**CHARACTERIZATION OF THE COMPACT MODEL GENOME OF THE JAPANESE PUFFER FISH (FUGU RUBRIPES) USING A COSMID SEQUENCE SCANNING APPROACH.**

ELGAR G.+, CLARK M.S., EDWARDS Y.J.K., MEEK S., SMITH S., UMRANIA Y., WARNER S., WILLIAMS G., BISHOP M.J.

UK Human Genome Mapping Project Resource Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SB, UK.;
e-mail: gelgar@hgmp.mrc.ac.uk;

+Corresponding author

Keywords: japanese puffer fish genome, cosmid sequence scanning, genome characterization, minimalist model genome

**Abstract**

In comparison with other vertebrate genomes, the genome of*Fugu rubripes*(*Fugu*) comprises less repetitive DNA sequence, smaller intergenic regions, smaller introns and is gene dense [1,2,3]. In the light of the above, a landmark map is being generated for the genome of*Fugu*.The aim of the*Fugu*Landmark Mapping Project is to generate partial sequences from 1000 cosmids using a random shotgun approach to provide reference points covering 10% of the*Fugu*genome. The*Fugu*Landmark Mapping Project was initiated in November 1996. This DNA sequence will cover 10% of the genome and provide data on short range linkage in*Fugu*and synteny with higher vertebrates. The*Fugu*Landmark Mapping Project has been running for over two years and 1000 cosmids will be completed around mid August 1998. The general characteristics of the minimalist model genome are described.

**1. Introduction**

*Fugu*has one of the smallest vertebrate genome which comprises approximately 400Mb. As an established model vertebrate genome, it is an economical tool on which to perform comparative sequence analyses both in hybridization experiments and molecular bioinformatics studies [1,2,3]. The*Fugu*Landmark Mapping Project primarily involves large scale single pass sequencing and sequence analysis on*Fugu*DNA. Sequence analyses have identified genomic regions likely to code for proteins, syntenic genes similar to that observed in the genome of other species, repetitive DNA sequences such as microsatellites and minisatellites and conserved regulatory regions in equivalent gene structures. The*Fugu*DNA sequence database at the time of writing this document contains over 22Mb of DNA and this is estimated to cover about 7% of the genome.

**2. Constructing the*Fugu*DNA Sequence Database**

Cosmids from the*Fugu*genomic library are selected randomly and sequenced using a shotgun scanning approach. DNA is prepared from cosmids by alkaline lysis and sonicated to an average size of 500-1000bp. These fragments are cloned into pBluescript II KS and sequenced from T3/T7 primer generated PCR products with KS primer giving an average clone length of 480bp. Approximately 50 clones are sequenced per cosmid. Primer and vector sequences are identified and clipped off. Sequences that comprise less than 60 bases or have sequence ambiguities of 10% or more are discarded. The sequences are screened for significant similarities using BLAST [4] matches against the EMBL, TrEMBL, and SWISSPROT databases. Identical or near identical matches with E. coli sequences are removed. The resulting cosmid-subcloned sequences are searched systematically for various types of protein coding regions, microsatellite, minisatellite sequences. Detailed information connected with the*Fugu*Landmark Mapping Project such as the project description, project protocols, cosmid sequences, clone sequences and the BLAST matches to EMBL, TrEMBL and SWISSPROT databases is available via the UK Human Genome Mapping Project Resource Centre’s World Wide Web site at the following URL[http://*Fugu*.hgmp.mrc.ac.uk](http://fugu.hgmp.mrc.ac.uk/).

**3. Identification and Characterization of Protein Coding Regions**

Each cosmid contains on average 50 clones approximating to 50% single pass coverage per cosmid. There are on average 3 strong gene hits per cosmid idetified using BLAST. Whilst the majority of searches reveal homology to mammalian genes, a significant proportion match*yeast S. cerevisiae*and*C. elegans*hypothetical genes. An increased number of conserved linkages in the genomes of*Fugu*and human have been discovered (Table 1) along with several evolutionary breakpoint regions.

Apart from the significant similarities using BLAST matches against the protein sequence databases, several gene prediction programs were tested on the*Fugu*Landmark Mapping DNA sequences. This type of analyses is useful to identify new genes likely to be present in the genomic sequences of*Fugu*. The Genemark program [5] contains a*Fugu-*DNA matrix that performed most successfully. A random subset of output files were subjected to further analysis using the information of known genes established from BLAST. Genemark proved effective at predicting the region of an ORF in the correct reading frame. This is particularly impressive considering that the maximum length of each clone is 650bp. Analysis of the Genemark ORFs enabled a re-estimation of the total percentage of protein coding regions in the*Fugu*genome. 14% of the genome is predicted to code for protein. This prediction is based on 529 cosmids (26,220 clones; 11.338Mb of*Fugu*DNA) which have 10,986 ORFs assigned by Genemark. The predicted quantity of protein coding in*Fugu*here is similar with a previous estimate in an earlier study in which 128Kb*Fugu*DNA was used [1]. This preliminary analysis of ORFs using the Genemark program indicates that about 60% of ORFs show no significant sequence homology to genes currently in any database and this finding is typical of other genomic sequencing projects; an example being the complete DNA sequence of yeast*Saccharomyces cerevisiae*[6].

Table 1. Examples of Conserved Synteny in*Fugu*

|  |  |  |
| --- | --- | --- |
| **Physically linked genes in *Fugu*** | **Distance apart (kb)** | **Human chromosome assignment** |
| Cannabis receptor | Both on one cosmida | 6q14-q15 |
| GABA rho receptor |   | 6q14-q21 |
| Tyrosine hydoxylase | 10kb of each otherb | 11p15.5 |
| Nucleosome assemby protein |   | 11p |
| Insulin-like growth factor receptor II |   | 11p15.5 |
| Phenylalanine hydroxylase | 20kb of each otherc | 12q22 |
| Insulin-like growth factor receptor I |   | 12q22-q24 |
| TSC2 | 1kb of each otherd | 16p13.3 |
| ADPKDI |   | 16p13.3 |
| wntI | All on one cosmide | 12q13 |
| erb-b3 |   | 12q13 |
| ADP ribosyl factor 3 |   | Unassigned |
| wnt 10b |   | Unassigned |
| fos | All on one cosmidf | 14q24.3 |
| s31 Golgi transport protein |   | 14q24.3 |
| s20i15 transcription factor |   | 14q24.3 |
| 7SL RNA gene |   | Unassigned |
| Activating factor 3 |   | Unassigned |
| fos-like gene |   | 14q24.3 |
| Dihydrolipoamide succinyl transferase |   | 14q24.3 |
| Hox b cluster (nine genes) | 90kbg | 17q21-22 |
| Hox c cluster (nine genes) | 66kbh | 12q12-q13 |
| Ig VH genes (at least 6 genes) | All on one cosmidi | 14q32.3 |
| Surf-2 | All on one cosmidj | 9q34 |
| Surf-4 |   | 9q34 |
| ASS |   | 9q34 |
| Dynamin |   | 9q34 |
| Golgin-95 |   | Unassigned |
| Basic transcription factor | All on one cosmid | 11p15 |
| Mucin 2 |   | 11p15.5 |
| Eosinophil peroxidase precursor |   | 11p12-p2 |
| 3-oxo-5-alpha-steroid-dehydrogenase I | All on one cosmid | 5p15 |
| Adenylate cyclase II |   | 5p15.2-15.1 |
| Wt I | All on one cosmidk | 11p13 |
| RCN |   | 11p13 |
| Pax 6 |   | 11p13 |
| 72KD type IV collagenase precursor | All on one cosmid | 16q |
| Sodium-dependentNor-adrenalin transporter |   | 16q |
| Tyrosine kinase CAK precursor | All on one cosmid | 1q33 |
| Regulator of G-protein signalling |   | 1q33 |
| Topoisomerase I | All on one cosmidl | 20q |
| Phospholipase C gamma I |   | 20q |
| KIAA0181 |   | 20q |
| Complement C8 alpha | All on one cosmidm | 1p32 |
| Complement C8 beta |   | 1p32 |
| MTF-1 | All on one cosmid | 1p32-34 |
| IT5-P |   | 1p34 |
| HCGV | All on two cosmids | 6p21.3 |
| Tenascin-like |   | 6 |
| Valyl tRNA synthetase |   | 6 |
| AA008813 |   | 6p21.3 |

Table References: aYamaguchi (unpublished data); bSandford (unpublished data); cSandford (unpublished data); dSandford et al (1996) Genomics 38:84-86; eGellner (unpublished data); fTrower et al (1996) Proc. Natl Acad. Sci. (USA) 93:1366-1369; gAparicio et al (1997) Nat. genet. 16:79-83; hAparicio et al (1997) Nat. genet. 16:79-83; iPeixoto (unpublished data); jBouchireb (unpublished data); kMiles et al (unpublished data); lSmith (unpublished data); mYeo (unpublished data).

**4. Identification and Characterization of Microsatellites and Minisatellites**

501 theoretically possible microsatellites with a repeat unit of 1-6 bp were used to query 11.338Mb of*Fugu*DNA [7] using a Smith Waterman based algorithm in Censor [8]. In decreasing order, the twenty most frequently occurring microsatellites are AC, A, C, AGG, AG, AGC, AAT, AAAT, ACAG, ACGC, ATCC, AAC, ATC, AGGG, AAAG, AAG, AAAC, AT, CCG and TTAGGG. The twenty most frequently occurring microsatellites represent 81.79% of the total microsatellites identified. One microsatellite occurs every 1.876Kb of DNA in*Fugu*. 11.55% of the microsatellites are detected in open reading frames (ORFs) that are predicted protein coding regions by Genemark [5]. With respect to the proportion of microsatellites present in ORFs and the total abundance (bp) of the total microsatellites, the genome of*Fugu*is similar to the genome of other vertebrate species. Previous estimates performed on genomic sequnces of primates, human, rat, rabbit, pig and chicken indicate that approximately 1% of many vertebrate genomes are comprised of microsatellites. However, many differences prevail in the abundance and frequency of the individual microsatellite classes. Many of the frequently occurring microsatellites in*Fugu*are known to code in other species for regions in proteins such as transcription factors, whilst others are associated with known functions, such as transcription factor binding sites in DNA sequences. Therefore, it is likely that microsatellites have a very significant and major role in evolution of genes, genomes and organisms.

The identification of microsatellite DNA in*Fugu*has proved essential for the identification and characterization of other repetitive structural units such as minisatellites. For the identification of similar sequences within the*Fugu*genome that are eukaryotic, vertebrate, fish or*Fugu*specific, cluster analysis is required. Such clustering of genomic sequences requires the identification and masking out of microsatellites which would otherwise hinder clustering studies. Clustering analysis is carried out using ICAtools suite of programs [9]. The minisatellite classes identified in*Fugu*appear few in number and repeated at low to medium numbers. These are being characterised further. The considerations given to the two broad classes of repeat structures in*Fugu*i.e., the minisatellites and transposable elements, suggest that these two main classes of repeats comprise less than 10% of the genome. It is these two classes of repetitive elements that are significantly reduced in*Fugu*compared with other vertebrate species such as human and primates, and not microsatellite sequences.

**5. Conclusions & The Future**

The general properties of the*Fugu*genome are well established and these include a conserved gene structure with other vertebrate species, small introns, similar codon usage and splice signals and homologous coding sequences. With respect to the analyses of the regulatory networks associated with equivalent and completely sequenced gene systems in different species, comparative sequence analysis are useful in understanding functions and functional relationships. For example, there are a growing number of examples where conserved elements in the 5’ regions of equivalent gene structures present in different species were subsequently shown to be functional regulatory regions in mammals and the pufferfish as seen in the Hox [10], the Otx2 [11] and the oxytocin gene structures [12]. The*Fugu*Landmark Mapping Project has been primarily concerned with the partial single-pass sequence scanning of cosmids and now the genomic research has moved towards the complete cosmid sequencing, assembling and finishing of key genomic regions identified from sequence scanning in*Fugu*to permit more detailed comparison with the human genome especially with regards to studying regulatory networks for genomic structure and syntenic genes.

**Acknowledgements**

This work is financially supported by the MRC grant awarded to Dr. Greg Elgar and Professor Sydney Brenner for the*Fugu*Landmark Mapping Project and the EC TRAnscription Databases and Analysis Tools (TRADAT) project (Contract: BIO4-CT95-0226).

**References**

1. S. Brenner, G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, & S. Aparicio. (1993). Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. Nature,**366**:265-268.
2. G. Elgar (1996). Quality not quantity: the pufferfish genome. Human Molecular Genetics,**5**: 1437-1442.
3. G. Elgar, R. Sandford, S. Aparicio, A. Macrae, B. Venkatesh, & S. Brenner (1996). Small is beautiful - comparative genomics with the pufferfish (Fugu rubripes). Trends in Genetics,**12**:145-150.
4. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, & D.J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol.,**215**:403-410.
5. M. Borodovsky & J. McIninch (1993). GENEMARK - Parallel gene recognition for both dna strands. Comp. Chem.,**17**:123-133.
6. D. Botstein, S.A. Chervitz, & J.M. Cherry (1997). Yeast as a model organism. Science,**277**:1259-1260.
7. Y.J.K. Edwards, G. Elgar, M.S. Clark & M.J. Bishop (1998). The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish,*Fugu rubripes*: perspectives in functional and comparative genomic analyses. J. Mol. Biol.,**278**:843-854.
8. J. Jurka, P. Klonowski, V. Dagman, & P. Pelton (1996). Censor - a program for identification and elimination of repetitive elements from DNA sequences. Comp. Chem.**20**:119-121.
9. 9a. J.D. Parsons, S. Brenner & M.J. Bishop (1992). Clustering cDNA sequences. CABIOS,**8**:461-466. 9b. J.D. Parsons (1995). Improved tools for DNA comparison and clustering. CABIOS,**11**:603-613.
10. S. Aparicio, A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf, & S. Brenner. (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish,*Fugu rubripes*. Proc. Natl. Acad. Sci. USA,**92**:1684-1688.
11. C. Kimura, N. Takeda, M. Suzuki, M. Oshimura, S. Aizawa & I. Matsuo. (1997). Cis-acting elements conserved between mouse and pufferfish Otx2 genes govern the expression in mesencephalic neural crest. Cells Development,**124**:3929-3941.
12. B. Venkatesh, S.L. Sihoe, D. Murphy & S. Brenner (1997). Transgenic rats reveal functional conservation of regulatory controls between the*Fugu*isotocin and rat oxytocin genes. Proc. Natl. Acad. Sci. USA,**94**:12462-12466.