



Common ground: relationships between different species are clearly written in their genes, but not in their chromosomes.

## All genomes great and small

Chromosome size and number can vary widely between closely related organisms. This poses a challenge for the evolutionary geneticists who are trying to make sense of genome structure, says Jonathan Knight.

**W**e are somewhat like bananas, at least as far as our genes are concerned. About a third of human genes are clearly related to those found in plants. Fruitflies, sitting closer to us on the tree of life, share about two-thirds of our genes. And mice, being fellow mammals, are at least 90% genetically similar to us.

But look at the size and structure of the chromosomes on which those genes sit, and the neat correlation between evolutionary relatedness and genomic similarity breaks down. Sheep have 27 pairs of chromosomes; the Indian muntjac deer has just 3. We have some 3 billion DNA base pairs; one species of amoeba has more than 600 billion.

It seems entirely random. But as the genomes of ever more organisms are sequenced in their entirety, trying to make sense of natural variability in genome structure has become a burgeoning field. "We're finally able to ask some key questions," says Daniel Hartl, an evolutionary geneticist at Harvard University.

Although it is still early days, comparisons of sequences from different species suggest that events such as bursts of activity of 'jumping genes', genetic duplications and chromosome fusions play a key role in evolution. Far from being a mass of junk DNA that holds a small but precious cargo of genes, geneticists are starting to see chromosomes as highly dynamic stages on which important evolutionary processes are played out.

### Jump to it

The mobile elements known as transposons are among the most powerful forces that shape chromosome evolution. These 'jumping genes' carry instructions for their own excision, duplication and insertion into the genome. It seems that, at certain periods in evolutionary history, transposon activity made chromosomes expand like accordions. As a result, most chromosomes are now filled with the silent remnants of transposons.

Evolutionary geneticists can work out how long ago a transposon became immobi-

lized by looking at the accumulation of mutations in its characteristic flanking sequences. Such studies have suggested that a flurry of transposon activity doubled the size of the maize genome from 1.2 billion to 2.4 billion bases in the past 3 million years<sup>1</sup>. In human evolution, transposons called long interspersed elements (LINEs) have expanded to about 100,000 copies in several bursts over the past 100 million years, the most recent event having occurred 25 million years ago<sup>2</sup>. LINEs now account for 15% of human DNA.

Why should mobile elements disperse themselves in bursts? One intriguing idea, says Hartl, is that most of the time cells repress transposon activity — a sensible strategy, given that a gene can be disabled if a transposon jumps into its sequence. But the costs and benefits may shift during periods of greater evolutionary stress. Increased rates of transposition might then be selected for, Hartl argues, because they may help organisms to adapt in tough times by increasing genetic variability.

	Genome size (base pairs)	Chromosome number ( <i>n</i> )
<i>Amoeba dubia</i>	670,000,000,000	Several hundred
Trumpet lily ( <i>Lilium longiflorum</i> )	90,000,000,000	12
Mouse ( <i>Mus musculus</i> )	3,454,200,000	20
Human ( <i>Homo sapiens</i> )	3,200,000,000	23
Carp ( <i>Cyprinus carpio</i> )	1,700,000,000	49
Chicken ( <i>Gallus gallus</i> )	1,200,000,000	39
Housefly ( <i>Musca domestica</i> )	900,000,000	6
Tomato ( <i>Lycopersicon esculentum</i> )	655,000,000	12

**What's in a number? Genome size and chromosome number seem unrelated to complexity. The figures above are for haploid genomes — most cells are diploid (2*n*), carrying two copies of each chromosome.**

Transposons aren't the only type of DNA that can be duplicated. If there is one thing DNA is good at, it is being copied. And as a result, duplications ranging from hundreds of bases to the cell's entire complement of chromosomes have figured heavily in the evolution of modern genomes.

The simplest duplications produce adjacent repeated sequences that are all oriented in the same way. The lengths of these 'tandem repeats' can vary greatly, and often entire genes are duplicated. The extra copy can then accumulate mutations; often these will render it useless, but a new and useful function can also emerge. Gene duplication is now widely considered to be the most likely origin of gene clusters, in which natural selection has shaped copies of one original gene to take on different functions. We owe our wide-ranging sense of smell, for instance, to the duplication and diversification of olfactory-receptor genes<sup>3</sup>.

Duplicated DNA need not end up close to its template. The human genome contains duplicated chunks of DNA hundreds of kilobases long at opposite ends of a chromosome or on different chromosomes altogether<sup>4</sup>. Evan Eichler, a genome researcher at Case Western Reserve University in Cleveland, Ohio, estimates that at least 5% of the human genome arose through this sort of duplication<sup>5</sup>. Indeed, the true figure may be much higher, as duplications in an ancestor that lived more than 40 million years ago would by now be undetectable because of the subsequent divergence of the copied regions.

Eichler and his colleagues have found that more than a third of duplicated regions that end up on another chromosome are found near the centromere, the anchor point for the



**Daniel Hartl** thinks deletion determines genome size.

protein filaments that yank freshly divided chromosomes apart during cell division. Centromeres consist mainly of tandem repeats, 171 bases long, known as  $\alpha$ -satellite DNA. This DNA tends to remain tightly packed in a structure called heterochromatin, where it is not

transcribed into RNA. But some centromeres in the human genome have an intermediate region between the satellite DNA and the adjacent, less closely packed 'euchromatin' that houses active genes. Eichler's group has found that duplicated segments may pop up as islands in a sea of  $\alpha$ -satellite DNA in a region known as the pericentromere<sup>6</sup>.

Some pericentromeres are part vacuum cleaner, part blender. They suck in duplications, chop them up and recombine them. As new insertions arrive, they drop in without respect for other insertions, sometimes splitting them in two. Once in a pericentromere, sections of duplicated DNA are clipped out, inverted and stuck in elsewhere. The result is a tossed salad of duplications — which Eichler suspects may serve as an incubator of new genes.

### Garbage or gold?

Many of the duplications have brought with them the regulatory sequences that allow them to be transcribed, and Eichler's team has detected RNA messages from pericentromeric regions<sup>7</sup>. Although most of the resulting transcripts are garbage, there is the possibility that one could occasionally code for a new and useful protein. So far, no example of a new gene being born this way has been found. But Eichler is examining the regions around a 'dead' centromere on human chromosome 2. Here there are higher rates of transcription than around the chromosome's active centromere, and Eichler hopes to find new and interesting genes.

Other researchers are studying analogous regions of repetitive DNA that sit adjacent to the telomeres that cap chromosomes' ends. Like pericentromeres, these 'subtelomeres'



**The yeast genome** was shaped by duplication, says Kenneth Wolfe.

seem to be zones of active genetic recombination. Researchers led by Barbara Trask of the Fred Hutchinson Cancer Research Center in Seattle have shown that sections of subtelomere some 50–100 kilobases long have moved from one chromosome to another in our species'

recent evolutionary history<sup>8</sup>.

Trask has also found several members of the family of olfactory-receptor genes hiding in human subtelomeres, at least one of which seems to be functional, being expressed in the olfactory epithelium<sup>9</sup>. Together, her discoveries suggest that, like pericentromeres, subtelomeres may be important regions for the mixing and matching of duplicated DNA to form new genes.

Intriguingly, in earlier studies carried out at the Lawrence Livermore National Laboratory in California, Trask found that most of the variability in overall genome size across human populations is due to variation in the sizes of pericentromeres and subtelomeres<sup>10</sup>. In the case of chromosome 21, this variation makes the chromosome 45% larger in some people than in others. The expansion of subtelomeres could be functionally and evolutionarily significant, Trask suggests.

### Doubling up

Such duplication and recombination events cannot explain the wide variety in chromosome number seen in nature. But individual chromosomes, and even entire sets, can also be duplicated. More than three decades ago, Susumu Ohno, a geneticist at the City of Hope Cancer Center in Los Angeles, proposed that genome duplication could account for the relatively rapid evolution of complexity in vertebrates<sup>11</sup>.

Most organisms are diploid — they carry two copies of each chromosome, one from their father, one from their mother. But an error in chromosome segregation can easily result in an individual being tetraploid, carrying twice the usual complement of chromosomes. Ohno argued that this was a key factor in vertebrate evolutionary history. Over time, mutations would accumulate in the duplicated chromosomes until they were so divergent that they were clearly distinct. The organism would then seem to be diploid once more — only now it would have twice the number of chromosomes.

Examples of genome duplication are easy enough to find, particularly in flowering plants, where only about half of all species are diploid. Wheat, for example, has six copies of each chromosome. And when plant biologists examined the genome sequence of the diploid



**Evan Eichler** is looking for new genes in a duplication salad.

thale cress (*Arabidopsis thaliana*), they concluded from extensive evidence of duplication that it must have evolved from a tetraploid ancestor that first arose around 112 million years ago<sup>12</sup>.

Similar polyploidy almost certainly figured in the evolution of the yeast *Saccha-*



The Indian (left) and Chinese (right) muntjacs have ended up with very different chromosome counts.

*romyces cerevisiae*, adds Kenneth Wolfe, a geneticist at Trinity College Dublin. In 1997, he reported that the sequences that flank the centromeres of *S. cerevisiae*'s 16 chromosomes are actually grouped into 8 pairs<sup>13</sup>. "What we are looking at are the evolutionary products of some kind of mistake in chromosome pairing that happened millions of years ago," says Wolfe. Although others have argued that Wolfe's observations could be the result of multiple duplications, Wolfe says that his unpublished analysis of other regions of the yeast genome suggests that it was a single, dramatic event.

### A theory with backbone

But what of Ohno's theory that genome duplications were critical to the evolution of vertebrates? The idea gained support in the late 1980s with the discovery in mice and humans of four nearly identical gene clusters that control the development of the body plan, known as the homeobox, or Hox, genes. Fruitflies — in which Hox genes were first discovered — and other invertebrates have only a single cluster, but the order of the genes within it is the same as in the mammalian clusters. So human Hox genes seem to be the product of two rounds of duplication of an ancestral Hox cluster<sup>14</sup>.

When geneticists discovered several other gene clusters that appear just once in invertebrate genomes but in quadruplicate in vertebrates, a consensus emerged that the entire genome must have been duplicated twice in some early ancestor within the vertebrate lineage<sup>15</sup>. But now that the human genome has been sequenced in draft form, the debate has opened up once more.

Austin Hughes, an evolutionary biologist at the University of South Carolina in Columbia, has found that fewer than 5% of human genes that have invertebrate homologues appear in quadruplicate. Furthermore, of 134 regions that do have four copies, only 30% are organized into two clusters of two, as would be expected if two successive genome duplications had occurred early in vertebrate

evolutionary history<sup>16</sup>. "To me that closes the book on it," says Hughes.

Not so for Wolfe. Mammalian and bird embryos cannot survive if they are tetraploid, so any genome duplications in the vertebrate lineage must have taken place more than 200 million years ago, before either of these groups had evolved. Such an ancient event would be extremely hard to spot in today's genomes, Wolfe argues. For one thing, many duplicate genes — possibly most of them — may have been lost. In yeast, where evidence of polyploidy is stronger, only 16% of genes seem to have a surviving homologous partner<sup>15</sup>. So for Wolfe, whole-genome duplication may still have played an important role in vertebrate evolution.

Species can also experience sudden reductions in chromosome number, as is clear from studies of muntjacs, a diminutive genus of deer. The Chinese muntjac (*Muntiacus reevesi*) has 23 pairs of chromosomes, but the Indian muntjac (*M. muntjac*) has only 3 pairs<sup>17</sup>. Other members of the genus have intermediate numbers of chromosomes. Yet the banding patterns of the chromosomes in all muntjac species suggest that they contain roughly the same genes.

### Shrinking genomes

Wen Wang and Hong Lan of the Chinese Academy of Sciences in Kunming, Yunnan Province, have produced an evolutionary tree of the seven muntjac species on the basis of the DNA found in their energy-generating mitochondria. Their results suggest that the ancestral muntjac species had a high chromosome number, which decreased in some lineages as a result of end-to-end fusions between different chromosomes<sup>18</sup>. And Fengtang Yang and Malcolm Ferguson-Smith at the University of Cambridge, UK, have used the colourful technique of chromosome painting to show that the Indian muntjac chromosome 3 is an assemblage of seven chromosomes from its Chinese cousin<sup>19</sup>.

Chromosomes can decrease in size as well

as in number — indeed, if there wasn't some shrinking, duplication events would cause genomes to grow inexorably over evolutionary time. In fact, Hartl suggests that spontaneous deletion may be the principal process that determines genome size. His team estimated the rate of deletions in several organisms by looking for evidence of excisions within dead transposons, the ages of which can be estimated by the accumulation of mutations in their flanking sequences. Hartl found that the fruitfly *Drosophila melanogaster* is losing DNA 60 times faster than mammals, which fits with its much smaller genome size. Hawaiian crickets (*Laupala* sp.), which have 11 times more DNA than fruitflies, lose DNA 40 times more slowly. And the grasshopper *Podisma pedestris*, whose genome is 100 times bigger than that of *Drosophila*, loses DNA more slowly still<sup>20</sup>.

Evolutionary geneticists are now pondering the significance of the processes that have shaped modern genomes. The more they learn, the less those vast stretches of non-coding DNA that litter most genomes seem like just junk. Pericentromeres and subtelomeres may be incubators for new genes. Duplications may help to generate genetic variation. And natural selection may continually force genomes to become leaner and fitter.

Undoubtedly these processes have also left behind heaps of rubbish. But just as the surface of the Moon holds a record of its impact history in its craters, those genetic wastelands can serve as records of the dynamic history of individual genomes. Now equipped with the best maps yet, the explorers are setting out. ■

Jonathan Knight writes for *Nature* from San Francisco.

1. San Miguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. *Nature Genet.* **20**, 43–45 (1998).
2. Sheen, F. M. et al. *Genome Res.* **10**, 1496–1508 (2000).
3. Glusman, G., Yanai, I., Rubín, I. & Lancet, D. *Genome Res.* **11**, 685–702 (2001).
4. The International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
5. Samonte, R. V. & Eichler, E. E. *Nature Rev. Genet.* **3**, 65–72 (2002).
6. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. *Genome Res.* **11**, 1005–1017 (2001).
7. Bailey, J. A. et al. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
8. Mefford, H. C. & Trask, B. J. *Nature Rev. Genet.* **3**, 91–102 (2002).
9. Linardopoulou, E. et al. *Hum. Mol. Genet.* **10**, 2373–2383 (2001).
10. Trask, B. J., van den Engh, G., Mayall, B. & Gray, J. W. *Am. J. Hum. Genet.* **45**, 739–752 (1989).
11. Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
12. The Arabidopsis Genome Initiative. *Nature* **408**, 796–815 (2000).
13. Wolfe, K. H. & Shields, D. C. *Nature* **387**, 708–713 (1997).
14. Schughart, K., Kappen, C. & Ruddle, F. H. *Br. J. Cancer Suppl.* **9**, 9–13 (1988).
15. Wolfe, K. H. *Nature Rev. Genet.* **2**, 333–341 (2001).
16. Friedman, R. & Hughes, A. L. *Genome Res.* **11**, 1842–1847 (2001).
17. Liming, S., Yingying, Y. & Xingsheng, D. *Cytogenet. Cell Genet.* **26**, 22–27 (1980).
18. Wang, W. & Lan, H. *Mol. Biol. Evol.* **17**, 1326–1333 (2000).
19. Yang, F., O'Brien, P. C. M., Wienberg, J. & Ferguson-Smith, M. A. *Chromosome Res.* **5**, 109–117 (1997).
20. Hartl, D. L. *Nature Rev. Genet.* **1**, 145–149 (2000).