WABA Success: A Tool for Sequence Comparison between Large Genomes

David L. Baillie[1] and Ann M. Rose[2,3]

[1]*Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6 Canada;* [2]*Department of Medical Genetics, University of British Columbia, Vancouver V6T 1Z3 Canada*

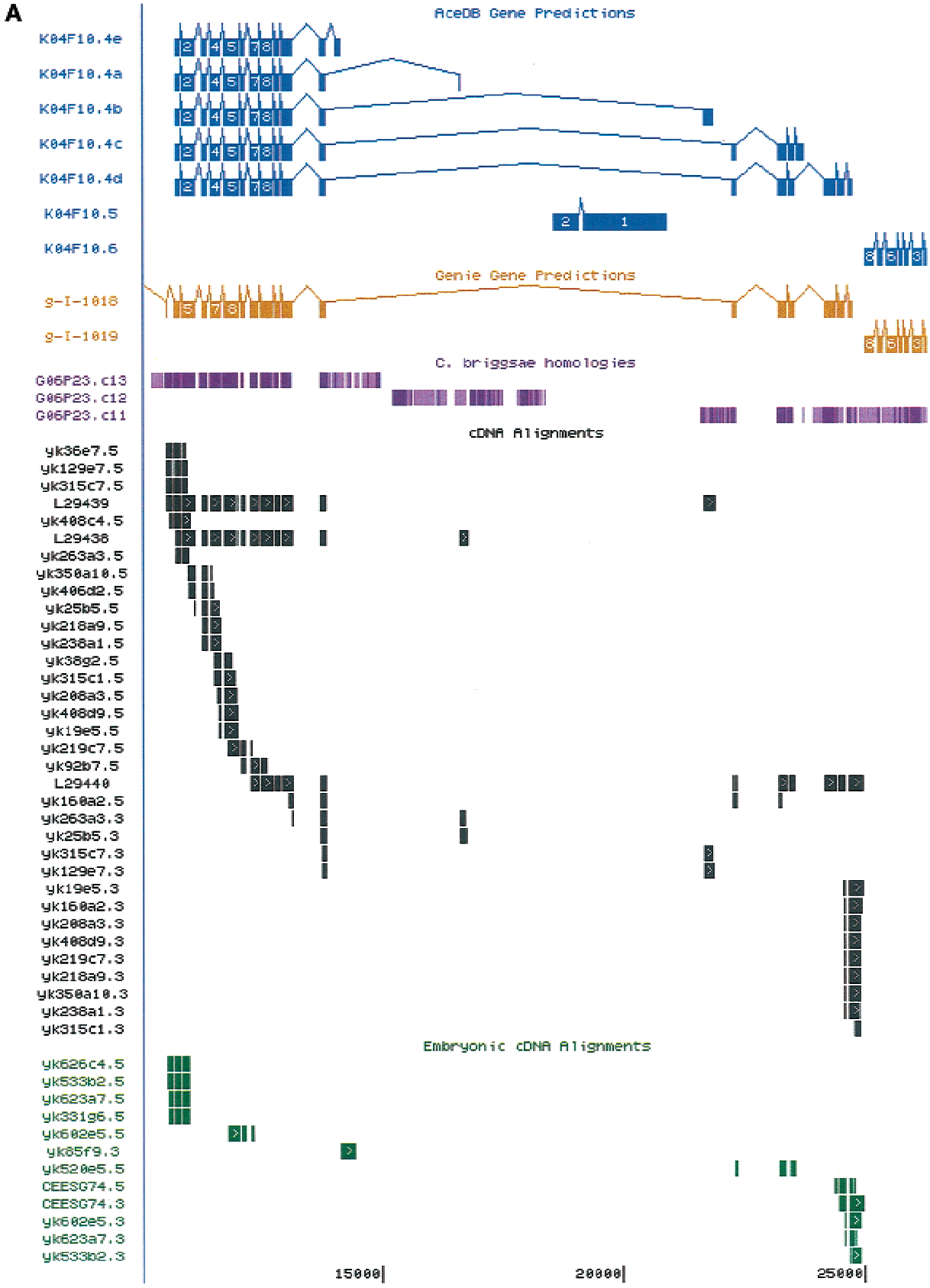
Whole-genome sequence comparisons between bacterial sequences are one thing, but try comparing two eukaryotic genomes, each containing tens or hundreds of millions of nucleotides. And try to do it on your desktop machine in your office or at home. That is what Kent and Zahler (2000) have tried, and the results are presented in this issue of *Genome Research*. The use of evolutionary conservation to unveil functional information contained within genomes is not new. In the case of the nematode, comparisons of *Caenorhabditis elegans* to its close relative *Caenorhabditis briggsae* go back as far as Emmons et al. (1979). Snutch (1984) made the first *C. briggsae* genomic library available to the research community. These two nematodes are almost identical in morphology and development, yet their genomes have been separated for a sufficient amount of time to allow intronic and intragenic sequences to become effectively randomized, while protein-encoding sequences and *cis*-linked regulatory elements are conserved. *C. briggsae* has diverged from *C. elegans* in regions of unselected sequence, the middle of large introns, between genes, and at the third nucleotide of synonymous codons in genes not abundantly translated. Based on the now-outdated concept of a constant evolutionary clock, these two species were estimated to have diverged some 30–60 million years ago. These facts prompted a *C. briggsae* genome sequencing effort to be initiated by The Washington University Genome Sequencing Center in St. Louis, Missouri. The project was given a jumpstart with the construction of a physical map (Marra et al. 1997) and an open invitation to the research community to participate in the selection of fosmids for sequence analysis. Researchers from around the world probed microarrayed fosmid filters with their favorite gene and contacted The Genome Sequencing Center with a request to sequence the identified fosmid. Currently, approximately 10% of the genome has now been completed and is available at http://www.genome.wustl.edu/pub/gsc1/sequence/st.louis/briggsae/finish/. Comparison of *C. elegans* and *C. briggsae* sequence has facilitated the identification of *cis*-linked regulatory elements (Heine and Blumenthal 1986), the cloning of genes as a result of their syntenic relationship (Kuwabara and Shah 1994), and interpretation of complex gene structure (Thacker et al. 1999). In many cases examined, adjacent genes are conserved both in position and orientation, with the occasional interruption caused by a transposable element or an apparent pseudogene.

Prior to the availability of *C. briggsae* sequence, gene feature identification was done by abinitio prediction. Many computer programs exist that attempt to predict the exon-intron structure of genes from genomic sequence. These programs vary in their accuracy and are at their worst when asked to predict the first and last exons of genes, often failing to identify correctly exons and introns outside the actual coding element. Messenger RNA and EST-based feature detection is often confounded by large transcripts, genes that have low levels of transcription, or genes lacking poly(A)$^+$-containing transcripts. Further complications adding to the problem of gene finding include the still-murky rules of sequence identification unknown for numerous important genomic elements (snRNAs, ribozymes, transcription factor binding sites, replication origins, chromatin folding and packaging signals, meiotic pairing information, etc.). Because of these issues, the initial annotation of the *C. elegans* genome was done interactively using GENEFINDER, a program written by Phil Green and LaDeana Hillier (unpubl.). The program, which was ahead of its time, did a remarkable job of gene prediction. Taken together with manual interpretation, the cDNA data (Kohara 1996) and related genomic sequence data from *C. briggsae*, gene structure predictions were made for > 19,000 *C. elegans* genes.

Kent and Zahler (2000) have taken advantage of the availability of genomic sequence from these two closely related nematodes to test the feasibility of doing large-scale alignments between genomic DNA of different species. They have developed an algorithm for sequence comparison in which every third base (the wobble position of the codon) is ignored. The wobble-aware bulk aligner (WABA) allows the sensitive identification of conserved coding regions. They use this algorithm in a three-tiered process to identify conserved regions between eight million base pairs of *C. briggsae* genomic sequence and the entire 97 million base pairs of *C. elegans*. Their analysis was performed on a readily available 450 mHz Intel-based machine. The results they have achieved are remarkable and will provide a useful resource for all in

[3]**Corresponding author.**
E-MAIL arose@gene.nce.ubc.ca; FAX (604) 822-5348.

10:1071–1073 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org


A
AceDB Gene Predictions
K04F10.4e
K04F10.4a
K04F10.4b
K04F10.4c
K04F10.4d
K04F10.5
K04F10.6
Genie Gene Predictions
g-I-1018
g-I-1019
C. briggsae homologies
G06P23.c13
G06P23.c12
G06P23.c11
cDNA Alignments
yk36e7.5
yk129e7.5
yk315c7.5
L29439
yk408c4.5
L29436
yk263a3.5
yk350a10.5
yk406d2.5
yk25b5.5
yk218a9.5
yk238a1.5
yk38g2.5
yk315c1.5
yk208a3.5
yk408d9.5
yk19e5.5
yk219c7.5
yk92b7.5
L29440
yk160a2.5
yk263a3.3
yk25b5.3
yk315c7.3
yk129e7.3
yk19e5.3
yk160a2.3
yk208a3.3
yk408d9.3
yk219c7.3
yk218a9.3
yk350a10.3
yk238a1.3
yk315c1.3
Embryonic cDNA Alignments
yk626c4.5
yk533b2.5
yk623a7.5
yk331g6.5
yk602e5.5
yk85f9.3
yk520e5.5
CEESG74.5
CEESG74.3
yk602e5.3
yk623a7.3
yk533b2.3
15000
20000
25000

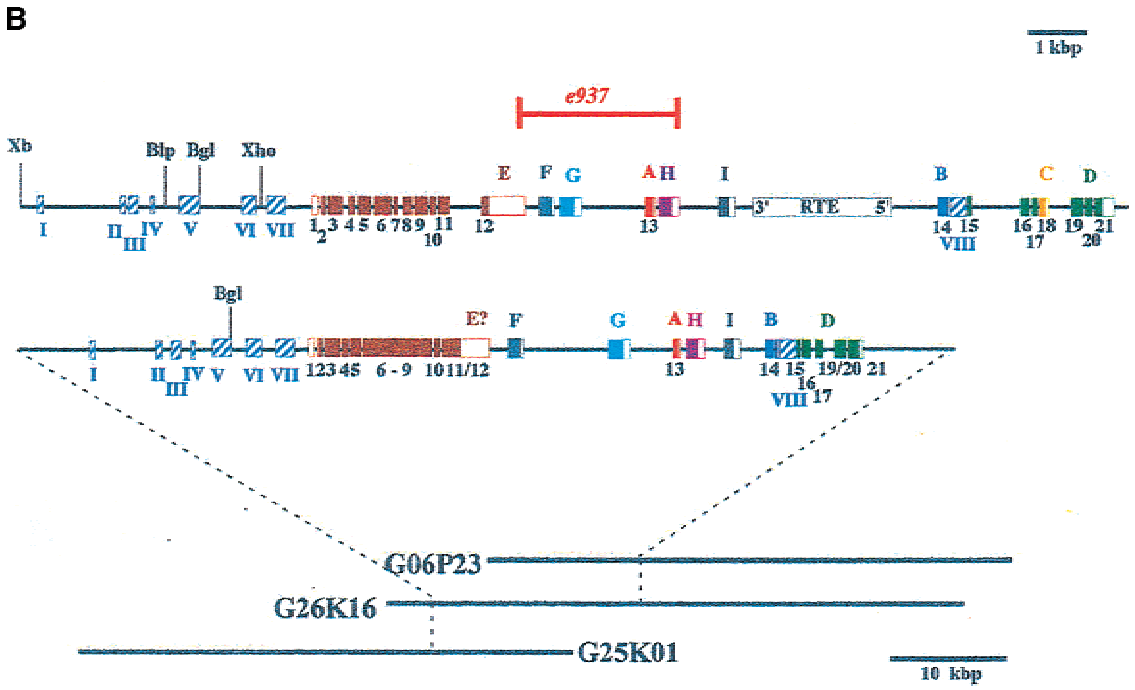


Figure 1 (*A*) Exon prediction in genomic regions near *bli-4* (also called K04F10; from Kent and Zahler at http://www.cse.ucsc.edu/~kent/cgi-bin/tracks.exe?where=K04F10). (*B*) Schematic representation of the *bli-4* gene from *C. elegans* (*top*) and *C. briggsae* (*bottom*). The common region(s) encoding the signal peptide, prodomain, protease domain, and middle domain are shown in brown. Alternatively spliced exons that encode carboxyl termini unique to the individual isoforms are labeled alphabetically and color coded. The position and extent of the *e937* 3325-bp deletion is indicated. Open boxes represent noncoding exons or untranslated regions; shaded boxes represent coding exons. Hatched boxes (labeled I–VIII) represent regions of nucleotide homology that may constitute regulatory elements, particularly those at the 5′ end. (Xb) *Xba*I; (Blp) *Blp*I; (Bgl) *Bgl*II; (Xho) *Xho*I. The approximate location of the *C. briggsae bli-4* gene is indicated by dashed lines on the fosmid clones used in Thacker et al. (1999). The entire sequence of clones G25K01 and G06P23 has been determined, whereas G26K16 was used solely for transformation rescue experiments. (Reprinted with permission from Thacker et al. 1999.)

the continuing analysis of genomes. One of the examples given by the authors is the *bli-4* (K04F10.4) gene. The structure of this gene was characterized previously by examination of the conservation between *C. elegans* and *C. briggsae* (Thacker et al. 1999; Fig. 1B). Traditionally, gene prediction programs have been forced to ignore the computationally complex problem of ab-initio prediction of splice variants. Neither ACEDB nor the Genie program predict the entire nine alternate transcripts described in Thacker et al. (1999). Figure 1A shows the WABA/intronator display for the gene. The areas of *C. briggsae* homology identified by the WABA program would greatly assist researchers by alerting them to the strong possibility that alternative transcripts may exist. The intronator display brings together gene predictions, *C. briggsae* homologies, and cDNA information in an easily used and intuitive fashion.

One surprising observation made by Kent and Zahler (2000) was the shortness of the syntenic segments, which is partly a consequence of the length of the *C. briggsae* clones. The number of fragments that clones are broken into by the alignment exhibited a bimodal distribution, which corresponds to some extent to the position of the sequence on the *C. elegans* autosomes. In the gene-rich central clusters, long alignments were observed, predicted to constitute approximately 40% of the genome. In contrast, the flanking arms were more susceptible to rearrangement. It has been known for some time that meiotic recombination is much higher on the chromosome arms. This fact and the observation that sequence similarities to organisms other than nematodes tends to be lower on the arms led the *C. elegans* Sequencing Consortium (1998) to suggest that the DNA in the arms might be evolving more rapidly than that in the central regions of the autosomes.

Computational tools, like WABA, will be essential for future analysis of large vertebrate genomes. Computational comparative genomics is expected to become an integral part of the analysis of human and mouse genomes and will be needed to identify coding elements and other functional components resistant to analysis by more conventional genetic and molecular approaches.

REFERENCES

C. elegans Sequencing Consortium. 1998. *Science* **282:** 2012–2018.

Emmons, S.W., Klass, M.R., and Hirsh, D. 1979. *Proc. Natl. Acad. Sci.* **76:** 1333–1337.

Heine, U. and Blumenthal, T. 1986. *J. Mol. Biol.* **188:** 189–198.

Kohara, Y. 1996. *Protein Nucleic Acid Enzyme* **41:** 715.

Kuwabara, P.E. and Shah, S. 1994. *Nucleic Acids Res.* **22:** 4414–4418.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. *Genome Res.* **7:** 1072–1084.

Snutch, T.P. 1984. Ph.D. thesis, Simon Fraser University, Burnaby, B.C.

Thacker, C., Marra, M.A., Jones, A., Baillie, D.L., and Rose, A.M. 1999. *Genome Res.* **9:** 348–359.