The pufferfish genome: Small is beautiful?

Philip Mileham and Stephen D. M. Brown

Nearly a quarter of a century ago, the study of genome size became a fashionable and important area of study[1]. Comparisons of genome size amongst both plant and animal species revealed an enormous diversity in the amount of DNA in the haploid genome of eukaryotic cells – the *C-value*[2]. Even within the vertebrates a very wide spectrum of genome size was observed with, for example, plethodontid salamanders having genome sizes up to twenty times that of man[3]. The apparent lack of correlation between genome size and so-called evolutionary complexity became known as the C-value paradox[2] and was a spur to the many investigations of genome structure and genome sequence complexity that subsequently developed. Further detailed investigations of genome content and complexity suggested that repeat sequences may be responsible for part of the variation in genome size. However, with unique sequences representing at least half of the genome in many organisms[4], it was clear that differences in repeat sequence content alone could not account for the wide variations in genome size. Nevertheless, these issues continually bring us back to the question of what minimum complement of DNA might be needed in any large class of organisms. How little of the genome can you get away with in order to construct a fully functional vertebrate? Recent work by Sydney Brenner and colleagues[5] on the organization and genome size of the tetraodontoid fish, *Fugu rubripes* – the pufferfish – provides some telling answers.

Along with many other organisms, 25 years ago, tetraodontoid fish too had been the object of genome size studies[6]. *Tetraodon fluviatilis* and *Spheroides maculatus* appeared to have DNA contents of 380 Mb and 480 Mb respectively – some one-eighth to one-sixth the size of the human genome. More recent studies[7] on the genome of *Arothron diadematus* using reassociation kinetics gave an estimate for its genome size of 470 Mb. Only 13% of *Arothron* DNA was repetitive. However, haploid DNA contents, determined by fluorometric analysis or via reassociation kinetics studies, can be unreliable. For these reasons, Brenner and colleagues turned their attention to analysing the complexity and genome size of the *Fugu* by a number of independent methods.

The first method was to analyse the complete sequence of 596 clones (average size 214 bp) derived from sonicated *Fugu* DNA. The derived sequences were scanned in two ways.

1. All the clones were analysed for repeat sequence motifs and ribosomal gene sequences. Overall, 7.6% of sequence in 108 of the clones was repetitive. The most abundant repeat sequence was a clustered tandemly repeated satellite sequence accounting for some 2% of the total genome. Though this satellite sequence had a homologue at much lower frequency in the human genome, no sequences homologous to human SINE (short interspersed repeat sequences e.g. *Alu*) or LINE (long interspersed repeat sequences e.g. *L1*) sequences[8] were found. Intriguingly, the overall number of microsatellites (short dinucleotide repeat sequences) in *Fugu* – around 100,000 – is comparable to that found in the human genome[9].
2. The available sequence was also translated and searched for homologies to known protein sequences. Overall, ten matches were found and in two of these exon-intron boundaries were conserved. The coding information accounted for 0.791% of the total *Fugu* sequence established. By comparison, the available non-redundant mammalian coding information in the protein databases is around 3.1×10^{-6} bp. Thus, given a human genome size of 3000 Mb, a search of random sequence from the human genome would be expected to identify 0.103% as coding sequence. It follows that the *Fugu* genome is some $0.791/0.103 = 7.68 \times$ smaller than the human genome, or around 400 Mb.

Finally, estimates of genome size were made by screening single-copy-gene probes against unamplified phage genomic libraries of *Fugu*. The results indicated that a single copy sequence was found on average once in every 24,625 clones screened. This number of clones therefore represents the equivalent of one *Fugu* genome. The average size of the clones examined was 16.4 kb, indicating that the size of the *Fugu* genome is around 404 Mb.

This combination of approaches demonstrates the small genome size of this vertebrate and also indicates the high complexity of the genome given its size – repetitive sequences are in relatively low abundance, though microsatellites are found in comparable numbers to the human genome, albeit at higher frequency. In addition, Brenner and colleagues allude to evidence that intron size is also correspondingly smaller in the *Fugu* (modal value, 80 bp). Overall, they conclude that the *Fugu* genome looks like a compressed version of the human genome with a lot of the 'junk' thrown out.

Given its small size and low repeat frequency, the *Fugu* is a potential, novel tool for genome mapping studies, particularly in the two vertebrate organisms – mouse and human – that are the major focus of mammalian genome efforts. The relatively large size of mammalian genomes – around 3000 Mb – creates two major problems for genome studies: (1) the establishment of clone contigs, overlapping arrays of clones, covering the entire genome or any relevant area of the genome that provides access to all of the underlying sequences, and (2) identification of coding sequences from within the clone contigs. The problem of genome mapping has been aided immensely by YAC cloning – the use of yeast artificial chromosome clones that can carry megabase inserts of mammalian DNA[10]. Nevertheless, the relatively compact size of the *Fugu* genome may help.

Firstly, it might (though see below) and the process of positional cloning[11] – the identification of a gene associated with a mutation on the basis of its position

Reproduced with permission from *BioEssays*, vol. 16, No. 3, March 1994. © ICSU Press.

in the genome. In many cases, the position of the relevant gene underlying the mutation is defined by closely flanking markers. Often the most closely flanking markers are still some distance from the mutation and contiging the region in between still requires enormous effort, despite the advent of YAC cloning technology. Contiging a comparable region of the *Fugu* genome would require fewer clones of any class – YACs, cosmids, etc. – than would be required in the human genome. Equally, a region spanned by a number of YACs in a mammalian genome could potentially be spanned by a similar number of cosmids in *Fugu*.

Secondly, the compact density of genes and the relatively low abundance of repeat sequences in the *Fugu* genome mean that present screening techniques for the identification of genes within clone contigs[12,13] should be more efficient. Most importantly, correspondingly fewer clones in any class will need to be handled from each region. But also, with the high gene density in *Fugu*, DNA sequencing strategies for gene identification become correspondingly more realistic.

There are two caveats to the *Fugu* as a tool for the study of mammalian genomes, both mentioned by Brenner and colleagues. The first is that we are presently unaware of how much linkage conservation exists between *Fugu* and the mammal. If genes have been sufficiently scrambled between these species over evolutionary time such that gene order is not conserved over the short range, then its use as an aid to positional cloning strategies will be limited. Secondly, the ultimate test of having correctly identified the relevant gene through a positional cloning experiment is to assay gene function in mammalian systems using transgenesis. In addition, transgenic experiments are a necessary step towards understanding gene function and interactions. We are unaware of how well *Fugu* genes will work in mammals – it seems likely that some will work and some won't. Nevertheless, in positional cloning experiments, most genes are confirmed as the correct candidate on the basis, firstly, of the likely function of proteins they encode, and secondly by screening for incumbent mutations. For the latter, it will be necessary to isolate and characterize the homologous mouse or human gene and this will be a suitable substrate for subsequent transgenic experiments.

In the final analysis, the study of tetraodontoid genomes may have a wider significance than providing us with a better way to pursue genome studies. *Fugu* may help furnish further clues to genome evolution. It seems we can make a perfectly good vertebrate without all the cumbersome repeat sequence apparatus that mammalian geneticists have come to know and love. So, is all that paraphenalia irrelevant – is it really all junk? Interestingly, microsatellite numbers in the *Fugu* genome are similar to mammals, so at least one aspect of the various mechanisms of sequence turnover that are known to occur in genomes[14] and that lead to repeat sequence turnover and evolution, i.e. replication slippage, is alive and kicking. It is unknown whether or not mechanisms that lead to interspersed repeats[8] and that can be partly responsible for increases in genome size are missing or defunct in *Fugu*. Nevertheless, it is evident that either *Fugu* has failed to accumulate extra DNA via repeat sequence amplification and other mechanisms, or has lost the bulk of the extra baggage normally seen in mammalian genomes.

Increases in genome size and the presence of interspersed repeats along with high chromosome number could lead to an evolutionary plasticity, which in certain circumstances, may confer a species selective advantage. This advantage may have been exploited by the mammals with their recent successful expansion in evolutionary time[15]. Maybe the *Fugu* is an evolutionary dead end. Its compact, ultra-efficient genome now no longer able to respond to the vagaries of new evolutionary opportunities. However, possibly *Fugu* will stimulate a new era of comparative genome mapping that not only uses maps for the correlation of genetic linkage groups, useful though that is, but also seeks, with the new genome analysis tools at hand, to investigate the nature of changes in overall genome structure through evolutionary time.

1. Britten, R. J. and Davidson, E. H., *Science*, 1969, **165**, 349–357.
2. John, B. and Miklos, G., in *The Eukaryote Genome in Development and Evolution*, Allen and Unwin, London, 1988.
3. Mizuno, S and Muegregor, H C., *Chromosoma*, 1974, **48**, 239–296.
4. Britten, R. J. and Kohne, D. E., *Science*, 1968, **161**, 529–540.
5. Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S., *Nature*, 1993, **366**, 265–268.
6. Hinegardner, R., *Am. Nat.*, 1968, **102**, 517–523.
7. Pizon, V., Cuny, G. and Bernardi, G., *Eur J. Biochem*, 1974, **140**, 25–30.
8. Deininger, P. L., Batzer, M A., Hutchison, C. A. and Edgell, M. H., *Trends Genet.*, 1992, **8**, 307–311
9. Beckmann, J. S. and Weber, J. L., *Genomics*, 1992, **12**, 627–631.
10. Anand, R., *Trends Biotech*, 1992, **10**, 35–40.
11. Collins, F., *Nature Genet*, 1992, **1**, 3–6.
12. Panmoo, S., Kolluri, R. and Weissman, S. M., *Nucl. Acids Res.*, 1993, **21**, 4422–4423.
13. Buckler, A. J., Chang, D. D., Graw, S. L. Brook, D. J., Haber, D. A., Sharp, P. A. and Housman, D. E., *Proc. Natl Acad. Sci. USA*, 1991, **88**, 4005–4009.
14. Dover, G. A., *Trends Genet*, 1986, **2**, 159–165.
15. Bush, G. L., Case, S. M., Wilson, A. C. and Patton, J. L., *Proc. Natl Acad. Sci. USA*, 1977, **74**, 3942–3946.

Philip Mileham and Stephen D. M. Brown are in the Department of Biochemistry and Molecular Genetics, St. Mary's Hospital Medical School, London W2 1PG, UK.