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. Edwards¹,*

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 database that represents the human transcriptome.
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.rdfgr.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.rdfgr.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.rdfgr.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 $\geq 30$, 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.rdfgr.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.
<table>
<thead>
<tr>
<th>Accession</th>
<th>Organism</th>
<th>GeneName</th>
<th>Product</th>
<th>Map position</th>
<th>Locus Link</th>
<th>Percent overlap</th>
<th>Blast Score</th>
<th>Match length</th>
<th>Query start</th>
<th>Query end</th>
<th>Percent id</th>
<th>DB Xrefence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EK002142</td>
<td>TCM4</td>
<td>target of myb1</td>
<td>22037-1</td>
<td>1041</td>
<td>58.20</td>
<td>0.160</td>
<td>1231</td>
<td>6520000</td>
<td>72381</td>
<td>67</td>
<td>ECTC45 (58200) EK25492</td>
</tr>
<tr>
<td>2</td>
<td>EK002144</td>
<td>NOS2</td>
<td>noggin precursor</td>
<td>17014-12</td>
<td>2849</td>
<td>80.80</td>
<td>0.389</td>
<td>893</td>
<td>47305</td>
<td>74694</td>
<td>30</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>3</td>
<td>EK002706</td>
<td>ASH1A1</td>
<td>erk like protein, ATP-binding, homolog 1</td>
<td>19853-12</td>
<td>856</td>
<td>80.30</td>
<td>0.736</td>
<td>6902</td>
<td>77323</td>
<td>77435</td>
<td>32</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>4</td>
<td>Ks.164529</td>
<td>ASH1A1</td>
<td>erk like protein, ATP-binding, homolog 1</td>
<td>19853-12</td>
<td>856</td>
<td>80.30</td>
<td>0.736</td>
<td>6902</td>
<td>77323</td>
<td>77435</td>
<td>32</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>5</td>
<td>Ks.000233</td>
<td>TMC2</td>
<td>vireno-induced dysphoric factor (best effect, best epitope)</td>
<td>1754-12</td>
<td>56597</td>
<td>88.30</td>
<td>0.224</td>
<td>224</td>
<td>97000</td>
<td>81741</td>
<td>59</td>
<td>C05930 (97000) EK25492</td>
</tr>
<tr>
<td>6</td>
<td>Ks.000245</td>
<td>TMC1A1</td>
<td>vireno-induced dysphoric factor 2-like protein 1</td>
<td>1953-12</td>
<td>5401</td>
<td>84.45</td>
<td>0.304</td>
<td>219</td>
<td>80324</td>
<td>91304</td>
<td>68</td>
<td>C05930 (97000) EK25492</td>
</tr>
<tr>
<td>7</td>
<td>M.301151</td>
<td>TMC2C1</td>
<td>vireno-induced dysphoric factor 2-like protein 2</td>
<td>1733-23</td>
<td>96672</td>
<td>97.30</td>
<td>0.116</td>
<td>116</td>
<td>80008</td>
<td>80080</td>
<td>88</td>
<td>C07370 (116000) EK25492</td>
</tr>
<tr>
<td>8</td>
<td>Ks.000243</td>
<td>TKOC1B</td>
<td>hypothetical protein</td>
<td>1341-12</td>
<td>6651</td>
<td>76.65</td>
<td>0.161</td>
<td>101</td>
<td>80301</td>
<td>93011</td>
<td>72</td>
<td>C05930 (97000) EK25492</td>
</tr>
<tr>
<td>9</td>
<td>Ks.000236</td>
<td>TUBR</td>
<td>jun B proto-oncogene</td>
<td>1953-12</td>
<td>2706</td>
<td>98.30</td>
<td>0.190</td>
<td>190</td>
<td>80390</td>
<td>90770</td>
<td>52</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>10</td>
<td>Ks.000255</td>
<td>TUBR</td>
<td>jun B proto-oncogene</td>
<td>1754-12</td>
<td>2706</td>
<td>98.30</td>
<td>0.190</td>
<td>190</td>
<td>80390</td>
<td>90770</td>
<td>52</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>11</td>
<td>Ks.000234</td>
<td>TUBR</td>
<td>jun D proto-oncogene</td>
<td>1953-12</td>
<td>2706</td>
<td>98.30</td>
<td>0.190</td>
<td>190</td>
<td>80390</td>
<td>90770</td>
<td>52</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>12</td>
<td>M.304148</td>
<td>TUBR</td>
<td>jun B proto-oncogene</td>
<td>1953-12</td>
<td>2706</td>
<td>98.30</td>
<td>0.190</td>
<td>190</td>
<td>80390</td>
<td>90770</td>
<td>52</td>
<td>C12231 (58200) EK25492</td>
</tr>
<tr>
<td>13</td>
<td>Ks.000245</td>
<td>TKOC1B</td>
<td>vireno-induced dysphoric factor (best effect, best epitope)</td>
<td>1341-12</td>
<td>6651</td>
<td>76.65</td>
<td>0.161</td>
<td>101</td>
<td>80301</td>
<td>93011</td>
<td>72</td>
<td>C05930 (97000) EK25492</td>
</tr>
<tr>
<td>14</td>
<td>Ks.000245</td>
<td>TKOC1B</td>
<td>vireno-induced dysphoric factor 2-like protein 1</td>
<td>1953-12</td>
<td>5401</td>
<td>84.45</td>
<td>0.304</td>
<td>219</td>
<td>80324</td>
<td>91304</td>
<td>68</td>
<td>C05930 (97000) EK25492</td>
</tr>
<tr>
<td>15</td>
<td>Ks.000245</td>
<td>TKOC1B</td>
<td>vireno-induced dysphoric factor 2-like protein 2</td>
<td>1733-23</td>
<td>96672</td>
<td>97.30</td>
<td>0.116</td>
<td>116</td>
<td>80008</td>
<td>80080</td>
<td>88</td>
<td>C07370 (116000) EK25492</td>
</tr>
<tr>
<td>16</td>
<td>Ks.000245</td>
<td>TKOC1B</td>
<td>vireno-induced dysphoric factor 2-like protein 3</td>
<td>1953-12</td>
<td>5401</td>
<td>84.45</td>
<td>0.304</td>
<td>219</td>
<td>80324</td>
<td>91304</td>
<td>68</td>
<td>C05930 (97000) EK25492</td>
</tr>
<tr>
<td>17</td>
<td>Ks.000245</td>
<td>TKOC1B</td>
<td>vireno-induced dysphoric factor 2-like protein 4</td>
<td>1953-12</td>
<td>5401</td>
<td>84.45</td>
<td>0.304</td>
<td>219</td>
<td>80324</td>
<td>91304</td>
<td>68</td>
<td>C05930 (97000) EK25492</td>
</tr>
</tbody>
</table>

**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 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.

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**