

Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence

Hugues Roest Crolius, Olivier Jaillon, Alain Bernot, Corinne Dasilva, Laurence Bouneau, Cécile Fischer, Cécile Fizames, Patrick Wincker, Philippe Brottier, Francis Quétier, William Saurin & Jean Weissenbach

The number of genes in the human genome is unknown, with estimates ranging from 50,000 to 90,000 (refs 1,2), and to more than 140,000 according to unpublished sources. We have developed 'Exofish', a procedure based on homology searches, to identify human genes quickly and reliably. This method relies on the sequence of another vertebrate, the pufferfish *Tetraodon nigroviridis*, to detect conserved sequences with a very low background. Similar to *Fugu rubripes*, a marine pufferfish proposed by Brenner *et al.*³ as a model for genomic studies, *T. nigroviridis* is a more practical alternative⁴ with a genome also eight times more compact than that of human. Many comparisons have been made between *F. rubripes* and human DNA that demonstrate the potential of comparative genomics using the pufferfish genome⁵. Application of Exofish to the December version of the working draft sequence of the human genome and to Unigene showed that the human genome contains 28,000–34,000 genes, and that Unigene contains less than 40% of the protein-coding fraction of the human genome.

To determine the conditions that would generate alignments in coding regions between human DNA and a pufferfish distant by 400 million years, we first tested a large number of BLAST conditions on a small set of 13 annotated human-pufferfish homologous genes (Table 1). We used *F. rubripes* genes because no complete *T. nigroviridis* gene sequence existed at the time of this work. We then applied the optimal conditions to a larger set of 322 annotated human genes and the partial *T. nigroviridis* genome sequence (33% of which has been determined), in which the positions of genes are unknown. We found that the existing sequence of the *T. nigroviridis* genome detects 26.5% of the 2,693 human exons in conditions in which no alignments fall in introns (Fig. 1a). The 724 exons detected are distributed in 64.9% of the genes (209/322). To estimate the influence of the amount of *T. nigroviridis* genome sequenced on the sensitivity of this approach

in detecting exons and genes in human DNA, we represented the fraction of exons and genes identified with increasing amounts of *T. nigroviridis* sequence (Fig. 1b). The fraction of human exons detected increases at a rate proportional to the amount of *T. nigroviridis* genome coverage generated. The probability of identifying a gene by at least one of its exons is higher because genes in general contain many exons, in addition to the fact that the random sequence tag (RST) database represents approximately 170,000 random sequences in the genome.

To reflect the fact that different *T. nigroviridis* sequences may generate overlapping alignments over the same exon and define a single, conserved human region, we defined the contiguous assembly of the different overlapping alignments as an 'ecore' (for evolutionary conserved region). In the set of 322 reference genes, the 209 genes (or 724 exons) that were detected by *T. nigroviridis* contained 831 ecores (2.58 ecores per gene). This result (Fig. 1a) provides a means to decide if new alignments between human and *T. nigroviridis* DNA overlap human exons, based on their length and percentage of identity. This criterion is the basis of the Exofish (for exon finding by sequence homology) selection mechanism (Fig. 2). To confirm the sensitivity of Exofish in detecting human genes, we performed a second comparison on a set of 4,888 complete human cDNA sequences extracted from Unigene version 105 (ref. 6). Using this set, 70% of the genes were identified, and each gene contained an average of 3.18 ecores (including the 30% of undetected genes). This ratio was used to derive a number of genes from a given number of ecores detected by Exofish.

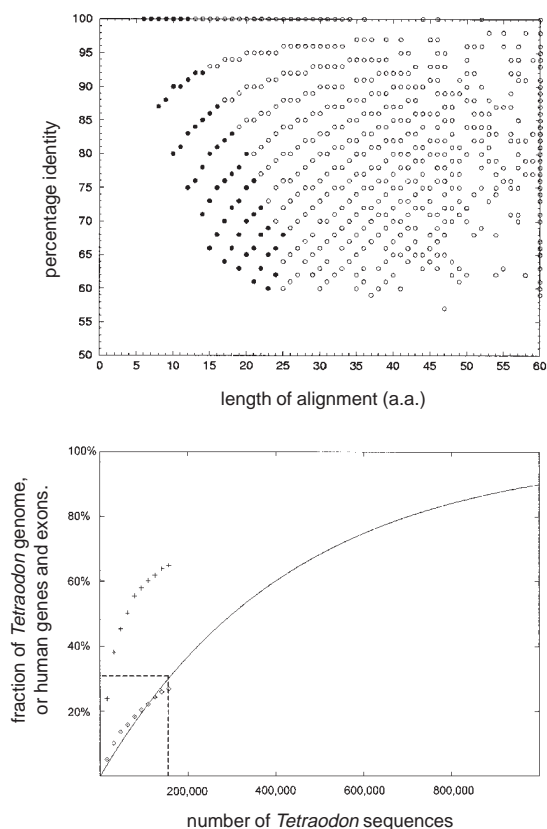
We analysed the sequence of chromosome 22 (ref. 7) with Exofish to estimate its capacity to confirm existing annotations and to detect new genes. We found 1,525 ecores over the complete length of the chromosome (Fig. 3). The distribution of ecores among the different types of annotated features showed

Table 1 • Performance of different BLAST configurations

Method	Matrix	W	X	L	I (%)	Sn (%)	Sp (%)	T (s)
BLASTN	NUC.4.4	8 bases	5	30 bases	70	66	93	4.8
BLASTN	NUC.4.4	8 bases	9	40 bases	70	76	94	5.7
BLASTN	NUC.4.4	10 bases	13	30 bases	70	68	40	4.3
TBLASTX	BLOSUM62	3 aa	9	13 aa	60	85	55	74.8
TBLASTX	BLOSUM62	4 aa	3	13 aa	70	80	94	1,065.2
TBLASTX	BLOSUM62	5 aa	1	13 aa	70	84	96	1,160.9
TBLASTX	CNS	4 aa	25	13 aa	70	85	96	10.0
TBLASTX	CNS	5 aa	13	13 aa	70	85	96	29.4
TBLASTX	CNS	5 aa	25	13 aa	70	89	94	29.3

Each program was run with 1,340 different conditions, and a representative selection of results is shown. A range of values for W (initial size of the search word) and X (threshold score for consecutive mismatching residues or bases) were tested. For amino acid alignments, a non-substitutive matrix (CNS, match = +15, mismatch = -12) was tested as well as the standard BLOSUM62 matrix. A minimal length (L) and percentage identity (I) were applied to select alignments for which a sensitivity (Sn) and specificity (Sp) were calculated in terms of numbers of overall matching exons. T indicates the time in seconds needed to compare the 13 homologues against each other. The last row shows the optimal performance that was retained for Exofish.

Genoscope and CNRS FRE2231, Evry cedex, France. Correspondence should be addressed to J.W. (jsbach@genoscope.cns.fr).



important variations (Table 2). Related genes (based on homologies to protein and genes from human and other species) and predicted genes (based on EST sequences) contained less ecores than known genes. These two categories of annotations also contained less annotated exons per gene, presumably because their respective counterparts in sequence databases are only partially homologous or are incomplete. In fact, 70 of 148 predicted genes consisted of a single exon. We estimate that approximately 50% of the 181 ecores that fell outside of annotations belonged to genes that are incompletely annotated or to pseudogenes. Therefore, the remaining 90 ecores corresponded to approximately 30 novel genes on chromosome 22 (Fig. 3*b,c*). We thus estimate that chromosome 22 contains less than 600 genes.

Of the 1,344 ecores that fell within the boundaries of the annotations, 1,197 (89%) corresponded to genes and 147 (11%) to pseudogenes. To estimate the sensitivity of Exofish in detecting genes on chromosome 22, we considered only the 247 known genes, because others are likely to be incomplete. Ecores were found in 32.0% of the 2,298 exons and 66.8% of the 247 known genes. These values are comparable to the 26.5% of exons and 64.5% of genes identified in the reference set of 322 human genes, and to the 70% sensitivity obtained on 4,888 full-length cDNA sequences. Exofish detects only 8% of the 325 genes predicted by Genscan that are not confirmed by homologies (compared with 64.5% for known genes), suggesting that most of these predictions are false positives.

It is possible to exploit the compactness of the *T. nigroviridis* genome to confirm that several neighbouring ecores that fell outside of existing annotations do belong to the same gene. For instance, the five isolated ecores (Fig. 3*c*) were joined by three *T. nigroviridis* RSTs. Subsequent to the release of the sequence of chromosome 22 (ref. 7), a human cDNA clone and a homologous gene in a *Caenorhabditis elegans* cosmid clone have confirmed that these ecores define a true gene. By contrast, ecores

Fig. 1 Construction of Exofish. **a**, Distribution of 8.3 million alignments generated by comparing the partial *T. nigroviridis* genome with a set of 322 human genes (2,693 exons). Each circle represents a population of alignments of a given length and a given percentage of identity, with a clear boundary between those which exclusively fall in exons (○) of human genes and those for which at least one alignment falls in an intron (●). This provides robust selection criteria to determine if any new alignment corresponds to a human exon, based on its length and identity with a *T. nigroviridis* sequence. For convenience, all alignments longer than 60 aa were arbitrarily drawn at 60, the longest measuring 245 aa. **b**, Evolution of the theoretical *T. nigroviridis* genome coverage (—) and observed sensitivity in gene detection (+) and exon detection (◊) by Exofish in the set of 322 human genes, as a function of *T. nigroviridis* sequences produced (10% increments). The dotted line is positioned at the current status of the sequencing project (33% of genome coverage). The theoretical coverage is calculated on the basis of a Poisson distribution of sequences of average size 886 bases on a genome of 385 Mb.

identified inside the boundaries of the 545 annotated genes, but outside exons (that is, in introns), would correspond to exons that remained undetected by other homology-based approaches, presumably because of alternative splicing. We found 25 ecores in the introns of 21 annotated genes, of which 19 were also predicted by Genscan (Fig. 3*d*). Approximately 50% of ecores that fell either in introns or outside of annotations have been confirmed as exons by the chromosome 22 annotation team at the Sanger Centre (J. Collins, D. Beare and I. Dunham, pers. comm.).

To estimate the number of genes in the human genome, we analysed the human working draft sequence with Exofish. In release 61 (December 1999) of the EMBL database, the publicly available human working draft sequence contained 1,272.3 Mb of non-redundant human DNA. Analysis of this fraction of the human genome (~42.4%) by Exofish generated 42,066 ecores. Results on human chromosome 22 indicated that 89% of ecores fell in genes, whereas the remaining 11% fell in pseudogenes. Based on the result that Exofish detects on average 3.18 ecores per human gene, the human genome would contain $(42,066 \times 0.89) / 0.424 = 88,299$ ecores and $88,299 / 3.18 = 27,767$ genes. We estimated the gene distribution for each chromosome and compared the results with the EST gene map of the human genome⁸ (Fig. 4). The gene-dense chromosomes (17, 19, 22) have an excess of ecores compared with ESTs, as does chromosome 16. To set an upper limit to our estimate, another calculation was based on the lower ratio of ecores per gene found in the initial gene test set and

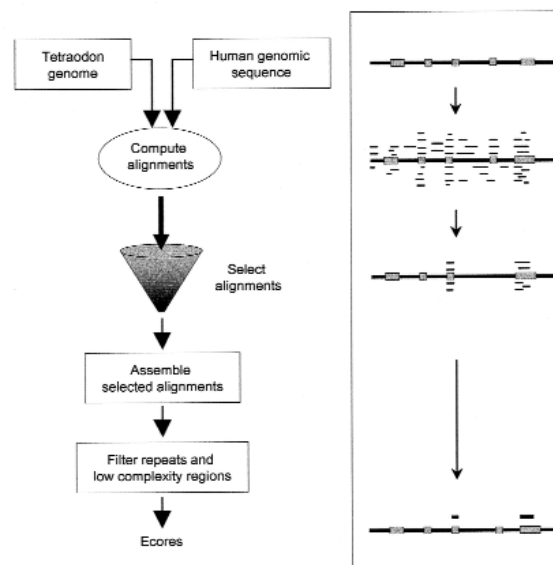


Fig. 2 Schematic of Exofish.

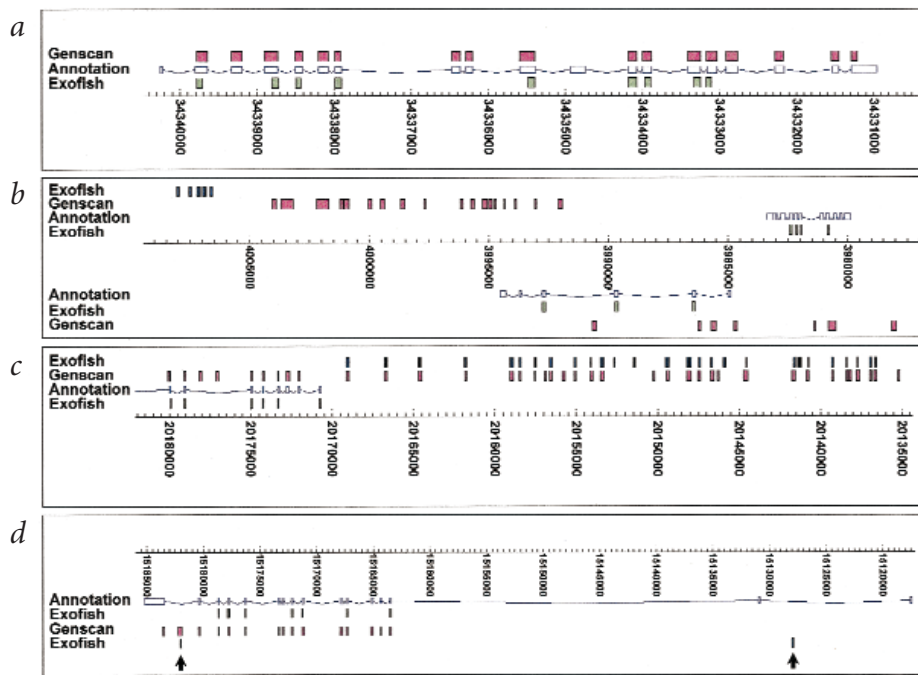


Fig. 3 Examples of chromosome 22 results. Open blue boxes linked by broken lines represent gene annotations from ref. 7. Red boxes represent exons predicted by Genscan. Green boxes represent exons predicted by Exofish that overlap gene annotations, and dark blue boxes represent exons that do not overlap annotations. The scale in nucleotides is relative to the sequence described in ref. 7. **a**, Typical result in which a known gene with 19 exons (encoding carnitine palmitoyltransferase I) is partially predicted by Genscan (17 exons) and Exofish (9 exons). **b**, On the 'Up' strand (above the scale), five exons indicate a new gene that is not predicted by Genscan, whereas two genes on the top and the bottom strand (similar to mouse *Htf9c* and *Ranbp1*, respectively) have several exons predicted by Exofish, whereas none are correctly predicted by Genscan. **c**, On the 'Up' strand, both Genscan and Exofish partially confirm a known gene (*HMG2L1*), whereas on the same strand a new gene seems to be predicted by both approaches. **d**, *LIMK2* has 16 annotated exons, of which 14 are predicted by Genscan and 6 by Exofish. Two additional exons that are presumably alternatively spliced are predicted by Exofish (arrows), one of which is also predicted by Genscan.

gave $88,299/2.58=34,224$ genes. We therefore estimate that the human genome contains 28,000–34,000 genes.

We used Exofish on the non-redundant set of human gene sequences represented by Unigene⁶ to estimate more accurately the fraction of protein-coding DNA present in publicly available databases. Release 105 of Unigene contains 10,501 clusters represented by known genes, whereas the remaining 82,430 clusters only contain EST sequences. When matched to the selected sequences representing Unigene clusters, Exofish detected 33,079 exons, which identified 62% of the 10,501 known genes and only 4.2% of the EST sequences. As the human genome is estimated to contain 88,299 exons, the 33,079 exons found in Unigene represent only 37.5% of the coding fraction of human genes. This result is coherent with the very low number of matches obtained on the EST sequences. Most selected ESTs representing Unigene clusters (87%) are 3' reads of cDNA clones, and most likely correspond to untranslated regions.

Because a genome is a finite entity that contains all genes with all exons necessary to express all the proteins required at any stage or in any tissue, the sensitivity of Exofish is not biased by the traditional problems encountered in cDNA databases, such as alternative splicing and varying gene-expression levels. This is confirmed by the fact that Exofish identifies the same fraction of genes ($\sim 2/3$) in three collections of human genes of diverse origins and characteristics. Our finding that the human genome contains only 28,000–34,000 genes is unexpected, considering

that it corresponds to just over twice the number of genes in the fly or worm. It is therefore to be expected that organismal complexity is not a direct consequence of gene number, but has its source in other mechanisms that may include alternative splicing and multi-domain proteins. As Unigene contains 92,000 clusters, and Exofish predicts 28,000–34,000 genes in the genome, Unigene is partially redundant and also contains mostly non-coding sequences. It is likely, however, that Unigene contains a 'tag' for most human genes and as such is an invaluable resource for gene identification. Exofish still cannot detect one-third of human genes (false negatives), including those for which the corresponding *T. nigroviridis* sequence is not yet known, those that evolve rapidly and for which protein sequence similarity is weak, and those that are strictly specific to mammals. It is likely, however, that smaller protein domains also participate in the detection process and enable Exofish to detect genes outside the limits of orthologous or paralogous genes. As described here, two immediate applications for Exofish include the annotation of genomic DNA and the estimation of the coding fraction in cDNA collections. Exofish also enables comparison of the *T. nigroviridis* genome with entire vertebrate genomes at the protein level in a few hours of computation time, and as such it is a powerful tool to explore new avenues in vertebrate genome research in a way so far only possible for bacteria or unicellular eukaryotes.

Table 2 • Distribution of exons in chromosome 22 annotations

Feature	No. of features on chromosome 22	No. of exons	Average no. of exons per feature	Average no. of exons per feature	No. of features identified by Exofish	% features identified by Exofish
known genes	247	848	3.44	9.11	165	66.8%
related genes	150	289	1.93	5.24	83	55.3%
predicted genes	148	60	0.41	3.03	22	14.8%
pseudogenes	134	147	1.10	1.66	62	46.3%
outside annotations	–	181	–	–	–	–
Genscan genes	817	1,330	1.63	8.17	307	37.6%
Genscan genes outside annotations	325	49	0.15	4.71	26	8.0%

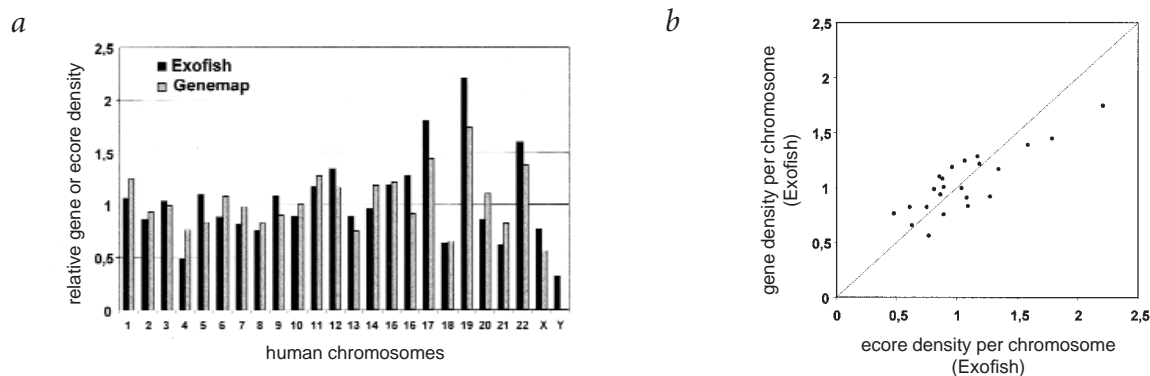


Fig. 4 Distribution of gene and ecores on individual human chromosomes according to the EST physical map⁸ and Exofish. **a**, Exofish confirms the density of genes obtained by EST mapping for most chromosomes, except chromosomes 16, 17, 19 and 22, and introduces an estimate for chromosome Y. **b**, The two independent data sets show a good correlation (correlation factor 0.88), which confirms (in most cases) the distribution obtained by physical mapping of ESTs.

Methods

Construction of Exofish. A summary of the approach used to select the optimal BLAST conditions for Exofish is shown (Table 1) and a full description is available (see Methods, http://genetics.nature.com/supplementary_info). Exofish is available as an annotation tool for human sequences (<http://www.genoscope.cns.fr/exofish>). We constructed a set of 322 complete human genes by global BLASTN alignments between a database of 10,067 human mRNA sequences and 3,930 genomic clones. Details of the parameters and selections used, as well as a file of the 322 human genes in fasta format, are available (see Methods, http://genetics.nature.com/supplementary_info).

***T. nigroviridis* genomic sequence.** BAC library construction and insert-end sequencing are described elsewhere⁹ Genomic DNA from a male *T. nigroviridis* specimen (ascertained as *T. nigroviridis* using morphological and mitochondrial DNA sequence characteristics) was extracted to construct a plasmid library. DNA was mechanically sheared, separated on a preparative agarose gel, and a size fraction corresponding to ~3 kb was excised, end-repaired and cloned in pcDNA2. After electroporation in DH10B electrocompetent cells, clones were plated on 2YT agar plates containing 70 µg/ml carbenicillin and 100,000 recombinants were robotically picked and replicated in microtitre plates. We sequenced 127,229 insert ends as described⁹. Including the BAC ends database, 174,828 sequences of an average useful length of 886 nt were produced, equivalent to 154.9 Mb of combined DNA. *T. nigroviridis* repeats (rRNAs, transposable elements, satellites) and microsatellite repeats were masked following BLASTN alignments against a *T. nigroviridis* repeat database⁹ and microsatellite database, respectively¹⁰. Minisatellites were identified by Tandem Repeat Finder¹¹ and subsequently masked. Microsatellites were further masked based on TBLASTX alignments, and low-complexity regions were identified and masked by RepeatMasker. In total, 11% of nucleotides were masked in the *T. nigroviridis* genomic sequence database.

Human working draft sequence and Unigene. We retrieved the entire HTG1, HTG2 and HTG3 sections, and sequences larger than 35 kb from

the human DNA section of EMBL release 61, and for each sequence the highest version number was retained to remove internal redundancy. Sequences were distributed as follows: HTG1, containing genomic clone sequences in unordered segments (726.9 Mb, 24.2%); HTG2, in which contigs were ordered within each genomic clone (55.5 Mb, 1.8%); HTG3, in which genomic clones are represented as a contiguous sequence (36.4 Mb, 1.2%); and HUM, in which sequences are considered finished, with an error rate of less than one in 10⁴ bases (480.9 Mb, 16.0%). All sequences were filtered to remove remaining cloning vector sequences (0.21%) and stretches of 'N' used to separate sequence contigs in HTG1 and HTG2 (1.87%). We used Unigene version 105 (January 2000) for comparisons with the *T. nigroviridis* genome.

Computing alignments. All alignments were computed with the suite of BLAST (ref. 12) algorithms or with the SMITH-WATERMAN (ref. 13) algorithm, implemented in LASSAP (Large Scale Sequence Comparison Package) version 1.1.5 (ref. 14). For all calculations, hardware consisted of four Digital quadriprocessor (AXP 21264 (EV6) at 525 MHz) computers (Compaq GS60) with 4 Go of memory each, except for comparison of the partial *T. nigroviridis* genome with the human working draft sequence, for which a SUN Enterprise 10000 server with 64 UltraSPARC-II (400 MHz) processors and 64 Go central memory were used with LASSAP version 1.2.0a.

Accession numbers. *T. nigroviridis* sequences, EMBL AL163976 to AL352938; human cDNA clone, AB033118.

Acknowledgements

We thank the sequencing and template preparation team at Genoscope; Sun Microsystems for access to the SUN benchmark centre; and F. Francis for critical reading of the manuscript. This work would not have been possible without the public availability of a large fraction of the sequence of the human genome, and we thank all contributing genome centres.

Received 10 March; accepted 2 May 2000.

- Fields, C., Adams, M.D., White, O. & Venter, J.C. How many genes in the human genome? *Nature Genet.* **7**, 345–346 (1994).
- Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90**, 11995–11999 (1993).
- Brenner, S. et al. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**, 265–268 (1993).
- Crnogorac-Jurcevic, T., Brown, J.R., Lehrach, H. & Schalkwyk, L.C. Tetraodon *fluviatilis*, a new puffer fish model for genome studies. *Genomics* **41**, 177–184 (1997).
- Elgar, G. et al. Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Res.* **9**, 960–971 (1999).
- Schuler, G.D. et al. A gene map of the human genome. *Science* **274**, 540–546 (1996).
- Dunham, I. et al. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Deloukas, P. et al. A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
- Roest Croliuis, H. et al. Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* (in press).
- Jin, L., Zhong, Y. & Chakraborty, R. The exact numbers of possible microsatellite motifs. *Am. J. Hum. Genet.* **55**, 582–583 (1994).
- Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
- Gleimet, E. & Codani, J. LASSAP, a large scale sequence comparisons package. *Comput. Appl. Biosci.* **13**, 137–143 (1997).