A Fugu–Human Genome Synteny Viewer: web software for graphical display and annotation reports of synteny between Fugu genomic sequence and human genes

Mark Halling-Brown, Clare Sansom, David S. Moss, Greg Elgar¹ and Yvonne J. K. Edwards^{1,*}

School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK and
¹MRC Rosalind Franklin Centre for Genomics Research, Genome Campus, Cambridge CB10 1SB, UK

Received February 17, 2004; Revised March 30, 2004; Accepted April 7, 2004

ABSTRACT

A web server has been developed to access annotation and graphical reports of synteny and gene order between the Fugu genome and human genes. In this system, the assembled Fugu genomic sequences (also known as scaffolds) are annotated. The annotations for each Fugu scaffold are computed, stored and made publicly available. The annotations describe matches to human homologous genes. For each significant human gene match on the Fugu scaffold, the corresponding human chromosome map and measures of the significance of each match are given. The web-based server provides public access to these annotations and graphical displays of the results. The user is provided with a selection of views including a chromosome-colour-coded image and a table containing the details of the matches. The Fugu–Human Genome Synteny Viewer has been tested by comparing results with examples from a paper that includes a study of transcription factors, Fos and Jun encoding regions. The Fugu–human genome synteny views are available for each Fugu scaffold through the clonesearch web page located at the Fugu Genomics website (<http://fugu.rfcgr.mrc.ac.uk/>).

INTRODUCTION

The *Fugu rubripes* (Fugu) genome is one of the smallest for all vertebrates and is an established tool for comparative genomics (1,2). Comparing the genomes of evolutionarily divergently evolved species, such as those of human and Fugu, has proven useful to identify sequences common to vertebrates that form the basis for functional studies (3,4). Several regions of conserved gene synteny have been reported between human and Fugu (1–10). Some comparative synteny studies have

aided the identification of new genes (5) and functional regulatory regions such as enhancers (6–10). It is the search for conserved non-coding elements (CNEs) that requires the identification of orthologues in multi-gene blocks shared between the sequences of human and Fugu. It is these CNEs that are screened for possible gene regulatory function. Synteny is important as some of these functional enhancers are located in the neighbouring genes of the gene they regulate (9). For these reasons and in order to aid researchers interested in studying synteny, gene order and functional genomics, we developed a system that reduces the time and effort to perform this analysis. We describe a web-based application in which gene synteny and gene order can be studied in Fugu and human. The application allows rapid comparisons of gene order between two highly evolutionarily divergent vertebrates and is publicly accessible via a web browser.

MATERIALS AND METHODS

Bioinformatics resources

The genome of Fugu has been sequenced to over 90% coverage and is publicly available (2). The current draft assembly consists of 8023 scaffolds ranging from 2 to 1100 Kbp. The synteny viewer provides annotation to an earlier release (2) and the current release of draft assemblies. These draft assemblies have been searched against protein and nucleotide public databases using BLAST (11) (version 2.0.12). These results were used to acquire the human gene matches to the individual Fugu scaffolds. The BLAST results used to decipher the significant matches came from two databases: International Protein Index (IPI) (<http://www.ebi.ac.uk/IPI/IPIhelp.html>) and UNIGENE (12). A match was classified as significant when the BLAST score, percentage identity and overlap were above a predefined threshold. These predefined thresholds were determined by experimentation with published results. IPI provides a top-level guide to the main databases that describe the human proteome: SWISS-PROT/TrEMBL (13), Ensembl (14) and RefSeq (15). UNIGENE is an experimental

*To whom correspondence should be addressed. Tel: +44 1223 494531; Fax: +44 1223 494512; Email: yjedward@rfcgr.mrc.ac.uk

system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UNIGENE cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and human chromosome map location (12).

The IPI database was searched using BLASTx and UNIGENE was searched using BLASTn. These BLAST results were run through the program MSPcrunch (16) to produce parsable CRUNCH files. The CRUNCH files contain a summary of the information from the BLAST results, filtered according to the *E*-values, and allow for easier processing. The Sequence Retrieval System (SRS) (17) was used to query the relevant databases in order to determine the significant matches to whole genes and obtain the required data for these results.

During development it was decided that matches provided by IPI and UNIGENE should be used. The system was limited to using these two databases as it was difficult to automate the analysis of results from the other databases. Furthermore, as the information stored in these databases used by the system is updated, the BLAST searches will be reapplied.

Description of the Fugu–Human Genome Synteny Viewer

This system is implemented in two parts. The first part produces and stores the annotation. The second comprises web-based scripts, which read the results from the stored files and display them in graphical and tabular form. The results are stored in XML format making it easy to read the data with little need for laborious parsing. The first section was designed and implemented as a Java application with the option of running it on individual Fugu scaffolds or as a batch process on all the Fugu scaffolds currently available. Its main roles are to read in the relevant BLAST results, query various databases to identify which of the matches refer to complete genes and acquire any useful information from these, filter through the results produced to remove low-scoring matches and produce an XML file containing the results. The second section of this system was designed and implemented as a PERL CGI script for use on the Fugu Genomics website located at <http://fugu.rfcgr.mrc.ac.uk>. This section of the system reads in the relevant XML file, produces an HTML table summarizing these results, and generates and displays a GIF image colour coded by chromosome to show the positions of the matches in relation to the Fugu scaffold in question (Fig. 1). A further script displays the results in a tab-delimited text format (Fig. 2).

Using the Fugu–Human Genome Synteny Viewer

The front end of the Fugu–Human Genome Synteny Viewer is provided by PERL CGI scripts. The ‘Human Synteny Viewer’ link may be accessed from the clonesearch page (<http://fugu.rfcgr.mrc.ac.uk/fugu-bin/clonesearch>) of the Fugu Genomics website. This accesses the results and displays them on the web browser. An online tutorial is accessible at the URL <http://fugu.rfcgr.mrc.ac.uk/News/> from the Fugu–Human Genome Synteny Viewer link.

Calculation of score, percentage overlap and percentage identity

Three important variables calculated for each match are the BLAST score, percentage overlap and percentage identity as these are used as the parameters for the cut-off threshold and filter, i.e. the limits placed on a match to define it as significant or not. A user can alter the threshold values via the form. This allows the user to increase or decrease the stringency of the threshold to display more or less significant matches. The cut-offs used during the processing stage were kept low: the BLAST score ≥ 30 , the percentage identity $\geq 30\%$ and the percentage overlap $\geq 10\%$. This increases the amount of insignificant matches, but also reduces information loss. For each match the BLAST score was obtained directly from the CRUNCH file, and the percentage identity was acquired by parsing the BLAST results for the relevant hit. The percentage overlap was then calculated by obtaining the length of the area of the gene matched and dividing it by the full length of that gene. In the cases when there were multiple matches of different exons from the same gene, these values were combined to provide single values for that gene. Each exon’s BLAST scores were summed to provide a single value. For example, with respect to the percentage identity calculation, the number of bases covered by each exon in the Fugu scaffold was summed to provide the total length of match covered by the gene. This value was divided by the total length of the gene to calculate the percentage overlap. In the case where genomic sequence was compared with protein sequence the numerator was divided by 3. In order to obtain a single value for the percentage identity, the individual values for each exon were averaged. The values for BLAST score, percentage identity and percentage overlap for each exon were stored alongside the overall values for that gene, as these details can be viewed via the web-based interface.

RESULTS

A useful method of demonstrating and evaluating this tool is to compare the results from the study of synteny and homology between specific regions or genes on the human and Fugu genomes using other methods. Examples from a paper by Cottage *et al.* (18) have been used to demonstrate this system. The results generated by Cottage and coworkers were obtained by querying the human protein sequences of *fos* and *jun* against the databases of Fugu scaffolds (<http://fugu.rfcgr.mrc.ac.uk/Analysis/>). Homologous sequences to human *fos* and *jun* genes were found in 11 Fugu scaffolds. This paper will consider one of these scaffolds to compare and contrast the findings with the results obtained through the Fugu–Human Genome Synteny Viewer.

Gene Content of the Fugu Jun encoding scaffold S000290

Thirteen genes were located on scaffold S000290 (18). These are RGS12, HD, GPRK2L, TOM1, NOG, ASNA, VMD2, JunB, PRDX2, WIZ, PTGER1, IL3 and ACP5. Six have known homologous sequences mapping to the human JunB locus 19p13.2: ASNA, VMD2, JUNB, PRDX2, ILF3 and ACP5. Two other genes map to the nearby locus 19p13.12: WIZ and PTGER1. This is an example where synteny is conserved between these genomes and not the exact gene order. This is commonly observed and has been reported previously (6). A

Figure 1. Screenshot of scaffold S000290 with no filters applied.

Fugu and human gene comparisons showing human map positions													
Remove	No	AccessionId	Genename	Product	Map position	Locus Link	Percent Overlap	Blast Score	Match length	Query start	Query end	Percent Id	DB Xreference
Positive strand													
☐	1	IP100023191	TOM1	target of myb1	22q13.1	10043	68.29	459	3281	69200	72481	67	O60784 604700 Hs 9482
☐	2	IP100012361	NOG	noggin precursor	17q21-q22	9241	40.95	78	339	74305	74664	30	Q13253 602991 Hs 248201
☐	3	IP100012466	ASNA1	arsA arsenite transporter, ATP-binding, homolog 1	19q13.3	439	107.47	609	2132	75323	77455	82	Q43681 601913 Hs 165439
☐	4	Hs 165439	ASNA1	arsA arsenite transporter, ATP-binding, homolog 1	19q13.3	439	64.39	717	1419	76002	77421	86	601913 Hs 165439
☐	5	IP100008823	VMD2	vitelliform macular dystrophy (Best disease, bestrophin)	11q13	7439	68.38	529	2241	79500	81741	59	Q76090 153700 Hs 167244
☐	6	IP100014945	FLJ20132	vitelliform macular dystrophy 2-like protein 1	19p13.2	54831	94.42	529	2199	79725	81924	68	Q9NXP0 607335 Hs 435611
☐	7	Hs 302513	VMD2L2	vitelliform macular dystrophy 2-like protein 2	1p33-p32.3	266675	13.4	158	612	80068	80680	88	607336 Hs 302513
☐	8	IP100013763	MGC13168	hypothetical protein MGC13168	12q14.3	84821	73.61	82	161	80140	80301	72	Q9BR80 Hs 412421
☐	9	IP100013439	JUNB	jun B proto-oncogene	19p13.2	3726	98.56	301	980	108790	109770	52	P17275 165161 Hs 400124
☐	10	IP100008965	JUN	v-jun avian sarcoma virus 17 oncogene homolog	1p32-p31	3725	99.7	257	992	108790	109782	45	P05412 165160 Hs 78463
☐	11	IP100015131	JUND	jun D proto-oncogene	19p13.2	3727	76.37	216	827	108955	109782	47	P17535 165162 Hs 2780
☐	12	Hs 400124	JUNB	jun B proto-oncogene	19p13.2	3726	14.8	239	654	109100	109754	86	165161 Hs 400124
☐	13	IP100018648	PTGER4	prostaglandin E receptor 4 (subtype EP4)	5p13.1	5734	51.64	125	1569	135723	137292	38	P35408 601586 Hs 199248
☐	14	IP100003081	PTGER1	prostaglandin E receptor 1, subtype EP1	19p13.1	5731	89.8	265	1617	135735	137352	42	P34995 176802 Hs 159360
☐	15	IP100015730	PTGER	prostaglandin F receptor (FP)	1p31.1	5737	82.45	252	1596	135768	137364	43	P43088 600563 Hs 89418
☐	16	IP100016313	PTGER3	prostaglandin E receptor 3, subtype EP3 isoform 1	1p31.2	5733	71.28	166	1530	135771	137301	43	P43115 176806 Hs 27860
☐	17	IP100016317	PTGIR	prostaglandin I2 (prostacyclin) receptor (IP)	19q13.3	5739	66.06	116	1581	135780	137361	32	P43119 600022 Hs 438324
☐	18	IP100013470	TBXA2R	thromboxane A2 receptor	19p13.3	6215	73.76	198	1479	135795	137274	56	188070 Hs 442530
☐	19	IP100010309	TBXA2R	thromboxane A2 receptor	19p13.3	6215	68.56	198	1479	135795	137274	56	P21731 188070 Hs 442530
☐	20	IP100007200	ZFR	zinc finger RNA binding protein	5p13.3	51663	57.42	309	6275	140118	146393	60	Q9Y687 Hs 173518
☐	21	IP100005199	ILF3	interleukin enhancer binding factor 3 isoform b	19p13.2	2602	76.75	449	7496	140118	147614	51	Q12906 603182 Hs 256583
☐	22	IP100020061	STRBP	spermatid perinuclear RNA-binding protein	9q33.3	55342	73.06	463	4208	143346	147554	65	Q9NUK4 Hs 287659
☐	23	IP100009937	ACP5	tartrate resistant acid phosphatase 5 precursor	19p13.3-p13.2	54	92.57	367	1119	150648	151767	53	P13686 171640 Hs 1211
Negative strand													
☐	1	IP100004087	RGS12	regulator of G-protein signalling 12 isoform 2	4p16.3	6002	44.33	563	4841	1586	6427	50	Q9UJN3 602512 Hs 434933
☐	2	IP100024714	RGS12	regulator of G-protein signalling 12 isoform 2	4p16.3	6002	42.16	563	4841	1586	6427	50	Q14924 602512 Hs 434933
☐	3	IP100019997	LIN-7-C	lin-7 homolog C	11p14	55327	40.1	50	224	6421	6197	38	Q9NUF9 Hs 91393
☐	4	IP100002335	HD	huntingtin	4p16.3	3064	106.2	4172	21729	11173	32902	66	P42838 143100 Hs 437162
☐	5	IP100027753	GPRK2L	G protein-coupled receptor kinase 2-like isoform 2	4p16.3	2868	97.4	766	11072	37261	48333	64	P32298 137026 Hs 32952
☐	6	IP100017330	GPRK6	G protein-coupled receptor kinase 6	5q35	2870	96.53	749	11039	37285	48324	63	P43250 600869 Hs 235116
☐	7	IP100002995	GPRK5	G protein-coupled receptor kinase 5	10q24-qter	2869	85.42	756	15060	37285	52345	71	P34947 600870 Hs 211569
☐	8	IP100019980	RHOK	rhodopsin kinase	13q34	6011	64.83	390	6499	37554	44053	50	Q13835 180381 Hs 103501
☐	9	IP100021981	GPRK2L	G protein-coupled receptor kinase 2-like isoform 2	4p16.3	2868	101.4	717	10773	37560	48333	65	137026 Hs 32952
☐	10	Hs 211569	GPRK5	G protein-coupled receptor kinase 5	10q24-qter	2869	10.09	211	25810	37636	63446	87	600870 Hs 211569
☐	11	Hs 435611	VMD2L1	vitelliform macular dystrophy 2-like protein 1	19p13.2	54831	11.11	54	73	81317	81390	83	607335 Hs 435611
☐	12	IP100027350	PRDX2	peroxiredoxin 2 isoform a	19p13.2	7001	100.0	337	1377	111306	112683	71	P32119 600538 Hs 432121
☐	13	IP100011937	AOE372	thioredoxin peroxidase	Xp22.13	10540	70.11	308	1362	111306	112668	67	Q13162 606596 Hs 83383
☐	14	IP100000874	PRDX1	peroxiredoxin 1	1p34.1	5052	99.5	337	1374	111309	112683	72	Q06830 176763 Hs 180909
☐	15	IP100024912	PRDX3	peroxiredoxin 3 isoform a precursor	10q25-q26	10935	68.36	236	1044	111648	112692	58	P30048 604769 Hs 327062
☐	16	Hs 432121	PRDX2	peroxiredoxin 2 isoform a	19p13.2	7001	11.95	115	826	111668	112494	87	600538 Hs 432121

Figure 2. Screenshot of scaffold S000290 table.

distinct evolutionary chromosomal breakpoint is observed. Three of the genes encoded map to the Huntington's disease locus 4p16.3: RGS12, GPRK2L and HD. The two other genes map to 17q23.2: TOM1 and NOG. The results produced by Fugu-Human Genome Synteny Viewer for this scaffold can be seen in Figures 1 and 2. The Fugu-Human Genome Synteny Viewer enabled the identification of 12 out of the 13 genes found by Cottage and coworkers including the three proteins that map to the Huntington's disease locus 4p16.3. TOM1 and NOG were both identified by the Fugu-Human Genome Synteny Viewer. NOG was determined to map to 17q23.2. TOM1 maps to 22q13.1. The six genes which Cottage and coworkers found to map to 19p13.2 were also identified by this software. JUNB, PRDX2, ASNA, ACP5 and ILF3 mapped to 19p13.2. A VMD2-like gene (VMD2L1) which maps to 19p13.2 was identified. PTGER1 was identified to map to 19q13.12. WIZ, however, was not shown, as the corresponding entry in the IPI database is not accompanied by a Locuslink identifier. In addition the predicted paralogues are shown for many of the genes discussed.

DISCUSSION

This system was designed to accept a Fugu scaffold identifier and determine matches to annotated human genes. It uses database cross-references to determine the significant matches to complete genes and then annotate these with relevant human chromosome mapping information. It frees the user from the tedious and repetitive tasks of searching, parsing and collating, allowing more time to be devoted to the important stages of expert manual verification of the results. This system is not a substitute for expert analysis. There are a number of features of the system which make it practical and useful, offering advantages over unassisted manual searching, collating and annotation. These are (i) automation, which provides the analysis methods in an objective and repeated fashion, (ii) use of multiple databases to acquire useful annotations, and (iii) provision of an online tool to examine results. All of these are prerequisites for the annotation of the very large number of genomic sequences currently available. Automation also permits this same system to be reapplied to the Fugu genome as the BLAST searches are reapplied or databases are updated.

Another system that provides related functionality in terms of annotation for Fugu scaffolds is Ensembl (14). Whilst Ensembl is a gold standard in vertebrate genome annotation, it does not currently provide a Fugu-human synteny viewer. There is more human-specific information on our Fugu-Human Genome Synteny Viewer such as the human chromosome map positions that are readily accessible. Our viewer provides quick and easily accessible synteny comparisons between Fugu scaffolds and human genes.

System limitations and future developments

The many advantages of an automated system for the annotation of a Fugu genomic sequence have been discussed. However, any manual or automatic procedure is prone to several types of error. Some of these errors are not addressed by this system. The most important of these are (i) false positives, where information is used based on a wrongly inferred homology, (ii) inaccurate positives, where the wrong information is used although the homology is correct, and (iii) inaccurate sources, where the database source is itself

misleading. These sorts of errors give greater emphasis to the need for manual scrutiny of the output to arrive at an expert adjudication. Although many levels of trust in the information stored in databases are implemented, the system assumes the accuracy and validity of the data. However, database entries may use heterogeneous nomenclature and contain incorrect annotation. This system presents the results from mapping human genes to the Fugu genome. A useful alternative would be to map Fugu genes to the human genome. This would provide the user with the opportunity to compare and contrast these two approaches.

REFERENCES

- Brenner, S., Elgar, G., Sandford, R., Macrae, A., Vankatesh, B. and Aparicio, S. (1993) Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature*, **366**, 265–268.
- Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**, 1301–1310.
- Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, **16**, 369–372.
- Nobrega, M.A. and Pennacchio, L.A. (2004) Comparative genomic analysis as a tool for biological discovery. *J. Physiol.*, **554**, 31–39.
- Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C. *et al.* (2003) Comparative analysis of multi-species sequences from targeted genome regions. *Nature*, **424**, 788–793.
- Smith, S.F., Snell, P., Gruetznier, F., Bench, A.J., Haaf, T., Metcalfe, J.A., Green, A.R. and Elgar, G. (2002) Analyses of the extent of shared synteny and conserved gene orders between the genome of *Fugu rubripes* and human 20q. *Genome Res.*, **12**, 776–784.
- Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
- Abrahams, B.S., Mak, G.M., Berry, M.L., Palmquist, D.L., Saionz, J.R., Tay, A., Tan, Y.H., Brenner, S., Simpson, E.M. and Venkatesh, B. (2002) Novel vertebrate genes and putative regulatory elements identified at kidney disease and NR2E1/fierce loci. *Genomics*, **80**, 45–53.
- Santagati, F., Abe, K., Schmidt, V., Schmitt-John, T., Suzuki, M., Yamamura, K. and Imai, K. (2003) Identification of *cis*-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny. *Genetics*, **165**, 235–242.
- Yu, W.P., Pallen, C.J., Tay, A., Jirik, F.R., Brenner, S., Tan, Y.H. and Venkatesh, B. (2001) Conserved synteny between the Fugu and human PTEN locus and the evolutionary conservation of vertebrate PTEN function. *Oncogene*, **20**, 5554–5561.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Pontius, J.U., Wagner, L. and Schuler, G.D. (2003) UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. National Center for Biotechnology Information, Bethesda, MD.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
- Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
- Sonnhammer, E.L. and Durbin, R. (1994) An expert system for processing sequence homology data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 363–368.
- Zdobnov, E.M., Lopez, R., Apweiler, R. and Etzold, T. (2002) The EBI SRS server—new features. *Bioinformatics*, **18**, 1149–1150.
- Cottage, A.J., Edwards, Y.J. and Elgar, G. (2003) API genes in Fugu indicate a divergent transcriptional control to that of mammals. *Mamm. Genome*, **14**, 514–525.