

Encyclopedia of Life Sciences

**The Impact of Gene Duplication on Human Genome Evolution**

Article Unique ID: A20841

**Cotton, James A**

James A Cotton  
School of Biological and Chemical Sciences  
Queen Mary, University of London  
Mile End Road  
London E1 4NS, UK

## **Definition**

Gene duplication has had a significant impact on all genomes and the human genome is no exception, as gene duplication contributes much of the raw material for natural selection to shape novel genes. In the context of human evolution, interest in gene duplication has been intense as the human genome is particularly rich in duplicated genomic regions. These duplicate genes contribute to genomic instability, leading to genome rearrangement and speciation. Recent evidence suggests that duplicated genes have undergone greater diversification than other loci in the human lineage, and so have been key in the evolution of uniquely human traits.

## **Key words** (up to 5)

2R hypothesis,

## **Contents**

1. Introduction
2. Gene duplication in genome evolution
3. The fate of duplicated genes
4. Rates and patterns of duplication in the human genome
5. The '2R' event of genome duplication
6. Duplication, rearrangement and speciation
7. Gene duplication and human traits
8. Conclusion

## 1. Introduction

A striking feature of the human genome is the high density of segments of duplicated DNA. A total of around 13.7% of the human genome is thought to consist of duplicated sequence. Most of this duplicated material is small, nonfunctional pieces of DNA that are likely to be rapidly deleted, but much of it consists of relatively large duplications that might contain intact functional elements. Pairs of genomic regions showing over 90% sequence similarity over at least 1kb make up about 5.4% of the human genome and most of this (5.0%) is in regions over 5kb in length. This content is certainly greater than that of most other vertebrate species, with around 2.7% of the mouse and chicken genomes and 1.6% of the rat genome being duplications over 1kb, and less than 6% being duplicated in total. Notably, the density of duplicated regions varies greatly across the human genome, varying between autosomes from 1.7% to 11.9%, and up to 50.4% for the Y chromosome (figures from Bailey and Eichler, 2006). The character and density of duplicates also varies between regions of chromosomes: the presence of duplications tends to increase the local rate of origin of new duplicates, leading to hotspots with high numbers of duplicated segments. There is also variation between chromosome regions: pericentromeric regions account for about a third of duplicated material in humans, and are particularly enriched for interchromosomal duplicates, and subtelomeric regions are similarly enriched, but to a far lesser extent. Duplicated regions across the rest of the genome are mostly intrachromosomal and particularly contain clusters of tandem duplicates. These duplications also tend to be younger than interchromosomal duplications and more gene-rich. **See also: 1699,5001,3855.**

Duplication of genetic material has probably played a major role in the evolution of all genomes. Gene duplication provides the raw material for the generation of new genes, and so is one of the principal drivers of evolutionary novelty at the molecular level. They also play a role in promoting genome rearrangement and, probably, in driving speciation. Here, I review the general role of gene duplication in genome evolution, drawing on data from a wide range of organisms. I then focus on the human genome, describing both the pattern of duplication in the human genome, and highlighting the particular roles gene duplication has probably played in human evolution. The large amount of duplicated material in the human genome suggests gene duplication may have been particularly important in human evolution, and there is evidence that it has been crucial in the evolution of a number of uniquely human traits during the most recent period of our evolutionary history. **See also: 1810.**

## 2. Gene duplication in genome evolution

The most important consequence of gene duplication is the production of new genes. Indeed, gene duplication is almost certainly the major source of novel genes capable of playing distinctive new roles in metabolism, cell function and gene regulation. There are now numerous examples of duplicated genes acting

as a substrate for natural selection to fashion novel proteins – indeed, it is hard to think of other mechanisms that could easily produce novel, functional genes. Most genes must have originated by modification of existing genes, but gene products have evolved to be well adapted to their current role, and most have been under strong purifying selection for most of their history, preventing them changing. It thus seems likely that processes like gene duplication have been essential in producing copies of existing genes, or of parts of existing genes (through processes like domain shuffling) that are free to significantly vary. **See also: 5087,5124,5096.**

Perhaps the most obvious result of gene duplication in eukaryotic genomes is that most genes are members of large families of related genes, and even of superfamilies of genes sharing structural or functional motifs. Many of these families are extremely large: the human genome contains around 339 functional olfactory receptor genes, together with 297 pseudogenes and there are at least 400 immunoglobulin genes in three different families. Protein superfamilies can be even larger: the entire complement of g-protein coupled receptors (GPCRs) in the human genome is probably around 1400. Large gene families also occur in bacterial genomes, albeit on a smaller scale, but there is evidence that bacterial genomes show rates of both gene duplication and gene loss, so that there is considerable turn-over of gene copies even in these smaller genomes. The birth and death of genes in gene families can lead to a complex tapestry of orthology and paralogy between gene copies in different organisms. **See also: 5045,1702,a000603,5125,5094,5072,5297,5298.**

### **3. The fate of duplicated genes**

There is little doubt that duplication mutations have the capacity to provide copious raw material for selection to work on, but the fate of newly-minted duplicates is less clear. For a duplication to persist through evolutionary time, an initially unique mutation must spread throughout a population. With two initially identical copies of a gene, one copy is probably more-or-less redundant, and so is a target for mutation, free from purifying selection. However, most mutations are deleterious, and many will disrupt the coding sequence or promoter elements of a gene, rendering it inactive or incapable and so producing a pseudo-gene. For a duplicated gene to evolve a new function, positively-selected mutations for the new function would need to occur before any loss-of-function mutation. Such mutations seem likely to be much more common than beneficial ones, leading to a slight mystery over how exactly duplicated genes have produced the remarkable diversity of existing gene families: while most gene families must be the result of gene duplication and subsequent selection to new functions, most duplicate genes are probably quite rapidly removed from the genome. **See also: 1737,5096,5125.**

An important solution to this conundrum was the suggestion that duplicate genes could be retained through a process that, at least initially, only involves

degenerative mutations. The duplications-degeneration-complementation (DDC, or subfunctionalisation) model proposes that duplicate copies of genes with multiple sites of expression, or with multiple functions, become fixed when different copies lose different regulatory elements or different functional sites. This model avoids the difficulty of requiring positively-selected gain-of-function mutations to occur before the effects of loss-of-function mutations are felt. The emphasis on regulatory elements also makes sense. As we learn more about eukaryotic gene regulation, it is becoming increasingly clear that regulatory elements may be as large, or larger, than the coding sequence for a particular gene, so these elements present a large target for mutation. A great deal of evidence supports this model: a number of gene pairs are known that have partitioned the expression pattern, splicing variants, or functions of a single-copy ancestor. Such evidence is slightly circumstantial, as such differences could have evolved following fixation by some other process – cases of exchange of expression patterns, at least, between paralogs, have been observed. In any case, theoretical models of population genetics also support the likelihood of the DDC model, and it must be seen as the likely explanation for the fixation of many, if not most, duplicate gene copies.

In principle, it should be easy to establish the relative contribution of subfunctionalisation and the alternative (neofunctionalisation) model: genes fixed by the latter process should show signs of positively selected substitutions, while those fixed by the former process should not. Unfortunately, things are not so easy – following the subfunctionalisation process, there is nothing to stop positive selected changes then occurring as the genes (now free from pleiotropic constraints) adapt to more specific roles. Indeed, this might be expected to be a very frequent occurrence. It can also be difficult to detect positive selection from the ratio of synonymous to non-synonymous substitutions – there is very little statistical power to detect higher ratios when few substitutions have occurred, and choosing between the models will depend upon detecting selection acting on the very first few substitutions, which may be difficult, or even impossible. A few studies have shown that dN/dS ratios are higher among young pairs of duplicates than older duplicates, but none have shown ratios as high as 1, which is generally taken to indicate positive selection – there is thus evidence for at least a relaxation of purifying selection early in duplicate evolution, but this is probably to be expected under either of the two main models. **See also: 5110.**

#### **4. Rates and patterns of duplication in the human genome**

Clearly, the potential importance of gene duplication in shaping the human genome will depend upon how frequently gene duplication mutations arise, and how rapidly they are lost through deletion mutations, or by becoming pseudogenes. There has been significant research interest in attempting to estimate the frequency of these events, particularly in the human genome. **See also: 4311.**

The earliest direct estimate of these rates appeared in an influential paper by Lynch and Conery in the year 2000, and subsequently updates as Lynch and Conery (2003), using data solely from paired duplicate genes, and assuming a strict molecular clock for silent substitutions. Lynch and Conery find a duplication rate for humans of around 0.009/gene/myr, significantly higher than for any of the other organisms they examined – which included two yeast species, *Arabidopsis* and the model animals *Caenorhabditis* and *Drosophila* – except the intracellular pathogen *Encephalitozoon cuniculi*. The suggestion is that both *E. cuniculi* and humans share a smaller effective population size, which would increase the chance of neutral mutations to become fixed. This, in itself, seems quite strong circumstantial evidence that neutral (or at least nearly neutral) processes dominate in the fixation of duplicate genes. Interestingly, Lynch and Conery also estimated the half-life of duplicates, which translates into an estimate of the rate of gene loss.

More recently, Lynch (2007, p. 198) has presented a different estimate of per-generation duplication rate that translate to a rates of around 0.00123/gene/myr, together with a loss rate almost twenty times higher. This estimate appears to be congruent with a number of estimates from other sources. Demuth et al. (2006) used a constant rate birth-death model of gene family size (and so discarding precise information about the actual ages of the duplications themselves). Their model constrains birth and death rates of new genes to be identical, but has the advantage that it is explicitly phylogenetic, increasing the likely precision of their estimates, and allowing them to estimate the rate across an entire phylogenetic tree for human, chimpanzee, mouse, rat and dog, encompassing at least 90 myrs of evolutionary history. Demuth et al. (2006) estimate a rate of 0.0016 duplications and losses/gene/myr.

Direct estimates of the distribution of duplication ages can be obtained using phylogenetic methods and relaxed molecular clock techniques, and using outgroup genes to calibrate the clock. A number of authors have constructed such distributions, for human gene duplicates as well as for a range of other species, and an example is shown in figure 1. More-or-less formal models of the birth-death process can be fitted to this kind of data, allowing a fairly direct estimate of rates of gene duplication and gene loss in the lineage of the genome in question. Using this kind of approach, Cotton and Page (2004) found values that appear to be very similar to other recent estimates, with a duplication rate of 0.00115/gene/myr and a loss rate of 0.00740/gene/myr in the human lineage over the last 200 myrs. **See also: 1669,5111,5135.**  
<figure 1 near here>

Interestingly, the age distribution produced by Lynch and Conery (2003) shows no sign of the “bump” of duplications observed in other data, perhaps because of their exclusive focus on pairs of duplicates rather than entire gene families. Another interesting pattern is that those methods allowing estimates of duplication and loss rates together suggest that loss is very much more rapid

than gene birth, underlining the relative rapidity with which most duplicate copies are removed from the genome. This rapid loss leads to the “hollowed-out exponential” rise in the number of gene copies of recent origin (figure 1). It is striking that similar data for segmental duplications do not show this pattern, as they show fewer very recent duplicates than older copies (figure 2). If this pattern is real – sequence assembly errors may have had an effect on these data, though it is unlikely to explain the entire effect – it suggests either a declining rate of segmental duplication, or increasing rate of deletion, during the last 40 myrs. **<figure 2 near here>**

## **5. The ‘2R’ event of genome duplication,**

A striking feature of figure 1 is the spike in duplicate copies formed around 450-500 million years ago. This certainly shows an episode of extensive gene duplication early in vertebrate evolution, and a large, and growing body of evidence suggests that these excess duplications are the result of a specific mode of duplication – the doubling of the entire genome through polyploidisation. Most authors agree that two polyploidy events may have occurred twice in quick succession. Polyploidy is now known to have occurred more-or-less frequently during the evolution of many groups, most notably allopolyploidy through hybridisation in flowering plants, but the suggestion that an ancestral vertebrate underwent two rounds of polyploidy was controversial when first put forward by Susumu Ohno in 1970. This was particularly because direct evidence for his hypothesis (since christened the ‘2R hypothesis’, as it proposes two rounds of genome duplication) waited largely until the availability of molecular data in the mid-to-late 1990s. It subsequently spawned a cottage industry of research papers proposing evidence both for and against the 2R event, and debating the timing and extent of the duplication events, with various papers proposing explanations from a simple expansion of smaller-scale duplication events, to 1, 2 and even 3 genome duplications, together with various mixtures of these possibilities. **See also: 1186, 5096, 5071.**

As this extensive literature suggests, the 2R hypothesis has proved difficult to test, because of the long period of time since the event. A number of different data might be useful in distinguishing a whole-genome duplication event from smaller, independent events. Following two rounds of duplication, ancestrally single-copy genes should be present in four copies, these duplicates should occur in large syntenic blocks, and the duplicates from each round should all have formed simultaneously. There are difficulties with all these lines of evidence. Background duplication and deletion has produced many more recent duplicates and removed many of the ‘2R’ duplicated copies, and genomic rearrangements have scrambled the duplicated blocks that are characteristic of large-scale duplication. Saturation of substitutions at silent sites make it difficult to accurately date duplication events over such long periods, so it is also difficult to identify temporal congruence. Nevertheless, this debate has quietened recently, with some high-profile papers making a strong case that two whole-

genome duplication events did indeed occur (e.g. Dehal and Boore, 2005). Crucially, these papers combine evidence from a number of different analytical approaches, for example including both map-based data (confirming that old duplicates are significantly clustered in the genome) and phylogenetic evidence (showing that the duplications occurred closely in time). Just as important has been the availability of genomic data from a growing number of vertebrate species, both from jawed vertebrates that post-date the 2R event and from stem chordates that provide an outgroup pre-dating the event.

An ancient event, probably occurring before the origin of vertebrate jaws, may seem to have little relevance for understanding the modern human genome, but a surprisingly large number of functionally important genes may descend from this event, leading a number of authors to suggest that this event may have been crucial in the evolution of complex vertebrates. In particular, a number of studies (e.g. Blomme et al., 2006) have shown a high retention of genes involved in transcriptional regulation, signaling, and development following genome duplication in vertebrates. In contrast, genes for processes like electron transport and 'amino acid and derivative metabolism' and 'RNA binding' appear to have been retained more often following duplication in small-scale events. Similar patterns of biased retention are found in plant genomes following polyploidy. The most likely explanation for these results is that it has something to do with dosage effects: following a whole-genome duplication, every gene is present in duplicate, so there is no change in the relative rates of transcription and translation, and stoichiometric interactions between different genes, which might be disrupted by duplication of some partners and not others, are unaffected.

The different patterns of duplicate retention following this mode of duplication may give it importance beyond the relatively few duplicates in modern genomes that were born during this event: less than one-third of gene families show even a single duplication from around the time of the 2R event in any modern vertebrate for which whole-genome sequence is available (Blomme et al., 2006). Changes in gene regulation were probably vital in the rapid evolution of development that occurred at the origin of vertebrates, producing such developmental novelties as neural crest and epidermal placodes. Indeed, the discovery that vertebrate *hox* genes occurred as four clusters of related genes was largely responsible for triggering renewed interest in the 2R hypothesis in the modern genomic era, and evolutionary developmental biologists have become particularly interested in early vertebrate genome evolution as a result. Kasahara (2007) reviews evidence that genes involved in the vertebrate immune system, particularly in the major histocompatibility complex, but also including natural killer receptors and some immunoglobins, were particularly expanded during this time. A more complex immune systems thus appears to be another innovation appearing at around the same time that is due to the 2R event. **See also: 1062, 1661,a0005921,a0006125.**

## **6. Duplication, rearrangement and speciation**

Gene duplication probably plays a major role in the movement of genetic material around the genome, by promoting genome rearrangements. The most common mechanisms for this rely on recombination between different parts of the genome rather than between the same allele on homologous chromosomes. Duplicated DNA regions lead to homologous sequences appearing at different loci, so that non-allelic homologous recombination (NAHR) can take place, leading to additional duplication, deletion of DNA, transposition of genetic material between chromosomes and inversions, depending on the relative position and orientation of the duplicated regions (figure 3). Even if duplicate loci are not involved in the mutational processes of genome rearrangement, gene duplication probably promotes the maintenance and spread of these mutations through populations, because translocations of genes may be more likely to be neutral, or less deleterious, if another copy remains intact and unmoved at the original locus elsewhere in the genomes. **See also: 1500,a0005798,1447.**

**<figure 3 near here>**

Apart from the above models of plausible molecular mechanisms, empirical evidence of a link between duplication and rearrangement comes from a number of different studies. The greatly increased content of duplicated segments that arose from whole-genome duplication is known to have particularly increased the rate of genome rearrangement in yeast, and is particularly well-known from comparative studies of the many polyploid flowering plants. In vertebrates, boundaries between blocks of syntenic genes rearranged between the human and mouse genomes are enriched for duplicated loci. Finally, there is also considerable evidence that duplicated genetic material accounts for much of the chromosome restructuring in the great ape lineage, and is a major source of genome structure variation within humans. Out of 11 large-scale rearrangements between chimpanzee and human genomes, 8 of them have breakpoints mapping to duplicated gene loci. The chimpanzee genome sequence reveals a similar picture at a finer scale, with many deletions, duplications and insertions unique to both genomes being associated with shared duplications. **See also: 5805.**

Genome rearrangement has been important in generating the diversity of genes in modern genomes, by allowing different parts of different genes to be shuffled around, creating novel combinations of functional domains and regulatory elements. It might have even more evolutionary importance, however, as an important driver of speciation. Genomic changes are thought to be of crucial importance in the evolution of post-mating reproductive isolation between incipient species, which leads to hybrids being infertile. Such post-zygotic isolating mechanisms are thought to have been important in driving speciation in a number of well-studied systems. Under the chromosomal model of speciation, large rearrangements lead to an increased chance of mis-segregation, where homologous chromosomes fail to pair properly prior to cell division. While this process relies on large-scale rearrangements, which may be promoted by the presence of duplicate genes, smaller scale events may lead to genetic

differences that, while not interfering with meiosis or mitosis, lead to reduced fitness of hybrids. Such “genetic incompatibilities” may be directly due to gene duplication and loss, if different and diverging copies of a duplicated locus are maintained in different populations (Lynch, 2007; pp228-235). There are more than 1000 rearrangements between human, macaque and chimpanzee genomes, so these kinds of events certainly could have played a role in recent speciation among our close ancestors. **See also: 1747,4168.**

## 7. Gene duplication and human traits

It is an oft-stated fact that humans and chimpanzees differ by around 1% at the DNA level (e.g. Mikkelsen et al 2005), so that this relatively small amount of coding sequence change must explain all the unique traits of the human lineage, such as bipedalism, hairlessness and greater cognitive ability. While some of this is undoubtedly human chauvinism – we probably see ourselves as more distinct from other primates than we really are, given that we share a common ancestor with the chimpanzee only 6 million years ago – this mystery has deepened with evidence that few genes have experienced positive selection since the divergence of humans and chimpanzees. An early study found 35 genes that showed significant positive selection, but none of these were statistically significant once corrected for their use of multiple statistical tests, suggesting that positive selection at the nucleotide level might have played no part in recent human evolution. **See also: 5300.**

Better resolution has been provided by the recent availability of the macaque genome sequence. The ancestor of macaques diverged from the hominid lineage around 25 million years ago, so this sequence acts as an outgroup, allowing the identification of genes under positive selection specifically on the human-chimp lineage. This has deepened the mystery still further, with both genes that are under positive selection on the branch leading to humans from our common ancestor with the chimpanzee apparently under selection generally across the great apes (Gibbs et al., 2007). Intraspecific, population-level data has the potential to provide far more power to detect interspecific positive selection, but the relatively limited amount of such data available seems to agree that positive selection on single-copy genes has played little role in the evolution of the human lineage. The ratio of replacement substitutions to silent substitutions between human and chimpanzees is almost identical to the ratio of replacement polymorphisms to silent polymorphisms in humans. Both the substitution and polymorphism figures should include both neutral (and mildly deleterious) mutations, but positively selected mutations will become fixed quickly and be missing from the polymorphisms. The fact that these ratios are almost identical thus suggests that positive selection has been responsible for very little of the evolution within the African apes, in contrast to results of similar studies comparing humans and monkeys, and in *Drosophila* (Mikkelsen et al., 2005; see discussion in Lynch, 2007, pp.81-82, for other primary references). **See also: 5449,1809.**

For methodological reasons, these genome-scale scans for signs of positive selection focus on single-copy genes, where orthology between copies can easily be established. They are thus missing a substantial amount of evolutionary novelty, and a possible resolution to these questions lies in the genes duplicated in the human lineage. Blomme et al. (2006) have estimated that a total of 396 gene duplications that left copies in the human genome occurred along the branch leading to humans since the divergence of primates and rodents around 100 million years ago, more than any other branch not associated with a whole-genome duplication event. Other genomic analyses have suggested that many more genes show increases in copy number specifically in the human genome than in the chimpanzee genome, leading to almost 3% divergence between the two species in the presence of large duplicated regions, a total of 76.3 Mb of genetic material. A different analysis of the chimpanzee draft genome suggests that as many as a third of recent duplications differ in copy number or content between human and chimpanzee.

Despite that lack of many genes with  $dN/dS$  significantly above 1, there is significant evidence for raised mean omega ( $dN/dS$  ratio) values in primates over other mammals, such as rodents, and this is particularly true in the chimpanzee and human lineages. There are two possible explanations for this pattern: the accumulation of duplications could be a sign that positive selection has favored the evolution of novel gene functions in the human lineage, or of almost the opposite effect: that the power of purifying selection to remove mildly deleterious duplications before they become fixed is reduced in the human lineage, perhaps due to population size bottlenecks. It is fairly uncontroversial that the modern human population is descended from a fairly ancestral population around 2 million years ago, but suggestions of a much more recent bottleneck have been controversial. It has been suggested that climate change following a large volcanic eruption at Lake Toba in Indonesia around 75,000 years ago might have reduced the human population to between 1,000 and 10,000 breeding pairs, but population genetic evidence for this event is at best indecisive. **See also: 1786.**

Demuth et al. (2006) also report a total of 414 gene families expanding along the primate lineage, with only 86 contracting. This large-scale turnover of gene copies has led to a 6.4% difference between human and chimpanzee genomes in gene content, far higher than the difference observed at the sequence level. A particular innovation in this work is their use of a statistical birth-death model to identify gene families showing statistically significant changes in copy number along a mammalian phylogeny. They identify 30 gene families that appear to have varied specifically within the human lineage, including a number of genes implicated in diseases like autism and autoimmune disorders. Intriguingly, the largest gene family created along the primate lineage (with 63 human and 46 chimpanzee copies) is largely of unknown function, but contains the gene *capase-7*, part of an apoptosis pathway involved in the development of the brain and nervous system. The same group has performed a similar analysis including

Macaque genome data (Gibbs et al., 2007), and found a similar pattern, identifying 20 gene families that have expanded specifically since our common ancestor with the chimpanzee.

The macaque genome analysis also provides evidence that positively selected single-copy genes tend to cluster near segmental duplications, and there is substantial evidence from other systems that selection acts more commonly on duplicated genes than single-copy genes. Taken together, this suggests that single-copy genes may be less important than variation in gene content in the very recent evolution of humans, and that gene duplication may have been the pre-eminent process in producing many uniquely human traits.

## **8. Conclusion**

The continual flux of gene birth and death plays an important role in providing the raw material from which natural selection can shape novel genes, and in promoting genome rearrangements that alter chromosomal structure and promote speciation. As Lynch (2007, p.194) explains, “the gene duplication process provides fuel for both of the major engines of evolution: adaptive phenotypic change within lineages and the creation of new lineages by speciation”. In the context of human evolution, both ancient and modern gene and genome duplications have contributed to the human genomic repertoire and may have played a key part in both producing the human species and in the evolution of many of the key traits setting us apart from our closest relatives.

## References

Bailey, J. A. and Eichler, E. E. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics* 7:552-564.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S. and Van de Peer, Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology* 7:R43.

Cotton, J.A. and Page, R.D.M. 2005. Rates and patterns of gene duplication and loss in the human genome. *Proceedings of the Royal Society of London, series B* 272:277-283.

Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3:e314.

Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N. and Hahn, M.W. 2006. The Evolution of Mammalian Gene Families. *PLoS One* 1:e85.

Kasahara, M. 2007, in press. The 2R hypothesis: an update. *Current Opinion in Immunology*. doi:10.1016/j.coi.2007.07.009.

Lynch, M. 2007. *The Origins of Genome Architecture*. Sinauer, Sunderland, MA.

Lynch, M. and Conery, J.S. 2003. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* 3:35-44.

Mikkelsen, T.J. et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genomes. *Nature* 437: 69-87.

Gibbs, R.A. et al. 2007. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science* 316:222-234.

### **Further reading (up to 10)**

Conrad, B. and Antonarakis, S.E. 2007. Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annual Review Genomics and Human Genetics* 8:2.1-2.19.

Donoghue, P.C.J. and Purnell, M.A. 2005. Genome duplication, extinction and vertebrate evolution. *Trends in Ecology and Evolution* 20:312-319.

Fortna, A., Kim, Y., MacLaren, W., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., Karimpour-Fard, A., Glueck, D., McGavran, L., Berry, R., Pollack, J., Sikela, J.M. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology* 2:e307.

Furlong, R.F. and Holland, P.W.H. 2002. Where vertebrates octoploid? *Philosophical Transactions of the Royal Society of London, series B* 357:531-544.

Lynch, M. 2007. *The Origins of Genome Architecture*. Sinauer, Sunderland, MA.

Ohno, S. 1970. *Evolution by Gene Duplication*. George Allen & Unwin, London.

Shaw, C.J. and Lupski, J.R. 2004. Implications of human genome architecture for rearrangement-based disorders: the genome basis of disease. *Human Molecular Genetics* 13:R57-R64.

Taylor, J.S. and Raes, J. 2004. Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics* 38:615-643.

## **Glossary**

**autosome** – any of the (22 pairs in humans) chromosomes not involved in sex determination – i.e. that are alike in male and female cells.

**ortholog, orthology** – genes in different species which are not the results of gene duplication are orthologs, they are said to be orthologous and show orthology to each other.

**paralog, paralogy** – genes that are related by gene duplication, whether in the same species or in different species.

**pericentromeric** – in the region adjacent to the centromere

**polyploidisation** – any process by which a polyploid, an organism with more than two sets of homologous chromosomes, is formed.

**segmental duplication** – any duplicated piece of DNA, whether it includes a gene or not, in contrast to a gene duplication.

**subtelomeric** – in the region just inside the telomere, at the end of eukaryotic chromosomes

**tandem duplication** – duplication event in which two identical chromosome segments are formed adjacently.

**figure 1** – The age distribution of human duplicate genes (from Cotton and Page, 2005).

**figure 2** – Approximate age distribution of human segmental duplications of at least 1kb length and 90% identity, assuming a strict molecular clock for silent-site substitutions. Data is from human genome segmental duplication database, <http://projects.tcag.ca/humandup/>.

**figure 3** – How duplicated genes promote genome rearrangement through homologous recombination. (a) Recombination between duplicate loci on homologous chromosomes can lead to further duplication and loss, where duplicates are in the same orientation. (b) Duplicates in opposite orientations can produce inversions by intra-chromosomal recombination, while (c) recombination between duplicates on non-homologous chromosomes can lead to translocation of material between them.





