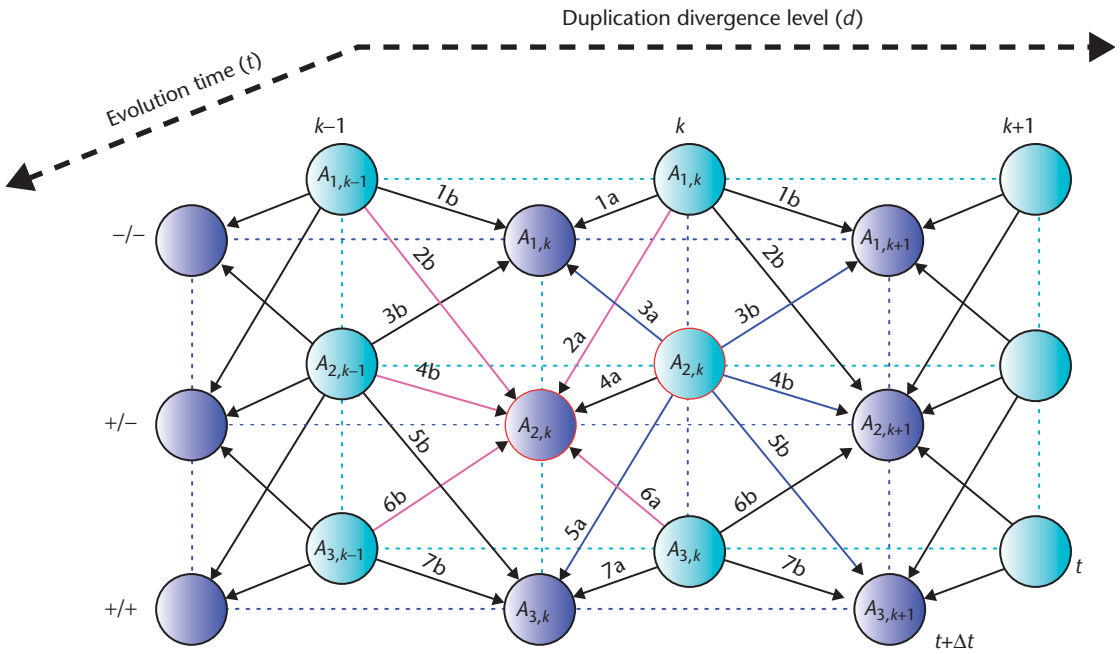
Top level subject categories: Evolution and Diversity of Life

Keywords: segmental duplications (SD) # low-copy repeats (LCR) # copy number variations (CNV) # transposable elements # rearrangement hotspots

AU:1

Glossary: None

Author(s) and affiliation(s): Mishra, Bud / New York University, New York, USA



f0001 **Figure 1** Zhou–Mishra model. The model formulates the changes in the distribution of flanking region pairs over different states as a Markov process over evolution time. At a particular evolution time, t, the flanking region pairs are distributed over different states (circles), defined by the configuration of the repeats in the flanking region (−/−, +/− or +/+) and the age group of the duplicated segments (k). During evolution, in each time interval Δt, the flanking region pairs may change its state through many possible transitions (arrows). The change in the distribution of the flanking region pairs in a particular state at time $t+\Delta t$ from time t depends on how much has entered into this state from other states, as well as how much has exited out of this state into other states in interval Δt since time t. At the same time, the flanking region pairs in other states can change into state $A_{2,k}$(dashed arrows). The difference between $A_{2,k}(t)$ and $A_{2,k}(t+\Delta t)$ can be calculated by taking the difference between the sum of the outflows (grey arrows) and inflows (dashed arrows). Reproduced from Zhou and Mishra (2005).

Author Query Form

Title: Encyclopedia of Life Sciences (ELS)

Entry Title/Article No: TRANSPOSABLE ELEMENT-DRIVEN DUPLICATIONS DURING HOMINOID GENOME EVOLUTION/a0020756

Dear Author,
During the preparation of your manuscript for typesetting some questions have arisen. These are listed below. Please check your typeset proof carefully and mark any corrections in the margin of the proof or compile them as a separate list.

Queries and/or remarks

AU:2	The references Armengol (2005), Cheng (2005), Armengol *et al*. (2005), Mishra and Zhou (2005) and Zhou (2006) are not listed in the reference list. Please provide complete bibliographic details.
AU:3	We notice that there are no cross-references to other ELS articles in this article. Please refer to your proofreading instructions if you would like to add cross-references to published ELS articles.
AU:4	Please confirm the call out of Figure 1.
AU:7	Sharp et al. (2005) has not been cited in the text. Please provide the citation for the same.
AU:5	Please provide the vol. no. in Mishra (2002).
AU:6	Please provide the place of publication in Ohno (1970).
AU:1	The abbreviation of 'transposable elements' has been deleted from the keywords list as it is not used in the text. Please check.