*Journal of Structural and Functional Genomics* **3:** 95–110, 2003.



# The 2R hypothesis and the human genome sequence

Karsten Hokamp[†], Aoife McLysaght[†] & Kenneth H. Wolfe[*]
*Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland*;
[*]*Author for correspondence: E-mail: khwolfe@tcd.ie*
[†]*These authors contributed equally to this work.*

Received 02.04.2002; accepted in final form 29.08.2002

*Key words:* human genome, paralogon, polyploidy, 2R hypothesis

**Abstract**

One theory formalised in 1970 proposes that the complexity of vertebrate genomes originated by means of genome duplication at the base of the vertebrate lineage. Since then, the theory has remained both popular and controversial. Here we review the theory, and present preliminary results from our analysis of duplications in the draft human genome sequence. We find evidence for extensive duplication of parts of the genome. We also question the validity of the 'parsimony test' that has been used in other analyses.

## The 2R hypothesis

In his 1970 book Susumu Ohno proposed that there may have been one or more whole genome duplications in the lineage leading to vertebrates. He postulated that genome duplication in the vertebrate lineage provided a platform for increasing the sophistication of the vertebrate genome and thus increasing morphological complexity. Genome duplication may be particularly powerful because all genes in a biochemical pathway will be duplicated simultaneously. Ohno was not specific about how many events occurred. The most popular form of this hypothesis is that there were **2 R**ounds of genome duplication early in the vertebrate lineage, as proposed by Holland *et al*. (1994). This has recently become known as the 2R hypothesis, an abbreviation attributable to Hughes (1999). There is no absolute consensus on the timing of these events, but the majority of references in the literature put one of these events immediately before, and one immediately after the divergence of agnathans from the lineage leading to tetrapods (see Figure 1 in Skrabanek and Wolfe, 1998). These timings are speculative and were probably chosen to coincide with major evolutionary transitions that they were thought to have facilitated. The lower limit on the timing of genome duplication is set by the observation of only a single Hox cluster in the invertebrate chordate amphioxus compared to four clusters in vertebrates (Garcia-Fernandez and Holland, 1994). As an upper limit, it seems unlikely that genome duplications would be viable in the mammalian lineage. Theory predicts that a genome duplication in an organism with a chromosomal basis of sex-determination (such as that of mammals) will result in sterility of the heterogametic sex, and thus inviability (Muller, 1925). Indeed the only known tetraploid mammal, a South American rodent, has duplicated copies of every chromosome except the sex chromosomes (Gallardo *et al*., 1999).

At the time of writing his book there was little evidence to support Ohno's claim. Very few protein sequences were known, and the hypothesis was based largely on genome size comparisons and matching patterns of cytogenetic bands. Much of the evidence which prompted Ohno to suggest a genome duplication event has lost merit in the light of our current understanding of genetics and genomes (Skrabanek and Wolfe, 1998). For example, differences in genome sizes are largely due to increased amounts of non-coding DNA rather than an increased number of genes; and cytogenetic bands, whose patterns were

used to list human chromosomes in pairs (Comings, 1972), are not indicative of the underlying gene content.

The debate on the 2R hypothesis to date has been a war of words (and limited data) between the phylogeneticists and the cartographers. As a general rule, analyses based on phylogenetic methods come out against the genome duplication hypothesis (e.g., Hughes, 1998; Hughes, 1999; Martin, 1999; Hughes *et al*., 2001; Martin, 2001), whereas map-based studies come out in favour (e.g., Lundin, 1993; Spring, 1997). Two main arguments have been advanced to support the theory of genome duplication in an early vertebrate: that there should be four vertebrate orthologues of each invertebrate gene, the so-called 'one-to-four rule' (Spring, 1997; Meyer and Schartl, 1999; Ohno, 1999); and that paralogous genes are clustered in a similar fashion in different regions of the genome (e.g., Martin *et al*., 1990; Lundin, 1993).

## The one-to-four rule

The one-to-four rule was first proposed by Jürg Spring (1997). He listed human paralogues present on different chromosomes and their *Drosophila* orthologues, and surmised that the maximum ratio of human to *Drosophila* genes was four. These 'tetralogues' seemed to bear the hallmark of a genome-wide event because they were discovered on all 23 female human chromosomes. The observation of some gene families with ratios of 2:6 or 2:5 *Drosophila*:human genes contradicts this hypothesis and Spring suggested that more complete genome sequences would provide data that could split these families into 'tetrapacks'.

The first extensive examination of the one-to-four rule using almost complete proteomes from *D. melanogaster*, *C. elegans*, and human, showed no excess of four-membered vertebrate gene families (see Fig. 49 of International Human Genome Sequencing Consortium [2001] and Fig. 12 of Venter *et al*. [2001]). Furthermore, the observation of gene families with five or more members directly contradicts the expectations of Spring (1997) that membership would be 'maximally four'. It appears that the one-to-four rule is an over-simplification of the history of the vertebrate genome. These data can of course be explained by hypothesising two genome duplications on a background of independent gene duplication and loss. However, as it is impossible to distinguish genome duplication from gene duplication on the basis of gene family size alone, this measure is simply uninformative.

## Paralogous chromosomal segments

The analysis of paralogous regions of the human genome is based on the assumption that, although it is expected that many rearrangements will have occurred in the time since the two duplication events envisaged by the 2R hypothesis, there should still be detectable remnants of the 4-way paralogy between some chromosomes, *i.e.*, some portions of some chromosomes should remain almost intact in four copies. This principle seems reasonable, though these studies have suffered for want of extensive genomic data. Finding as few as two genes in several linked clusters in a genome of over 30,000 is hardly overwhelming evidence for a genome duplication event (e.g., Martin *et al*., 1990). Objections that these observations can easily be explained by regional duplications of segments of chromosomes must be entertained.

### *HSA 1, 6, 9, and 19*

The observation of paralogous regions (around the MHC locus) on human chromosomes 1, 6, 9, and 19, led to the suggestion that these were duplicated by whole genome duplication events at the base of the vertebrate lineage (Kasahara *et al*., 1996; Katsanis *et al*., 1996; Kasahara, 1997). This was further supported by the finding of only a single related cluster in amphioxus (Flajnik and Kasahara, 2001). Ten members of particular gene families are present on chromosomes 6 and 9, and four of these are also represented on chromosome 1. The claim that this arrangement resulted from several rounds of polyploidy was refuted by Hughes (1998) using phylogenetic analysis of the nine families with sufficient data (Retinoid X receptor (RXR); $\alpha$ pro-collagen (COL); ATP-binding cassette (ABC) transporter; Proteasome component $\beta$ (PSMB); Notch; Pre-B-cell-leukemia transcription factor (PBX); Tenascin (TEN); C3/C4/C5 complement components; Heat shock protein 70 (HSP70)). However, Hughes' analysis did indicate that this arrangement could be partly due to block duplication. Trees of these families showed that five (RXR, COL, PBX, TEN, C3/4/5) of the nine families with sufficient phylogenetic information could have duplicated simultaneously, and that this timing was

consistent with a duplication in early vertebrate history 550–700 Mya. The phylogenetic analysis indicated that the four genes on chromosome 1 probably duplicated as a block. Similarly, a phylogenetic analysis by Endo *et al*. (1997) rejected the hypothesis that the 11 gene pairs on chromosomes 6 and 9 were duplicated in a single event, but did support the simultaneous duplication of six of the pairs. However, analysis of the remaining genes showed that the ABC transporter genes diverged before the origin of eukaryotes, the PSMB and the HSP70 gene families both originated before the divergence of animals and fungi, and the Notch genes diverged before the origin of deuterostomes (Hughes, 1998). Obviously these gene families did not arise as part of a block duplication event at the base of the vertebrate lineage. However, it can still be argued that these results are consistent with block duplication of this region if one assumes that there was an ancient tandem duplication of some of these genes, and after block duplication there was differential loss of one of the tandems, so that the divergence date of paralogues on two different chromosomes is that of the tandem duplication event rather than of the block duplication event (Kasahara *et al*., 1996; Smith *et al*., 1999).

*HSA 4, 5, 8, and 10*

Pébusque *et al*. (1998) reported the presence of paralogous genes on human chromosomes 4, 5, 8, and 10. In contrast to the analysis of the genes around the MHC discussed above, this study was based on a combination of phylogenetic and map-based methods. These genes are linked on the human chromosomes, with the exception that there is one family member on each of chromosomes 2 and 20, which require genome rearrangements to be reconciled with a block duplication event. The phylogenetic analyses consistently showed that these gene family members diverged in the vertebrate lineage and so are consistent with the 2R hypothesis of genome duplication. This conclusion was criticised by Martin (1999) who pointed out that the gene trees of the ankyrin family and the EGR (early growth response) family indicated different histories for their host chromosomes. The ankyrin gene tree groups chromosome 4 and 10 to the exclusion of chromosome 8, whereas the EGR gene tree groups 8 and 10 to the exclusion of all others. This contradicts the expectation that the family members on each chromosome have had a shared history since the block duplication event.

*HSA 2, 7, 12, and 17*

The quadruplication of the Hox cluster is the touchstone of the 2R hypothesis. There are four colinear Hox clusters in the vertebrate genome (Kappen *et al*., 1989), but only one in the invertebrate chordate amphioxus (Garcia-Fernandez and Holland, 1994). Phylogenetic analysis of the clusters showed that they duplicated early in vertebrate history. It seems certain that these clusters duplicated *en bloc*. The question is whether they arose by genome duplication events, or by sub-genomic duplication events, or a mixture of both. In the analysis of Zhang and Nei (1996) Hox clusters C and D were grouped with a high bootstrap, but there is not enough information in the alignments of the 61 amino acids of the homeodomain to resolve the phylogeny further. Instead, Bailey *et al*. (1997) analysed the relationship of the linked fibrillar-type collagen genes, which presumably shared the same duplication history. Assuming the collagen genes have a shared history with the Hox clusters, then the results can be interpreted as a topology (outgroup(HoxD(HoxA(HoxB,HoxC)))), which contradicts the grouping of HoxC and HoxD found by Zhang and Nei (1996). Furthermore, this is contrary to the expectations of the 2R hypothesis, which predicts a symmetric topology, but may be explained by three rounds of genome duplication with loss of 4 clusters, or by independent cluster duplications (Bailey *et al*., 1997).

In a phylogenetic analysis of the human Hox-bearing chromosomes (2, 7, 12, 17) Hughes *et al*. (2001) examined 35 gene families with members on at least two of the Hox chromosomes. 15 of these families could be classified as either pre-vertebrate, or post-mammalian duplicates and so are inconsistent with the 2R hypothesis. For the remaining 17 gene families the tree topologies did not exclude duplication at the same time as the Hox clusters. There were 15 of these for which the molecular clock was not rejected and estimates for the divergence dates of these gene families were calculated. Six of the gene families were dated to within the time of divergence of the Hox clusters, 528–750 Mya (as defined by lineage divergences), and two others had divergence estimates that were not significantly different from the time of Hox duplication. Phylogenies of gene families with members on at least three of the four Hox bearing chromosomes did not reveal a common topology for the relationship of these chromosomes.

*Other regions*

Some of the supposed paralogous regions of the vertebrate genome that can be found listed in the literature are based on rather sparse evidence. For example Gibson and Spring (2000) list human chromosomes X, 4, 5, and 11 as a possible paralogous quartet based only on the presence of members of two gene families (alpha-amino−3-hydroxy−5-methyl−4-isoxazole-propionic acid (AMPA) and androgen / mineralocorticoid / glucocorticoid / progesterone nuclear receptors) on all of these chromosomes.

## Testing the (AB)(CD) topology prediction

In its simplest form, the hypothesis of two rounds of genome duplication predicts a symmetric (A,B)(C,D) phylogenetic tree topology (where A, B, C, D, represent any four-membered gene family), with the age of the AB split the same as the age of the CD split, thus displaying the history of successive genome duplications. The alternative hypothesis, that of sequential gene duplication, will not always predict a symmetric topology. Under a sequential duplication model a four-membered family must arise from the duplication of one member of a three membered family. There is only one possible topology for three sequences, namely (A(C,D)) (Figure 1). Duplication of gene A will result in a symmetric topology, and duplication of either C or D will result in an asymmetric topology. Assuming that all three genes are equally likely to be duplicated, sequential gene duplication will give rise to a symmetric (A,B)(C,D) topology 1/3 of the time, and an asymmetric topology (A((B,C)D)) or (A(C(B,D))) the remaining 2/3 of the time. Thus, the null hypothesis is that the symmetric topology should be found in 1/3 of trees, and not 1/5 as proposed by Gibson and Spring (2000).

Hughes (1999) and Martin (2001) employed similar methodologies to test the phylogenies of gene families listed as exemplars of the one-to-four rule (Sidow, 1996; Spring, 1997) for congruence with the 2R hypothesis (*i.e.*, whether or not they displayed a symmetric topology, and if they duplicated in the vertebrate lineage). The symmetric topology was only observed in a small minority of the cases (one out of nine trees in Hughes [1999]; and two out of ten trees in Martin [2001]), although in Martin's analysis seven of the eight minimum-length trees that were not symmetric were not significantly shorter than a symmetric tree.

Variations on the 2R hypothesis result in different predictions for the phylogenies of vertebrate gene families. For example, if vertebrate genome doubling occurred by allopolyploidy (*i.e.*, hybridisation of two species, as has been suggested; (Spring, 1997)) or by segmental allopolyploidy (*i.e.*, behaving as an autopolyploid at some loci, and as an allopolyploid at others) then a single genome doubling event will produce paralogues with two different coalescence dates (Gaut and Doebley, 1997; Wolfe, 2001). Alternative models hypothesise that the two rounds of genome duplication may have occurred in short succession and thus not allowing the diploidisation procedure time to complete before the second genome duplication event. This would result in some tetrasomic loci, and some octasomic loci, in the quadruplicated genome (Gibson and Spring, 2000).

## Diploidisation

Diploidisation is a natural consequence of polyploidy. With some rare exceptions (*e.g.*, some loci of recent salmonid tetraploids; Allendorf and Thorgaard, 1984) all hypothesised paleopolyploid genomes have reverted to disomic inheritance at all loci. There is an increased incidence of non-disjunction of chromosomes when they form multivalents rather than divalents, so selection for increased fertility probably causes the reinstatement of disomic inheritance (Allendorf and Thorgaard, 1984).

Immediately after autotetraploidy all loci in the genome will be tetrasomic. These duplicated genes will not separate into two independently diverging loci until disomic inheritance is established (Ohno, 1970). This is important for our interpretation of what a paleopolyploid genome should look like because one of the properties we test in assessing genome duplication is the synchronicity of divergence of duplicated loci. Depending on the manner and speed of diploidisation this may or may not be an appropriate test for a paleopolyploid genome. In a diploid organism, chromosomes are arranged in pairs at meiosis (*i.e.*, chromosomes are bivalent). These pairs can exchange segments of DNA by recombination, and drift and gene conversion maintain a high degree of similarity between most alleles. In a tetraploid genome, chromosomes are arranged in tetravalents,

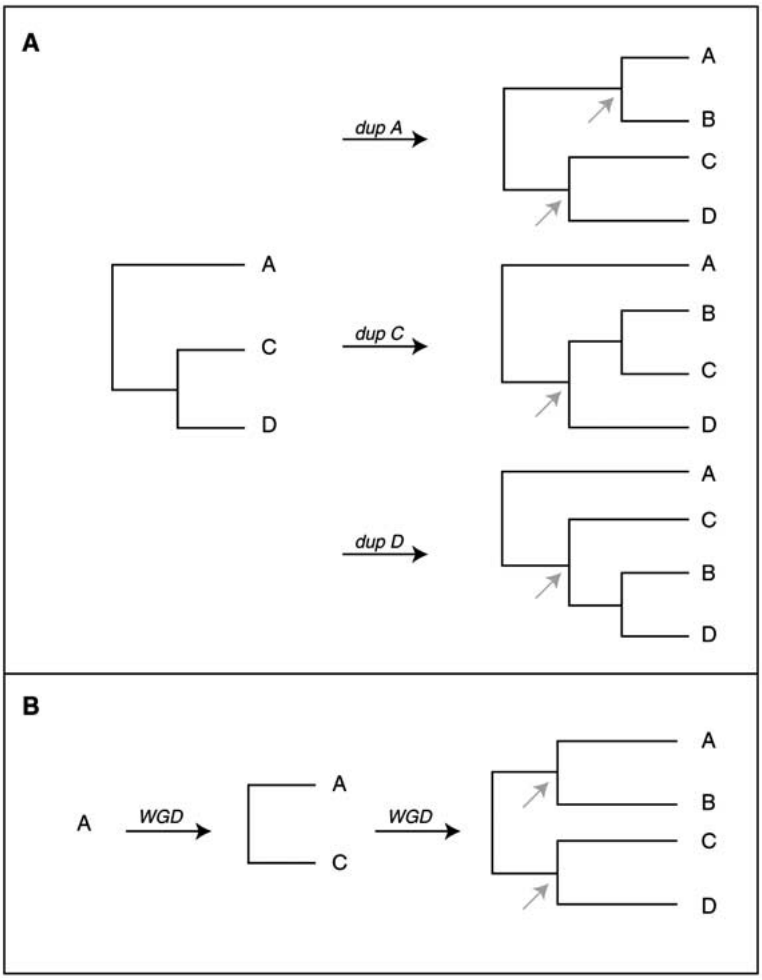


*Figure 1*. Alternative phylogenetic tree topologies of four-membered families resulting from sequential gene duplication or genome duplication. Grey arrows indicate the nodes that are critical to define the symmetry or asymmetry of the topology. (A) Phylogenetic tree topologies resulting from duplication of one member of a three-membered gene family. Three different trees result. The tree from the duplication of gene C and that from the duplication of gene D have asymmetric topologies. (B) Phylogenetic tree topologies resulting from two whole genome duplication (WGD) events. All genes are duplicated at each step, resulting in a symmetric tree topology.

rather than pairs, at meiosis. Diploidisation can be reduced to a problem of chromosome association. By what mechanism does a genome convert from forming chromosome quartets to forming chromosome pairs, *i.e.*, from tetraploid to diploid behaviour?

The answer to this question probably lies in a deeper understanding of the mechanisms of chromosome association. Is chromosome sequence divergence a cause or a consequence of diploidisation? If chromosome association occurs by homologous sequence attraction, then sequence divergence (by chromosome rearrangements) will cause diploidisation of chromosomes. On the other hand, if chromosome association is controlled by some other mechanism, such as attraction of homologous centromeres or telomeres, then chromosomal rearrangements may allow the independent evolution of the relocated loci and their previous partners in a tetrasomic locus, as separate loci without actually causing the diploidisation of the chromosomes in question.

The mammalian Y chromosome may serve as a model for this process. It is an unusual chromosome because it is partially diploid (at the pseudoautosomal region), and the rest is haploid. Lahn and Page (1999) identified homologous genes on the human X and Y chromosomes, which would have been part of the same locus when these chromosomes behaved autosomally (the sex chromosomes are thought to have

evolved from autosomes; (Graves, 1996)). They measured the amount of divergence at synonymous sites ($K_s$) between homologous gene pairs. From this they found that the homologues were in four age classes arranged sequentially along the X chromosome. They interpreted this as the result of inversions of large sections of the Y chromosome, leaving the X intact, which had the effect of suppressing recombination between these portions of the chromosomes. These chromosomes have diverged substantially, and most of the Y chromosome loci are haploid. The X and Y chromosome still pair at meiosis (at the pseudoautosomal region), and thus behave like diploid chromosomes, yet most of the loci are haploid. It may be the case that chromosomal tetravalency and locus tetrasomy can be separated in the same way.

The wheat genome (*Triticum aestivum*) is hexaploid, with three contributory genomes (*A*, *B*, and *D*). There is evidence for genetic control of chromosome association in wheat through the *Ph1* locus on chromosome V of the B genome (Riley and Kempanna, 1963). In the presence, but not the absence, of a particular allele of this locus, non-homologous associated centromeres separate at the beginning of meiosis (Martinez-Perez *et al*., 2001). The *Ph1* locus probably acts to amplify the differences between non-homologous chromosomes.

The most widely accepted hypothesis is that diploidisation proceeds by structural divergence of chromosomes. Allendorf and Thorgaard (1984) discuss a model whereby some loci may appear disomic while others apparently segregate tetrasomically. In their model they assumed that chromosome pairing occurred at the telomeres, but it can easily be modified to assume centromere association as indicated from the wheat study (Martinez-Perez *et al*., 2001). The model of residual tetrasomic inheritance hypothesises that there are two stages of chromosome pairing. The first stage will allow pairing between homoeologous chromosomes (partially similar chromosomes), thereby allowing recombination events between paralogous loci on different chromosomes. The second stage of pairing in this hypothesis resolves non-homologous chromosome pairing, and ensures that each gamete receives one copy of each chromosome in the normal manner. Evidence in support of this model comes from the observation of Martinez-Perez *et al*. (2001) that some non-homologous centromeres are associated just before the beginning of meiosis. This model predicts that loci closer to the point of association of the chromosomes (*i.e.*, closer to the centromere) will retain residual tetrasomic inheritance longer than others. For any locus, the likelihood that it behaves disomically rather than tetrasomically in a particular meiosis will be correlated with its distance from the centromere.

## Paralogon searches in the human genome sequence

In our laboratory we have begun to analyse the draft sequence of the human genome for evidence of ancient large-scale duplications, such as might be expected under the 2R hypothesis, and describe our approach here. The detection of paralogous chromosomal blocks (termed ‘paralogons’ by Popovici *et al*. (2001)) essentially involves the search for closely grouped sets of genes with homologues that occur in close vicinity in one or more other locations within the genome. Thus, the basic requirements for paralogon detection consist of the position and the sequence information of all genes. For a thorough and precise analysis the most complete and accurate set of human genes together with their map position is desirable. Further annotational data such as gene descriptions can add valuable background information, particularly about gene functions. A final version of the human genome sequence will probably not be available until 2003 (International Human Genome Sequencing Consortium, 2001), but several groups are striving to annotate the genome sequence in its current state. We have chosen to use data releases from Ensembl (Hubbard *et al*., 2002), a joint project between EMBL-EBI and the Sanger Institute which aims to develop a software system for automatic annotation of eukaryotic genomes, because it stands out in several respects:

– Ensembl employs an ‘open source’ philosophy, which allows public insight into every detail of the data assembly procedure. This can be very valuable to trace back steps that lead to decisions in the gene prediction process.
– Strategies and plans are discussed openly through a mailing list. This offers the chance to receive early information on data-related issues and also allows for interaction with the developers.
– The data are provided in the common SQL database format, which facilitates their integration into a local computer system.
– Frequent version releases provide a constant update of information.

– Most importantly, all data and programs are freely available and can be used without any restrictions.

In the genome annotation process, Ensembl integrates information of known proteins from SP-TREMBL (Bairoch and Apweiler, 2000). The GeneWise program (Birney and Durbin, 2000) converts these to exon structures on the human sequence. Additionally, new genes are detected *ab initio* through the GenScan program (Burge and Karlin, 1997). Functional annotation is derived from the InterPro, OMIM and SAGE databases. This results in a set of predicted and confirmed genes, the latter of which numbered 27,615 in Ensembl data release version 1.0. Since sequencing of the human genome is still in progress, this also affects the annotation process, which leads to regular data updates. Ensembl provides chromosomal locations for most of the confirmed genes through integration of mapping data from the 'Golden Path', an arrangement of BAC-cloned human sequences assembled and maintained by the University of California at Santa Cruz (Kent and Haussler, 2001). The underlying data comprise approximately 830 Mb of finished sequences and 2,300 Mb of draft sequences. A further 100 Mb had not been sequenced at the time of the data freeze. Data in the draft stage have only been sequenced once or twice and contain basepair ambiguities, gaps and segments with unknown order or orientation. High quality sequence data are expected from tenfold coverage.

*Sequence similarity search*

Establishing homology relationships among genes in an automated fashion is based on sequence similarity searches. The first step in the analysis therefore requires the comparison of all human proteins with each other. We included invertebrate proteomes (nematode and fly) in the search database to act as an approximate natural orthology threshold (any human proteins that are less similar than an invertebrate protein to the human query protein probably duplicated before the invertebrate-vertebrate lineage divergence, and so are not relevant to the 2R hypothesis). We carried out BLASTP searches, running on a 20-node Beowulf cluster, with the SEG filter to exclude low complexity regions. For organising the large volume of resulting query/hit pairings, we found the freely available MySQL database system very useful.

*ORF collapsing*

Tandem gene duplications are usually relatively recent evolutionary events and occur quite frequently. They can artificially inflate gaps between pairs of paralogues, and can inflate the number of hits reported between different chromosomes (or chromosomal regions). We therefore attempted to detect and collapse tandem duplicates. Similar to the method applied by Vision *et al*. (2000) in an analysis of the *A. thaliana* genome, all genes within a close neighbourhood that show strong sequence similarity to each other were removed from the map and replaced by a single representative, *i.e.*, the longest peptide of each group was retained. Any BLASTP hits to a gene that is part of a tandem array were 'redirected' to the single remaining gene representing the array.

*Parameter optimisation*

The paralogon detection process has to take into account evolutionary events like inversions, rearrangements, deletions and mutations, which are likely to obfuscate traces of genome duplication. A program was developed that is controlled by parameters to deal with gaps between pairs of duplicates, high-copy protein families, and the distinction between spurious similarities and true homologies. The values for these parameters greatly determine the outcome of the program. Too strict a set of values will mostly show highly conserved or recent blocks, where only closely grouped genes with very strong similarity are detected. By contrast, a relaxed definition of paralogons can lead to inflated block sizes and numbers as a result of inclusion of insignificant pairings. In the end a trade-off between selectivity and sensitivity must be chosen. We carried out extensive tests with different combinations of parameters to determine suitable parameter values.

## Paralogon detection results

A paralogon was 'built' starting from an anchor: a pair of homologous genes at different chromosomal locations. This was extended by including protein pairs on these chromosomes that were positioned no further than 30 genes distance from another protein included in the paralogon. Hits with a BLASTP expectation threshold higher than 1e−7 where excluded, as well as proteins with more than 20 hits.

The resulting paralogons range in sizes from only two pairs of duplicated genes to up to 29. Some of these paralogons, particularly the smaller ones, might have arisen from chance constellation of similar genes. To determine statistical significance, the paralogon detection method was applied to an artificial genome in which chromosome number and size has been retained but where genes have been assigned random locations. Using the same sets of parameters and repeating the shuffling and detection 1000 times allowed an estimation of significant block sizes. The results indicate that paralogons occur much more frequently in the real genome than expected by chance. Paralogons defined by at least six duplicated genes were in excess of 50 standard deviations more frequent in the real genome than expected from the simulations. The only alternative hypothesis that could fit these data is selection for clustering of these genes on a chromosome, as has been suggested for the mammalian MHC gene complex and the Surfeit locus (Hughes, 1999).

A web-based, interactive user interface was developed to allow navigation of duplicated regions and zooming between chromosomal and gene level. An example of the graphical presentation of a paralogon is shown in Figure 2. The web graphic contains integrated links that lead to more detailed information for genes and their similarity search results, as well as to external databases like Ensembl or GenBank. Only the paired duplicates are shown but intermittent genes can be switched on for closer inspection. The number of paralogues within a block lies typically between 9 and 23 percent of the genes that are covered by the region. This gives an indication of the large amount of evolutionary changes that happened since the occurrence of the duplication event.

Blocks with sizes of six or greater cover more than 44% of the human genome. Some of the largest regions are found on the four Hox chromosomes 2/7/12/17. This is also one of the few examples where duplications among four locations were found. The graphical overview of these chromosomes shown in Figure 3 proves the effectiveness of the block detection algorithm: the paralogons exactly cover the position of the Hox clusters, which are often used as a prime example of duplications within the human genome. Additionally, the covered areas expand previously reported regions and indicate more extensive duplications than formerly estimated.

Hughes *et al*. (2001) recently reported widely differing duplication dates for 42 gene families having members on the Hox-bearing chromosomes. Comparison of their results to ours shows that 139 of the 175 genes used in their study were present in our genome dataset, but only 31 of these genes (and a further 9 associated tandem repeats) form links that make up our paralogons; the other possible pairs were removed by our chordate-specific filters. Our paralogons (containing three or more duplicated genes) on the Hox chromosomes include a total of 426 duplicated genes (*i.e.*, 395 genes not included in Hughes *et al*.). There is no disagreement between the two sets of observations; chromosomes 2/7/12/17 contain some large paralogons formed by chordate-specific duplications, as well as many members of gene families formed by older duplications.

Beyond known examples, our analysis also uncovered interesting new duplications such as a region shared between chromosomes 8/14/16/20 in which copines and matrix metalloproteinases are found in close vicinity. Figure 4 provides a detailed view of the core areas of the duplicated regions. Copines form a recently discovered family of calcium-dependent, phospholipid-binding proteins that are suggested to be involved in membrane trafficking (Tomsig and Creutz, 2000). The metalloproteinases found in their vicinity are classified as the transmembrane subtype of the membrane-type MMPs (MT-MMPs) (Sato *et al*., 1997; Kojima *et al*., 2000). A literature search produced no results that indicate a connection between these two groups, which is not surprising, considering that research on copines is in its early stages. Their colocation in the same blocks, together with their membrane-association, suggests some kind of relationship, in particular because the copine and transmembrane MT-MMPs gene families are both small. These highly significant results provide an interesting base for a separate research project.

*Comparison with Celera data*

The only equally comprehensive report on paralogous blocks in human so far can be found in a study which is part of the private-sector human genome project led by Celera (Venter *et al*., 2001). A version of the program MUMmer (Delcher *et al*., 1999), modified to align protein sequences, was used to carry out intragenomic comparison based on the Celera sequence data. Results are presented as one large graphic (Fig. 13 in Venter *et al*., 2001), which shows paralogous regions for each chromosome. Blocks were defined by at least three linked genes. Due to

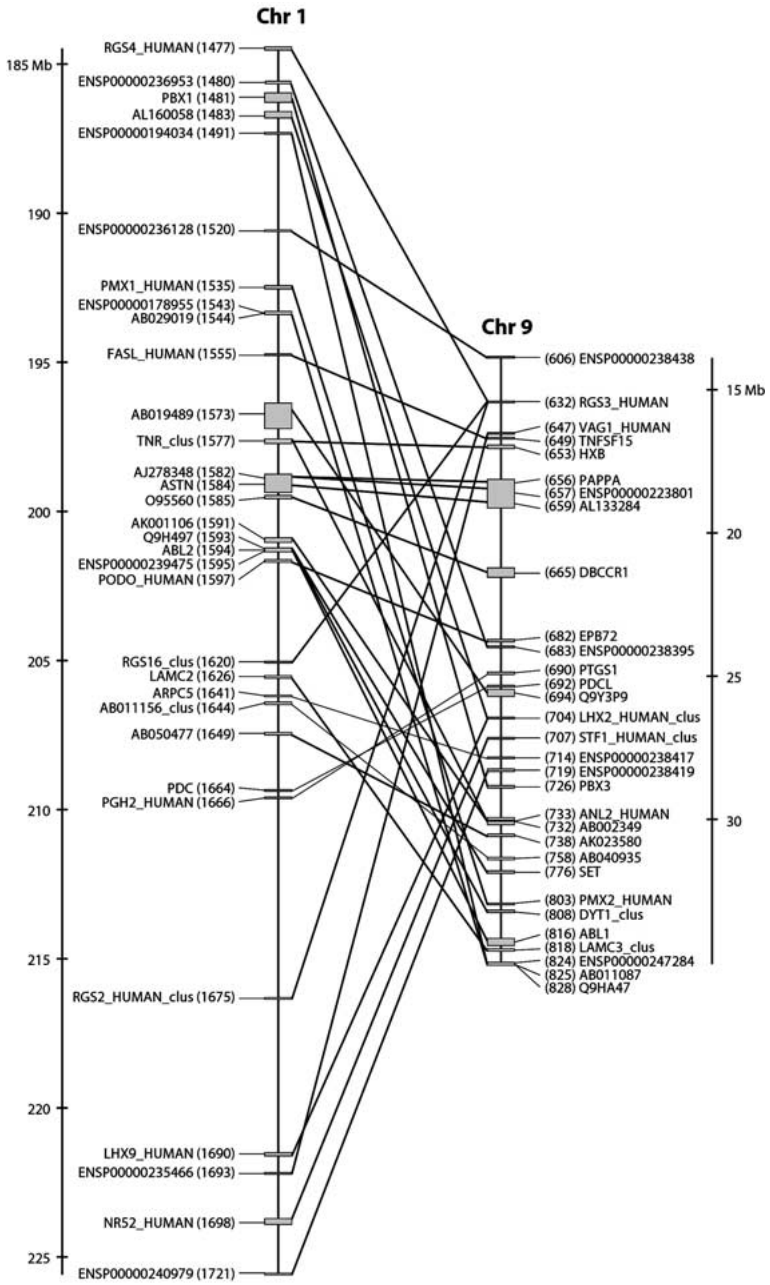


*Figure 2*. Paralogous block between human chromosomes 1 and 9. This is the largest paralogous block detected in the human genome including 29 duplicated genes. Blocks can be viewed at wolfe.gen.tcd.ie/dup.

lack of details a comprehensive comparison with our paralogons is not possible. The only feasible method consists of counting the presence or absence of links between each pair of chromosomes for both data sets: Of the 276 possible chromosome pairs, our method detected 151. We detected 55 regions that were not found in the analysis by Venter *et al*., and we did not detect any relationship between 21 pairs of chromosomes for which they illustrated pairings. Pairings between chromosomes 18 and 20 were provided in more detail and are shown in Figure 5 together with the corresponding blocks detected by our method. The overall appearance of cross-links seems to be the same except for a region near the centre of chromo-

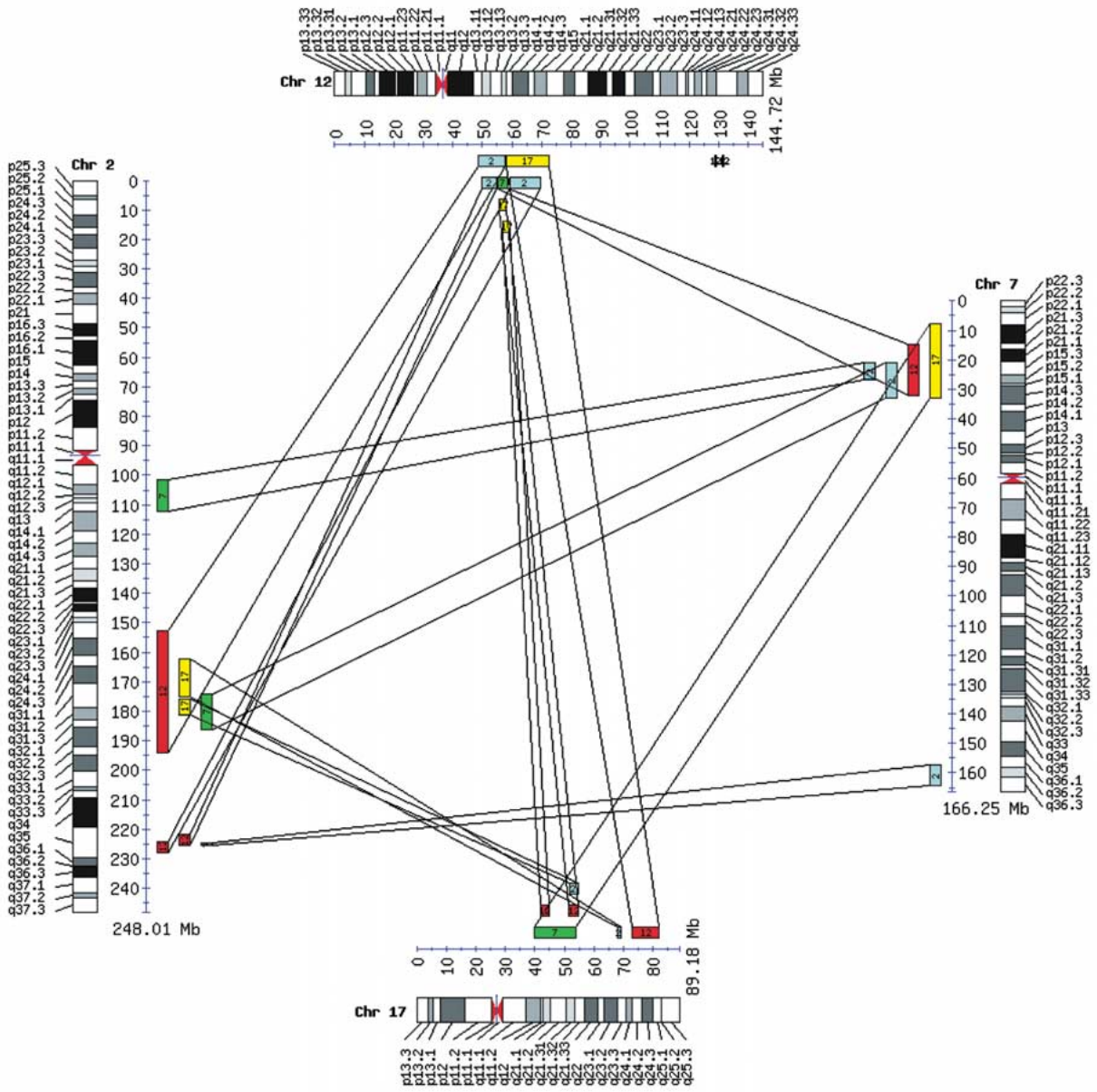


*Figure 3*. Blocks between Hox chromosomes. All detected paralogous blocks containing 6 or more duplicated genes between human chromosomes 2, 7, 12 and 17 are shown. Blocks can be viewed at wolfe.gen.ted.ie/dup.

some 20. The segment from gene 'GATA rel' to 'Krup rel' on chromosome 20 in the MUMmer graph might correspond to the far end of chromosome 20 ($>$ 64 Mb) in our graph, because both seem to be connected to roughly the same region on chromosome 18. A translocation such as this could occur from differences in the assembly process. Large discrepancies exist in the underlying data: Celera reports 217 protein assignments on chromosome 18 and 322 on chromosome 20. This corresponds to 388 and 748 proteins in the Ensembl data. Venter *et al*. state that their analysis found 64 protein pairs in the blocks between chromosome 18 and 20, and that these blocks have a duplicate gene density of 20–30%. In our case four blocks of sizes 6, 7, 7 and 8 are detected which link a total of 29 and 28 genes on chromosome 18 and 20, respectively. The density of involved genes ranges from 12–39% with a median at 19.7%. Unfortunately,

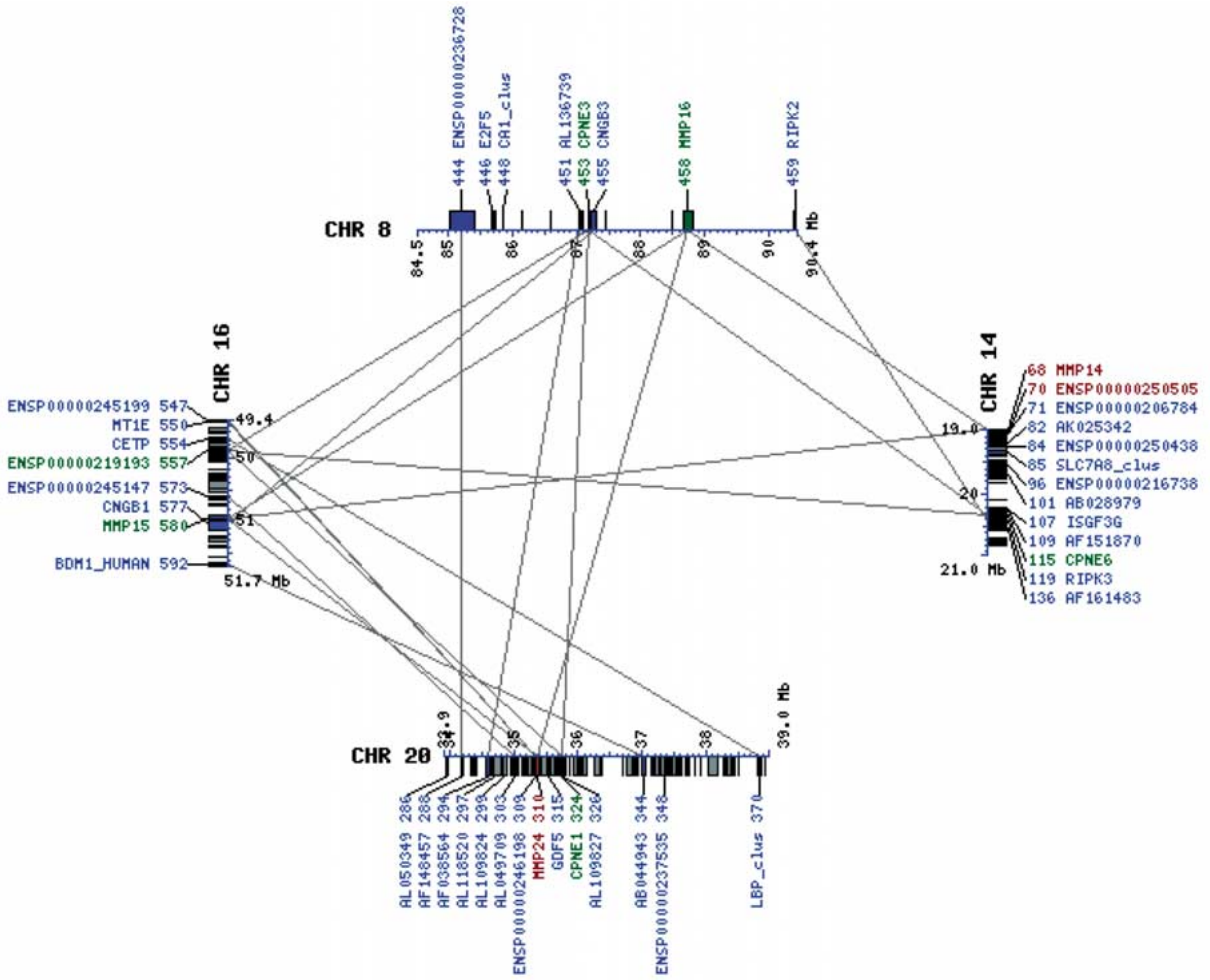


*Figure 4*. Blocks with copines and MMPs. The core areas of duplicated regions between human chromosomes 8, 14, 16 and 20 are shown. A copine (CPNE) and a matrix metalloproteinase (MMP) can be found in each of them in close vicinity (protein ENSP00000219193 on chromosome 16 is predicted to be a copine by the Ensembl automated annotation system). Linked genes are drawn and labelled in blue. Genes with links to two or three other blocks are highlighted in red and green, respectively. Intermediate genes are filled with grey.

only 7 of the reported gene names (TGIF, VAPA, VAPB, NFATC1, NFATC2, KCNG2, KCNB1) match between the two data sets so a more detailed comparison was not possible. A possible explanation for differences between both graphs can be found in the recent finishing of chromosome 20 by Deloukas *et al.* (2001), where discrepancies between the private and the public sequence affecting order of genes and chromosomal blocks were discussed.

## A question of parsimony

One way in which the genome duplication hypothesis is more parsimonious than alternative hypotheses that explain the distribution of paralogues in the genome is in the number of words it takes to describe it, a fact which may be related to its popularity as a hypothesis. Austin Hughes has challenged the assumption that block duplication is the most parsimonious way to generate paralogous regions within a genome using a parsimony statistic. The statistic considers the relative parsimony of the hypothesis that paralogous regions were made by a block duplication event (perhaps as part of a whole genome duplication event), or the alternative hypothesis that they are the result of tandem duplication of genes followed by translocation (Hughes, 1998; Hughes *et al*., 2001). Following Gu and Huang (2002) we refer to these as the ‘BD’ (block duplication) model and ‘TD’ (tandem duplication)

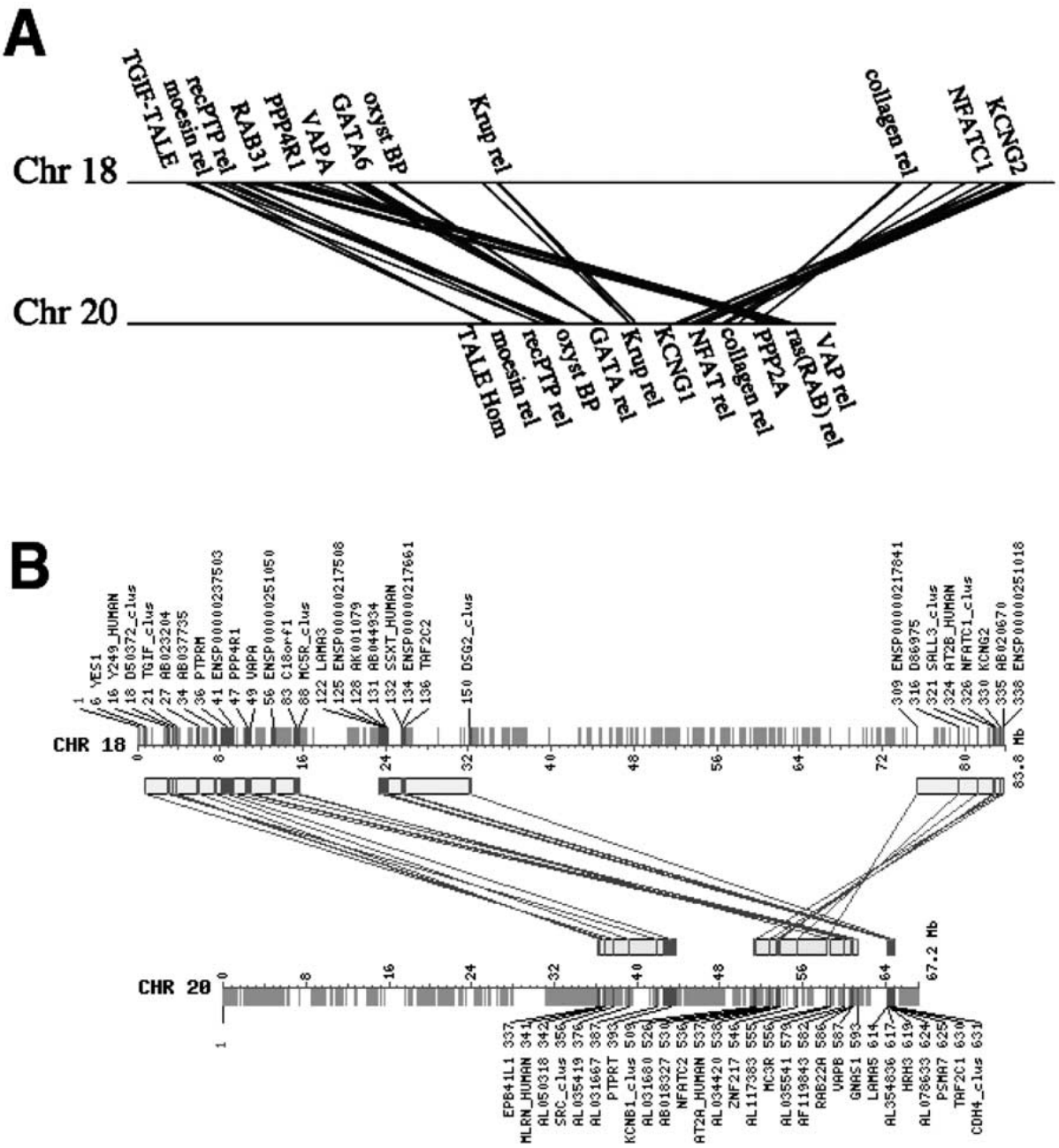


*Figure 5*. Paralogous regions between human chromosomes 18 and 20 defined by two different methods. Lines drawn between the chromosomes connect paralogues. (A) From Venter *et al*. (2001), showing labels for a selection of genes. (B) From our analyses with all duplicated genes labelled.

model, respectively. Hughes found for both the Hox cluster regions and the chromosome 1/6/9/19 region that the TD model was more parsimonious than the BD model as an explanation for the observed gene orders and phylogenetic trees. However, his reasoning may have been flawed as shown below. The applicability of the parsimony statistic to the paleopolyploid *Arabidopsis* genome has also been challenged by Gu and Huang (2002).

Here we consider the simple case of a single genome duplication. The BD model has an inbuilt disadvantage in the parsimony count method of Hughes (1998; 2001). Each gene that is no longer present in duplicate is counted as an individual deletion event (or equally, as a single translocation event removing it from the scope of detection as part of a paralogous region). By contrast, the method is very generous to the TD model, assuming that a single translocation event brings each gene to its current position within a paralogous region. In fact, as shown below, Hughes' TD model will always require fewer events than a BD model, so long as fewer than 1/3 of genes are retained in duplicate.

Let $G$ be the number of genes in the pre-duplication genome. Let $q$ be the proportion of the pre-duplication genome retained in duplicate in the modern genome, and $p$ be the proportion in single copy.

$$\text{Then } q + p = 1 \tag{1}$$

$$\text{and } Gq + Gp = G \tag{2}$$

$Gq$ is the number of genes retained in duplicate.

The TD model requires $Gq$ tandem duplication events and a further $Gq$ translocation events, totalling $2Gq$ events. The BD model requires one large duplication event (in this example, a genome duplication) followed by $Gp$ gene deletion events (the number of genes seen in single copy). For these two hypotheses to have an equal number of events (*i.e.*, to be equally parsimonious) then:

$$2Gq = 1 + Gp \tag{3}$$

Replace $p$ with $1-q$ from Equation 1:

$$\rightarrow 2Gq = 1 + G(1 - q) \tag{4}$$

$$\rightarrow q = 1/3G + 1/3 \tag{5}$$

For genomes with a large number of genes (*e.g.,* $G = 6000$ for yeast):

$$\rightarrow q \approx 1/3 \tag{6}$$

The TD model will be more parsimonious than the genome duplication (BD) model if $q < 1/3$, *i.e.*, whenever the retention of genes in duplicate is less than 1/3 of the pre-duplication genome. This result is also apparent from the work of Gu and Huang (2002) who separately analysed 103 duplicated blocks in the *Arabidopsis* genome. Their Figure 2 shows empirically that the TD model is more parsimonious in the 94 blocks having $q < 1/3$, whereas the BD model is more parsimonious only in the remaining 9 blocks with $q > 1/3$. A retention frequency ($q$) of 1/3 corresponds to a duplication level of 50% in the post-duplication genome (*i.e.*, 50% of genes in the modern genome will have polyploidy-derived paralogues).

*Table 1.* Crossover values for $q$ (proportion of genes retained in duplicate) and $d$ (average number of genes deleted in a single event) at which the genome duplication and TD models are equally parsimonious (calculated from equation 9).

| $q$ | $d$ |
|---|---|
| 0.33 | 1.0 |
| 0.20 | 2.0 |
| 0.10 | 4.5 |
| 0.08 | 5.8 |
| 0.05 | 9.5 |
| 0.01 | 49.5 |

The above calculations were based on Hughes' assumption that genes are deleted individually, *i.e.,* that only one gene is deleted per deletion event. It may be more biologically realistic to allow for several neighbouring genes to be duplicated in a single event. If $d$ is the average number of genes deleted in a gene deletion event, then Equation 4 can be rephrased as:

$$2q = 1/G + (1 - q)/d \tag{7}$$

$$\rightarrow 2q \approx (1 - q)/d \tag{8}$$

$$\rightarrow d \approx 1/2q - 1/2 \tag{9}$$

Solving Equation 9 for different values of $q$ shows the average size of a deletion event that is required for the two hypotheses to have equal probability for different frequencies of duplicate gene retention (Table 1).

One of the observations of the well-documented case of paleopolyploidy in yeast was that only 8% of the pre-duplication genome was retained in duplicate (Seoighe and Wolfe, 1998). For $q = 0.08$ the average size of a deletion event ($d$) needs to be 6 genes or larger (Table 1) to favour the genome duplication hypothesis by the simple parsimony statistic. Intuitively this seems like a biologically feasible size. Indeed it seems more acceptable than another assumption built-in to the alternative tandem-duplica-

tion and translocation model, *i.e.*, that selection can create genomic regions with similar gene contents by favouring particular translocations (Hughes, 1999).

We consider an example from the yeast genome. Within a block first described by Pohlmann and Philippsen (1996), and later numbered block 39 by Wolfe and Shields (1997) there are six duplicated genes and 20 unduplicated genes (eight on chromosome XIV and 12 on chromosome IX). Under Hughes' TD model the formation of this block would require 12 steps (six tandem duplications, and six translocations). Under a whole genome duplication model, with each gene deleted individually, the formation of this block would require 21 events (one whole genome duplication, and 20 gene deletions) and would thus be less parsimonious by this statistic. However, if each deletion event included on average three genes then only seven deletion events would be required to explain the current state of this paralogous region, and the block duplication model would be more parsimonious. Thus it appears that, even in the well-documented case of yeast, which Hughes and colleagues agree is a likely polyploid (Friedman and Hughes, 2001), this simple parsimony statistic is not appropriate to determine the relative probability of paralogous region formation by block duplication versus tandem duplication and translocation.

## Discussion

The approach described here acknowledges two important aspects of the genome duplication hypothesis. Ones is that it proposes a whole genome event, and anecdotal evidence for individual gene families is unlikely to result in a firm conclusion – we have therefore analysed the whole human genome. The other important aspect is that genome duplication is not proposed as the only mechanism of gene family expansion – we have therefore removed obvious tandem duplicates from the analysis, and deliberately limited this study to look at the mechanisms of gene-family expansion in the vertebrate lineage. Nobody realistically expects that all gene families that exist in the vertebrate genome were singletons before vertebrate origins. What is detectable as a gene family, i.e., a group of paralogous genes, will often include members representing a long evolutionary history, only some of which is vertebrate specific. Studies where gene families are included indiscriminately will doubtless include diverse evolutionary histories, and closer inspection will reveal a straw man easy to knock down.

Our map-based method found paralogons covering over 44% of the human genome. These are most probably vertebrate specific, and are distributed throughout the genome. This can be explained by either extensive sub-genomic duplications, or by polyploidy. An additional phylogenetic analysis carried out recently on the same data set indicates a significant accumulation of duplication activity during a relatively short period between 350 and 650 Mya (McLysaght *et al.*, 2002). Both findings together lend support to the hypothesis of one polyploidy event early in the vertebrate lineage. There is no specific evidence for two as opposed to one tetraploidy event, or for auto- rather than allotetraploidy. Neither do current findings from investigations of the one-to-four rule or tree topologies give clear signals for any of these scenarios. Further refinement of the human genome sequence and complete gene identification will hopefully enable more precise analyses in the future. It is also likely that the genome sequencing project for the tunicate *Ciona* will contribute useful data in the near future. However, in view of the immense time span under consideration and the huge complexity of genomic changes that might have occurred, it remains unsure if the complete history of ancient duplication events that led to the current shape of the human genome will ever be revealed.

## References


Allendorf, F.W. and Thorgaard, G.H. (1984) Tetraploidy and the evolution of salmonid fishes. In *Evolutionary Genetics of Fishes* (Ed. Turner, B.), Plenum Press, New York, NY, pp. 1–46.

Bailey, W.J., Kim, J., Wagner, G.P. and Ruddle, F.H. (1997) Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.*, **14,** 843–853.

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28,** 45–48.

Birney, E. and Durbin, R. (2000) Using GeneWise in the Drosophila annotation experiment. *Genome Res.*, **10,** 547–548.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268,** 78–94.

Comings, D.E. (1972) Evidence for ancient tetraploidy and conservation of linkage groups in mammalian chromosomes. *Nature*, **238,** 455–457.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27,** 2369–2376.

Deloukas, P., *et al*. (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414,** 865–871.

.Endo, T., Imanishi, T., Gojobori, T. and Inoko, H. (1997) Evolutionary significance of intra-genome duplications on human chromosomes. *Gene*, **205,** 19–27.

Flajnik, M.F. and Kasahara, M. (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity*, **15,** 351–362.

Friedman, R. and Hughes, A.L. (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res*., **11,** 373–381.

Gallardo, M.H., Bickham, J.W., Honeycutt, R.L., Ojeda, R.A. and Kohler, N. (1999) Discovery of tetraploidy in a mammal. *Nature*, **401,** 341.

Garcia-Fernandez, J. and Holland, P.W. (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature*, **370,** 563–566.

Gaut, B.S. and Doebley, J.F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA*, **94,** 6809–6814.

Gibson, T.J. and Spring, J. (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans*., **28,** 259–264.

Graves, J.A. (1996) Mammals that break the rules: genetics of marsupials and monotremes. *Annu. Rev. Genet*., **30,** 233–260.

Gu, X. and Huang, W. (2002) Testing the parsimony test of genome duplications: a counterexample. *Genome Res*., **12,** 1–2.

Holland, P.W.H., Garcia-Fernandez, J., Williams, N.A. and Sidow, A. (1994) Gene duplications and the origins of vertebrate development. *Development*, **Suppl. 1994,** 125–133.

Hubbard, T., *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res*., **30,** 38–41.

Hughes, A.L. (1998) Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol*., **15,** 854–870.

Hughes, A.L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol*., **48,** 565–576.

Hughes, A.L., da Silva, J. and Friedman, R. (2001) Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res*., **11,** 771–780.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Kappen, C., Schughart, K. and Ruddle, F.H. (1989) Two steps in the evolution of Antennapedia-class vertebrate homeobox genes. *Proc. Natl. Acad. Sci. USA*, **86,** 5459–5463.

Kasahara, M. (1997) New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas*, **127,** 59–65.

Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T. and Ishibashi, T. (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA*, **93,** 9096–9101.

Katsanis, N., Fitzgibbon, J. and Fisher, E.M.C. (1996) Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics*, **35,** 101–108.

Kent, W.J. and Haussler, D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res*., **11,** 1541–1548.

Kojima, S., Itoh, Y., Matsumoto, S., Masuho, Y. and Seiki, M. (2000) Membrane-type 6 matrix metalloproteinase (MT6-MMP, MMP–25) is the second glycosyl-phosphatidyl inositol (GPI)-anchored MMP. *FEBS Lett*., **480,** 142–146.

Lahn, B.T. and Page, D.C. (1999) Four evolutionary strata on the human X chromosome. *Science*, **286,** 964–967.

Lundin, L.G. (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, **16,** 1–19.

Martin, A. (2001) Is tetralogy true? Lack of support for the 'one-to-four' rule. *Mol. Biol. Evol*., **18,** 89–93.

Martin, A.P. (1999) Increasing genomic complexity by gene duplication and the origin of vertebrates. *Amer. Nat*., **154,** 111–128.

Martin, G.R., Richman, M., Reinsch, S., Nadeau, J.H. and Joyner, A. (1990) Mapping of the two mouse engrailed-like genes: close linkage of En–1 to dominant hemimelia (Dh) on chromosome 1 and of En–2 to hemimelic extratoes (Hx) on chromosome 5. *Genomics*, **6,** 302–308.

Martinez-Perez, E., Shaw, P. and Moore, G. (2001) The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature*, **411,** 204–207.

McLysaght, A., Hokamp, K. and Wolfe, K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nature Genet.*, **31**, 200–204.

Meyer, A. and Schartl, M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol*., **11,** 699–704.

Muller, H.J. (1925) Why polyploidy is rarer in animals than in plants. *Amer. Nat*., **9,** 346–353.

Ohno, S. (1970) *Evolution by Gene Duplication*, George Allen and Unwin, London, UK.

Ohno, S. (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. *Semin. Cell Devel. Biol*., **10,** 517–522.

Pébusque, M.-J., Coulier, F., Birnbaum, D. and Pontarotti, P. (1998) Ancient large scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol*., **15,** 1145–1159.

Pohlmann, R. and Philippsen, P. (1996) Sequencing a cosmid clone of *Saccharomyces cerevisiae* chromosome XIV reveals 12 new open reading frames (ORFs) and an ancient duplication of six ORFs. *Yeast*, **12,** 391–402.

Popovici, C., Leveugle, M., Birnbaum, D. and Coulier, F. (2001) Coparalogy: Physical and functional clusterings in the human genome. *Biochem. Biophys. Res. Commun*., **288,** 362–370.

Riley, R. and Kempanna, C. (1963) The homeologous nature of the non-homologous meiotic pairing in *Triticum aestivum* deficient for chromosome V (5B) *Heredity*, **18,** 287–306.

Sato, H., Tanaka, M., Takino, T., Inoue, M. and Seiki, M. (1997) Assignment of the human genes for membrane-type–1, –2, and –3 matrix metalloproteinases (MMP14, MMP15, and MMP16) to 14q12.2, 16q12.2-q21, and 8q21, respectively, by in situ hybridization. *Genomics*, **39,** 412–413.

Seoighe, C. and Wolfe, K.H. (1998) Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl. Acad. Sci. USA*, **95,** 4447–4452.

Sidow, A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Devel*., **6,** 715–722.

Skrabanek, L. and Wolfe, K.H. (1998) Eukaryote genome duplication – where's the evidence? *Curr. Opin. Genet. Devel*., **8,** 694–700.

Smith, N.G.C., Knight, R. and Hurst, L.D. (1999) Vertebrate genome evolution: a slow shuffle or a big bang? *BioEssays*, **21,** 697–703.

Spring, J. (1997) Vertebrate evolution by interspecific hybridisation – are we polyploid? *FEBS Lett*., **400,** 2–8.

Tomsig, J.L. and Creutz, C.E. (2000) Biochemical characterization of copine: a ubiquitous $Ca^{2+}$-dependent, phospholipid-binding protein. *Biochemistry*, **39,** 16163–16175.

Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291,** 1304–1351.

Vision, T.J., Brown, D.G. and Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science*, **290,** 2114–2117.

Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genet*., **2,** 333–341.

Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387,** 708–713.

Zhang, J. and Nei, M. (1996) Evolution of Antennapedia-class homeobox genes. *Genetics*, **142,** 295–303.