TREBALL DE RECERCA

PER OPTAR AL

DIPLOMA D'ESTUDIS AVANÇATS

(DEA)

Josep Francesc ABRIL FERRANDO †

— 15 de març de 2002—

Genome Informatics Research Lab
Grup de Recerca en Infomàtica Biomèdica
Institut Municipal d'Investigació Mèdica
Universitat Pompeu Fabra

† e-mail: **jabril@imim.es**

Índex

< Id: main.tex,v 1.3 2002/03/15 08:33:54 jabril Exp >

ARTICLES PUBLICATS

A continuació es presenta una llista dels articles publicats en els que entre els autors hi figura el doctorand. Primerament s'indica la pàgina d'aquest document on es poden veure inclosses les dues primeres pàgines de cada article, ja que donat l'extensió d'alguns d'ells s'ha cregut convenient que es poden obtenir a través de la seva referència bibliogràfica; ara bé, el doctorand pot entregar una separata dels mateixos si fòra necessari. S'han inclòs també les figures principals corresponents a algun dels articles per destacar la feina en la que es va participar, aquestes figures estan escalades a la mida de la pàgina amb el que no es pot apreciar el detall de les mateixes, encara que es poden recuperar a través de la referència bibliogràfica corresponent.
Es pot accedir a més informació sobre les publicacions, incloent-hi els seus abstracts, així com els pòsters presentats a congressos científics, amb el corresponent *link* a la versió electrònica, des de la següent adreça d'internet:

`http://www1.imim.es/~jabril/index.html#PAPER`

✯ Visualitzant anotacions genòmiques amb el `gff2ps` [pàgina 7]

Abril, J. F. and Guigó, R. (2000). `gff2ps`: visualizing genomic annotations. *Bioinformatics*, 16(8):743–744

Aquest programa genera, a partir d'un conjunt d'anotacions genòmiques en format GFF[1], una figura en format POSTSCRIPT on es resumeix de manera molt intuitiva la informació. Els avantatges de treballar amb aquest format de gràfics vectorials es poden resumir en els punts següents: és un format independent de dispositiu, al ser de tipus vectorial els gràfics es poden ampliar o reduir amb facilitat i permet treballar amb gran quantitat de dades de manera eficient. Malgrat que la seva publicació va ser a l'agost de l'any 2000, la primera versió operativa del programa ja era accessible a través d'internet des de principis de gener del mateix any, la qual cosa ens va permetre participar el la fase d'anotació del genoma de la mosca del vinagre (*Drosophila melanogaster*). En les següents tres publicacions es va participar directament en la representació gràfica de les anotacions disponibles per a les corresponents seqüències genòmiques. A més, el programa està sent utilitzat pel consorci públic que està seqüenciant i anotant el genoma de *Dictyostelium discoideum*[2], pels investigadors que estan desenvolupant el programa de predicció computacional de gens anomenat `Phat`[3], etc...

Informació complementària:

- ➢ Adreça a internet del programa:
  `http://www1.imim.es/software/gfftools/GFF2PS.html`
- ➢ Manual de l'usuari:
  `http://www1.imim.es/software/gfftools/gff2ps_docs/manual/MANUAL_GFF2PS_v0.96.ps.gz`

✯ Avaluant l'anotació de la regió de 3Mbp *Adh* a *Drosophila melanogaster* [pàgina 8]

Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F., and Lewis, S. E. (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Research*, 10:483–501

En aquest article es comparaven alguns dels mètodes computacionals de predicció de gens amb l'objectiu d'explorar la fiabilitat dels programes al treballar amb seqüències genòmiques d'organismes eucariotes. Els genomes eucariotes, a diferència dels procariotes, presenten la dificultat afegida de que la seva seqüència de DNA no codifica de manera contínua els gens i aquests a la

[1] `http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml`
[2] `http://genome.imb-jena.de/dictyostelium/`
[3] `http://stat-www.berkeley.edu/users/scawley/Phat/README.html`

seva vegada tampoc estan codificats de manera contínua, però a més, la distribució dels gens al llarg de la seqüència tampoc és uniforme, ni tan sols les longituts dels introns. Fins la publicació d'aquest article no es dispossava de seqüències llargues que continguessin múltiples gens i que haguèssin estat revisades amb cura per experts. Dotze grups, entre ells el nostre, van participar en el projecte d'avaluació d'anotacions genòmiques (GASP[4]) sobre un contig de 3Mbp del braç del cromosoma 2L de *D. melanogaster*, la regió on es troba el gen de la *Adh* i que conté uns 222 gens.

Informació complementària:

- ➢ La pàgina dedicada al pòster que es va presentar al ISMB'99:
  `http://www1.imim.es/software/gfftools/GFF2PS-ADHposter.html`
- ➢ Versió electrònica del pòster, "*Drosophila* Genome Annotation Assessment Project".
  `http://www.genome.org/content/vol10/issue4/images/data/483/DC1/GR10n4_poster.zip`

✯ Anàlisi del genoma de *Drosophila melanogaster* [pàgina 10]

Adams, M. D. *et al.* (including Josep F. Abril) (2000). The Genome Sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195

Aquest article conté l'anàlisi de la primera versió ensamblada del genoma de *D. melanogaster*, que van elaborar conjuntament el consorci públic de seqüenciació del genoma de la mosca, encapçalat pel *Berkeley Drosophila Genome Project* de la universitat de Berkeley a Califòrnia (BDGP[5]), i la empresa nordamericana Celera Genomics, dirigida pel Dr. Craig Venter. Va ser la proba de que el mètode de seqüenciació per fragments a l'atzar (*shotgun sequencing*) era aplicable a genomes d'organismes eucariotes complexos, alhora que era el genoma més gran seqüenciat fins a la data, 120Mbp de seqüència repartits en 4 cromosomes. L'anàlisi computacional d'aquest genoma va revelar l'existència d'uns 13000 gens, que es van representar en la figura central d'aquest article (veieu la figura 2, a la pàgina 11 d'aquest informe).

Informació complementària:

- ➢ Versió electrònica de la figura 4, "Coding content of the fly genome".
  `http://www.sciencemag.org/feature/data/genomes/2000/drosophila.shl`

✯ Anàlisi del genoma humà [pàgina 12]

Venter, J. C. *et al.* (including J. F. Abril and R. Guigó) (2001). The Sequence of the Human Genome. *Science*, 291(5507):1304–1351

El febrer del 2001 es van publicar de manera simultània els dos esborranys de la seqüència del genoma humà, l'un va aparèixer a la revista *Nature* i estava signat pel consorci públic, mentre que l'altre ho feia a la revista *Science* i venia signat per la empresa Celera Genomics. Donat que ja haviem col.laborat amb Celera en l'etapa d'anotació del genoma de *D. melanogaster*, ens van oferir l'oportunitat de poder treballar amb les dades del genoma humà de que disposaven. El genoma humà, de 3Gbp, era 30 vegades més gran que el de la mosca del vinagre, es treballava amb uns 35000 gens i tota la resta d'informació adicional (contingut en G+C, regions repetitives, etc...). Per complicar encara més les cosses, la seqüència del genoma humà tan sols codifica en un 1%, la distribució dels gens al llarg de la seqüència dels cromosomes és molt variable així com la seva densitat (per exemple, el cromosoma 19 és un dels que té més gens per kilobase). Això ens va obligar a treballar amb dues escales, una per representar les regions on els gens mapaven sobre la seqüència dels cromosomes, i un altra, a mena d'expansions que es projectaven a partir d'aquelles regions, per poder visualitzar les estructures exòniques d'aquests gens. Alhora, es va haver de paral.lelitzar l'execució del programa `gff2ps` (corrent el programa amb els mateixos

[4] http://www.fruitfly.org/GASP1/
[5] `http://www.fruitfly.org/`

paràmetres de configuració independenment per cada cromosoma i fussionant els resultats en una única figura) per poder obtenir esborranys de la figura en un temps raonable.

Informació complementària:

➢ Versió electrònica de la figura 1, "Annotation of the Celera Human Genome Assembly".
`http://www.sciencemag.org/cgi/content/full/291/5507/1304/DC2`

✯ Fiabilitat dels programes de predicció computacional de gens en seqüències genòmiques [pàgina 14]

Guigó, R., Agarwal, P., Abril, J. F., Burset, M., and Fickett, J. W. (2000). Gene prediction accuracy in large DNA sequences. *Genome Research*, 10:1631–1642

Donada la mancança de conjunts de dades per seqüències a escala genòmica, era necessari poder avaluar la fiabilitat i la precissió dels programes de predicció computacional de gens quan aquests s'enfrontessin amb aquestes dades. Aquest punt és força important en quant que el primer resultat que en resulta de seqüenciar el genoma humà és el catàleg de gens codificats en la seva seqüència. Per que aquest conjunt de gens tingui una utilitat a nivell biomèdic ha de ser el més acurat i complet que es pugui. Es va treballar amb dos conjunts de dades:
a) un extracte de la base de dades de seqüències EMBL, amb 178 seqüències genòmiques que contenien sencer un únic gen (i del qual es coneixien les coordenades del seu mRNA i dels exons que el composaven). La seva longitut mitjana era de 7Kbp.
b) sobre un fons de seqüència generat a l'atzar per simular els espais intergènics es van combinar conjunts de les seqüències anteriors. La longitut mitjana d'aquests pseudo-BACs era d'uns 180Kbp.
Els resultats suggerien que les eines de predicció computacional de gens a escala genòmica encara no són capaces de recuperar l'estructura exònica exacta de cadascun dels gens codificats en el genoma humà.

Informació complementària:

➢ El conjunt de seqüències de proba està disponible a:
`http://www1.imim.es/databases/gpeval2000/index.html`

PERSPECTIVES DE RECERCA

✯ Predicció computacional de gens basada en l'homologia a nivell de seqüència

L'anàlisi comparativa de genomes permetrà una millora de les predicions computacionals dels gens codificats en els mateixos. Emprant la informació que ens aporta l'homologia a nivell de seqüència entre espècies filogenèticament properes es poden abordar dos problemes que avui en dia encara no s'han resolt i que estan força interrel.lacionats: per una banda, completar el catàleg de gens per a una espècie en concret —en el nostre cas aplicant les tècniques per determinar les regions codificants del genoma humà—, i per l'altra, poder acostar-nos a la determinació del patrò de processat alternatiu dels exons que conformen l'estructura d'un gen (el que s'anomena *alternative splicing*). La nostra aproximació es basa en utilitzar les dades de seqüència del genoma de ratolí (*Mus musculus*), tant dels contigs provinents del procés de seqüenciació per *shotgun* com de les seqüències genòmiques ensamblades dels seus cromosomes —col.laborant estretament amb el consorci públic del genoma del ratolí—, per reanotar el genoma humà.

✯ Anàlisi automatitzada del genoma humà.

Per poder fer front als requeriments que sorgeixen de l'apartat anterior, ens cal desenvolupar una sèrie de programes que s'executin de manera automàtica cada cop que vagin sortint noves versions tant a nivell de dades de seqüència com de les anotacions disponibles. Això ens permetrà adaptar-nos a uns conjunts de dades que van canviant al llarg del temps i, a la vegada, incrementar la fiabilitat dels nostres resultats al minimitzar les errades humanes que poden apareixer en escalar les aplicacions a un problema tan complexe com la reanotació d'un genoma.

✯ Articles en preparació.

La feina que s'està duent a terme actualment es centra en la implementació de diferents programes per fer manipular i analitzar les dades d'origen genòmic, en concret:

- ➢ `gff2aplot`: visualitzant anotacions sobre alineaments de parells de seqüències (*pairwise alignments*). A partir d'un o més fitxers en format GFF[6], que contenen informació tant de l'homologia entre parells de seqüències com de les anotacions per cadascuna de les seqüències, es genera un fitxer en POSTSCRIPT. Els principals avantatges respecte altres aplicacions actualment disponibles per visualitzar alineaments, sòn: que pot ser fàcilment incorporat en els programes que emprem per automatitzar tasques —donat que treballem amb el format GFF per intercanviar dades entre diferents tasques—, que pot incorporar molta informació sobre les anotacions de les seqüències analitzades i per la gran qualitat que ens proporcionen els gràfics vectorials amb format POSTSCRIPT.

- ➢ SGP-2: de *Syntenic Gene Prediction tool*, aquesta eina combina la informació de l'homologia entre seqüències de dues espècies que s'obté amb el programa TBLASTX —que forma part de la aplicació BLAST desenvolupada a Washington University[7]—, amb els algoritmes de predicció computacional de gens que implementa el programa `geneid`[8,9]. La seva potència es basa en millorar la puntuació dels exons predits que solapen regions conservades (els anomenats HSPs o *High-scoring Segment Pairs*), tot i projectant la puntuació d'aquestes per aprofitar la seva informació, afegir aquestes puntuacions a les calculades pel `geneid` per cada exò abans d'agrupar-los dintre d'estructures gèniques.

[6] Veieu la nota a peu de pàgina número 1 de la pàgina 2

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410

[8] Guigó, R., Knudsen, S., Drake, N., and Smith, T. F. (1992). Prediction of gene structure. *Journal of Molecular Biology*, 226:141–157

[9] Parra, G., Blanco, E., and Guigó, R. (2000). Geneid in *Drosophila*. *Genome Research*, 10:511–515

Un cop tinguem versions plenament operatives de tots dos programes, s'enviarà per la seva publicació un article per cadascun d'ells. Tots dos articles estan en aquests moments en fase de preparació; en el cas del SGP-2 l'article corresponent detallarà els algoritmes i els paràmetres amb els que haurem treballat sobre els conjunts de seqüències de proba.

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 16 no. 8 2000*
Pages 743–744

gff2ps: visualizing genomic annotations

Josep F. Abril and Roderic Guigó*

Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF), C/ Dr. Aiguader, 80. 08003—Barcelona, Spain

Received on December 15, 1999; revised on February 18, 2000; accepted on February 24, 2000

Abstract

Summary: *gff2ps is a program for visualizing annotations of genomic sequences. The program takes the annotated features on a genomic sequence in GFF format as input, and produces a visual output in PostScript. While it can be used in a very simple way, it also allows for a great degree of customization through a number of options and/or customization files.*

Availability: *gff2ps is freely available at http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS.html*

Contact: *jabril@imim.es*

Supplementary information: *http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS.html*

As genomic sequences accumulate, visualization tools are becoming essential for the analysis and interpretation of sequence data. Recently, a format has been proposed for specifying genes and other features associated with genomic sequences, the General Feature Format (GFF, proposed by Durbin and Haussler, http://www.sanger.ac.uk/Software/GFF/). In GFF each feature on the genomic sequence is described in a single-line record that essentially specifies the type and position of the feature on the genomic sequence. A grouping field allows to define sets of features within the GFF file. A number of tools have already been developed to deal with GFF files (see also at GFF URL). We have developed a tool, gff2ps, which allows for visualization of GFF files. gff2ps is a program written in GNU awk (http://www.gnu.org/software/gawk/gawk.html) and PostScript, running on UNIX platforms, that generates a PostScript file given a GFF file.

The page description language PostScript is recognized as the current *de facto* industry standard for high-quality printing. PostScript provides both a printer-independent and a computer-system-independent means to describe integrated text and graphics, which can be put out on a variety of printers, plotters and workstation screens. The generation of PostScript output is very common in sequence analysis tools. Notably, we can cite the RSVP package by Searls (1993).

*To whom correspondence should be addressed.

gff2ps plots the features from different sources specified on a GFF file in a number of parallel rows (the so-called tracks here) along the length of the output page(s) (see Figure 1 for examples). Actually these are 'virtual' pages (the so-called blocks here) allowing for several blocks to be included in a single physical page, or for splitting a single block in a number of physical pages. Features can be plotted in a variety of colors and shapes and those grouped together can be visually linked in a number of ways.

gff2ps allows for a substantial amount of customization through command line options, and configuration files. However, meaningful output in most cases, meaningful output can be obtained without the need of any customization, by simply calling gff2ps with the input GFF file. gff2ps assumes, by default, that the GFF file itself carries enough formatting information. The examples in the figure show the versatility of gff2ps. Additional examples can be found at the gff2ps web page, as well as a detailed User Manual.

One of the main advantages of gff2ps is its ability to manage many physical page formats, including user-defined ones. This allows, for instance, the generation of poster size genomic maps. As an example, we used gff2ps to display at the ISMB'99 meeting, the predictions submitted to the Genome Annotation Assessment Project (GASP1) (http://www.fruitfly.org/GASP1/). The GASP1 plot was generated on three B0 size posters from a GFF file of over 50 000 feature records. The program has also been used to obtain the poster figures of recent relevant papers in genomic research (Adams *et al.*, 2000; Reese, 2000).

Acknowledgments

We thank Moisès Burset and Genís Parra (IMIM) for their useful comments, Richard Bruskiewich (Sanger Center) for his helpful hints on the GFF format, also Elena Casacuberta and Amparo Monfort (CSIC) for motivating us to develop this tool. This work is supported by a grant from Plan Nacional de I+D, BIO98-0443-C02-01, and from a fellowship to J.A. from the Instituto de Salud Carlos III, 99/9345.

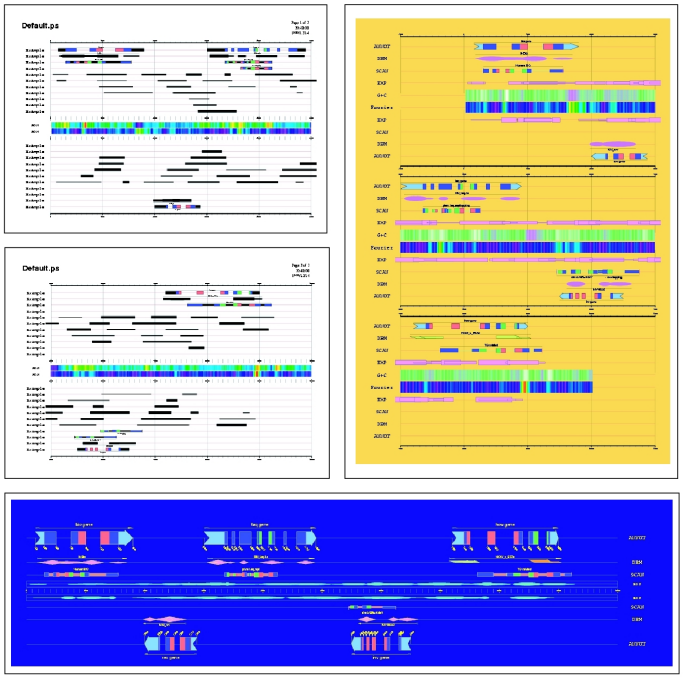

Fig. 1. Different views of the same input GFF file, using gff2ps with different configuration files and command-line options. The top two pages on the left were obtained using the default configuration (specifying only the number of pages, and output media size). By default, gff2ps makes a number of assumptions. Among others: (i) Features grouped from the GFF input file (ungrouped features are treated as a single element group) within the same source are plotted in the minimum number of tracks, guaranteeing that different groups do not overlap. (ii) The plot is fitted into a single block (assuming the length of the sequence to be the end of the most downstream feature), and the block is printed into a single physical page. (iii) Features for which the frame is specified are plotted using a two color code schema. The upstream half of the graphical element representing the frame of the feature and the downstream half the complement modulus 3 of its remainder. This is useful to check frame consistency between adjacent features (for instance, predicted exons). Two adjacent features are frame compatible when the color of the downstream half of the upstream feature matches the color of the upstream half of the downstream feature. (iv) If a score is provided for a feature, the feature is plotted with a height proportional to its score. (v) Obviously, all these default options can be overridden by the user. Notes: The real size color plots, the input GFF files, the configuration files and command line options used in each case, as well as additional examples can be found at: http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS-Snapshots.html.

References

Adams,M.D. and Abril,J.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**(5461), 2185–2195.

Reese,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and Lewis,S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.

Searls,D.B. (1993) Doing sequence analysis with your printer. *Comput. Appl. Biosci.*, **9**(4), 421–426.

Letter

Genome Annotation Assessment in *Drosophila melanogaster*

Martin G. Reese,[1,4] George Hartzell,[1] Nomi L. Harris,[1] Uwe Ohler,[1,2] Josep F. Abril,[3] and Suzanna E. Lewis[1]

[1]*Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3200 USA;* [2]*Chair for Pattern Recognition, University of Erlangen–Nuremberg, D-91058 Erlangen, Germany;*[3]*Institut Municipal d'Investigació Médica—Universitat Pompeu Fabra, Department of Medical Informatics (IMIM—UPF), 08003 Barcelona, Spain*

Computational methods for automated genome annotation are critical to our community's ability to make full use of the large volume of genomic sequence being generated and released. To explore the accuracy of these automated feature prediction tools in the genomes of higher organisms, we evaluated their performance on a large, well-characterized sequence contig from the *Adh* region of *Drosophila melanogaster*. This experiment, known as the Genome Annotation Assessment Project (GASP), was launched in May 1999. Twelve groups, applying state-of-the-art tools, contributed predictions for features including gene structure, protein homologies, promoter sites, and repeat elements. We evaluated these predictions using two standards, one based on previously unreleased high-quality full-length cDNA sequences and a second based on the set of annotations generated as part of an in-depth study of the region by a group of *Drosophila* experts. Although these standard sets only approximate the unknown distribution of features in this region, we believe that when taken in context the results of an evaluation based on them are meaningful. The results were presented as a tutorial at the conference on Intelligent Systems in Molecular Biology (ISMB-99) in August 1999. Over 95% of the coding nucleotides in the region were correctly identified by the majority of the gene finders, and the correct intron/exon structures were predicted for >40% of the genes. Homology-based annotation techniques recognized and associated functions with almost half of the genes in the region; the remainder were only identified by the ab initio techniques. This experiment also presents the first assessment of promoter prediction techniques for a significant number of genes in a large contiguous region. We discovered that the promoter predictors' high false-positive rates make their predictions difficult to use. Integrating gene finding and cDNA/EST alignments with promoter predictions decreases the number of false-positive classifications but discovers less than one-third of the promoters in the region. We believe that by establishing standards for evaluating genomic annotations and by assessing the performance of existing automated genome annotation tools, this experiment establishes a baseline that contributes to the value of ongoing large-scale annotation projects and should guide further research in genome informatics.

Genome annotation is a rapidly evolving field in genomics made possible by the large-scale generation of genomic sequences and driven predominantly by computational tools. The goal of the annotation process is to assign as much information as possible to the raw sequence of complete genomes with an emphasis on the location and structure of the genes. This can be accomplished by ab initio gene finding, by identifying homologies to known genes from other organisms, by the alignment of full-length or partial mRNA sequences to the genomic DNA, or through combinations of such methods. Related techniques can also be used to identify other features, such as the location of regulatory elements or repetitive sequence elements. The ultimate goal of genome annotation, the functional classification of all the identified genes, currently depends on discovering homologies to genes with known functions.

We are interested in an objective assessment of the state of the art in automated tools and techniques for annotating complete genomes. The Genome Annotation Assessment Project (GASP) was organized to formulate guidelines and accuracy standards for evaluating computational tools and to encourage the development of new models and the improvement of existing approaches through a careful assessment and comparison of the predictions made by current state-of-the-art programs.

The GASP experiment, the first of its kind, was similar in many ways to the CASP (Critical Assessment of techniques for protein Structure Prediction) contests for protein structure prediction (Dunbrack et al. 1997; Levitt 1997; Moult et al. 1997, 1999; Sippl et al. 1999; Zemla et al. 1999), described at http://predictioncenter.llnl.gov. However, unlike the CASP contest, GASP was promoted as a collaboration to evaluate various techniques for genome annotation.

[4]Corresponding author.
E-MAIL mgreese@lbl.gov; FAX (510) 486-6798.

10:483–501 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

The GASP experiment consisted of the following stages: (1) Training data for the *Adh* region, including 2.9 Mb of *Drosophila melanogaster* genomic sequence, was collected by the organizers and provided to the participants; (2) a set of standards was developed to evaluate submissions while the participating groups produced and submitted their annotations for the region; and (3) the participating groups' predictions were compared with the standards, a team of independent assessors evaluated the results of the comparison, and the results were presented as a tutorial at ISMB-99(Reese et al. 1999).

Participants were given the finished sequence for the *Adh* region and some related training data, but they did not have access to the full-length cDNA sequences that were sequenced for the paper by Ashburner et al. (1999b) that describes the *Adh* region in depth. The experiment was widely announced and open to any participants. Submitters were allowed to use any available technologies and were encouraged to disclose their methods. Because we were fortunate to attract a large group of participants who provided a wide variety of annotations, we believe that our evaluation addresses the state of art in genome annotation.

Twelve groups participated in GASP, submitting annotations in one or more of six categories: ab initio gene finding, promoter recognition, EST/cDNA alignment, protein similarity, repetitive sequence identification, and gene function. Table 1 lists each participating group, the names of the programs or systems it used, and which of the six classes of annotations it submitted (see enclosed poster in this issue for a graphic overview of all the groups' results). Additional papers in this issue are written by the participants themselves and describe their methods and results in detail.

It should be noted that the lack of a standard that is absolutely correct makes evaluating predictions problematic. The expert annotations described by the *Drosophila* experts (Ashburner et al. 1999b) are our best available resource, but their accuracy will certainly improve as more data becomes available. At best, the data we had in hand is representative of the true situation, and our conclusions would be unchanged by using a more complete data set. At worst, there is a bias in the available data that makes our conclusions significantly misleading. We believe that the data is not unreasonable and that conclusions based on it are correct enough to be valuable as the basis for discussion and future development. We do not believe that the values for the various statistics introduced below are precisely what they would be using the extra information, and we emphasize that they should always be considered in the context of this particular annotated data set [for a further detailed discussion of evaluating these predictions, see Birney and Durbin (2000)].

In the next section we describe the target genomic sequence and the auxiliary data, including a critical discussion of our standard sets. Methods gives a short

Table 1. Participating Groups and Associated Annotation Categories

	Program name	Gene finding	Promoter recognition	EST/c DNA alignment	Protein similarity	Repeat	Gene function
Mural et al. Oakridge, US	GRAIL	X		X			X
Parra et al. Barcelona, ES	GeneID	X					
Krogh Copenhagen, DK	HMMGene	X					
Henikoff et al. Seattle, US	BLOCKS				X		X
Solovyev et al. Sanger, UK	FGenes	X					
Gaasterland et al. Rockefeller, US	MAGPIE	X	X	X		X	X
Benson et al. Mount Sinai, US	TRF					X	
Werner et al. Munich, GER	CoreInspector		X				
Ohler et al. Nuremberg, GER	MCPromoter		X				
Birney Sanger, UK	GeneWise				X		X
Reese et al. Berkeley/Santa Cruz, US	Genie	X	X				

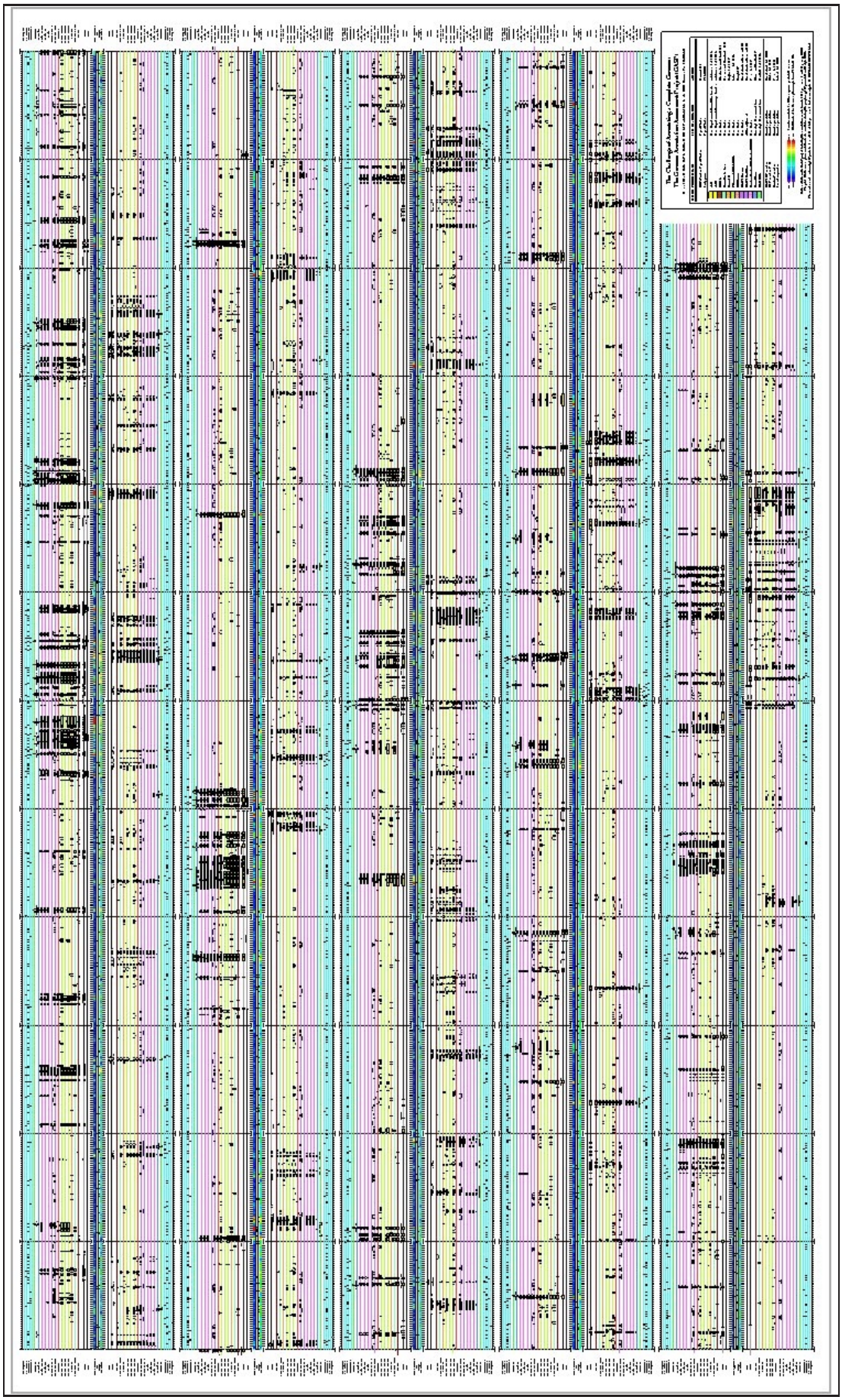

Figura 1: "*Drosophila* Genome Annotation Assessment Project".
La mida real d'aquesta figura és de 100x60cm.

REVIEW

The Genome Sequence of *Drosophila melanogaster*

Mark D. Adams,[1]* Susan E. Celniker,[2] Robert A. Holt,[1] Cheryl A. Evans,[1] Jeannine D. Gocayne,[1] Peter G. Amanatides,[1] Steven E. Scherer,[3] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[2] Reed A. George,[2] Suzanna E. Lewis,[4] Stephen Richards,[2] Michael Ashburner,[5] Scott N. Henderson,[1] Granger G. Sutton,[1] Jennifer R. Wortman,[1] Mark D. Yandell,[1] Qing Zhang,[1] Lin X. Chen,[1] Rhonda C. Brandon,[1] Yu-Hui C. Rogers,[1] Robert G. Blazej,[2] Mark Champe,[2] Barret D. Pfeiffer,[2] Kenneth H. Wan,[2] Clare Doyle,[2] Evan G. Baxter,[2] Gregg Helt,[6] Catherine R. Nelson,[4] George L. Gabor Miklos,[7] Josep F. Abril,[8] Anna Agbayani,[2] Hui-Jin An,[1] Cynthia Andrews-Pfannkoch,[1] Danita Baldwin,[1] Richard M. Ballew,[1] Anand Basu,[1] James Baxendale,[1] Leyla Bayraktaroglu,[9] Ellen M. Beasley,[1] Karen Y. Beeson,[1] P. V. Benos,[10] Benjamin P. Berman,[2] Deepali Bhandari,[1] Slava Bolshakov,[11] Dana Borkova,[12] Michael R. Botchan,[13] John Bouck,[3] Peter Brokstein,[4] Phillipe Brottier,[14] Kenneth C. Burtis,[15] Dana A. Busam,[1] Heather Butler,[16] Edouard Cadieu,[17] Angela Center,[1] Ishwar Chandra,[1] J. Michael Cherry,[18] Simon Cawley,[19] Carl Dahlke,[1] Lionel B. Davenport,[1] Peter Davies,[1] Beatriz de Pablos,[20] Arthur Delcher,[1] Zuoming Deng,[1] Anne Deslattes Mays,[1] Ian Dew,[1] Suzanne M. Dietz,[1] Kristina Dodson,[1] Lisa E. Doup,[1] Michael Downes,[21] Shannon Dugan-Rocha,[3] Boris C. Dunkov,[22] Patrick Dunn,[1] Kenneth J. Durbin,[3] Carlos C. Evangelista,[1] Concepcion Ferraz,[23] Steven Ferriera,[1] Wolfgang Fleischmann,[5] Carl Fosler,[1] Andrei E. Gabrielian,[1] Neha S. Garg,[1] William M. Gelbart,[9] Ken Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[3] Zhiping Gu,[1] Ping Guan,[1] Michael Harris,[1] Nomi L. Harris,[2] Damon Harvey,[4] Thomas J. Heiman,[1] Judith R. Hernandez,[3] Jarrett Houck,[1] Damon Hostin,[1] Kathryn A. Houston,[2] Timothy J. Howland,[1] Ming-Hui Wei,[1] Chinyere Ibegwam,[1] Mena Jalali,[1] Francis Kalush,[1] Gary H. Karpen,[21] Zhaoxi Ke,[1] James A. Kennison,[24] Karen A. Ketchum,[1] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[6] Zhongwu Lai,[1] Paul Lasko,[25] Yiding Lei,[1] Alexander A. Levitsky,[1] Jiayin Li,[1] Zhenya Li,[1] Yong Liang,[1] Xiaoying Lin,[26] Xiangjun Liu,[1] Bettina Mattei,[1] Tina C. McIntosh,[1] Michael P. McLeod,[3] Duncan McPherson,[1] Gennady Merkulov,[1] Natalia V. Milshina,[1] Clark Mobarry,[1] Joe Morris,[6] Ali Moshrefi,[2] Stephen M. Mount,[27] Mee Moy,[1] Brian Murphy,[1] Lee Murphy,[28] Donna M. Muzny,[3] David L. Nelson,[3] David R. Nelson,[29] Keith A. Nelson,[1] Katherine Nixon,[2] Deborah R. Nusskern,[1] Joanne M. Pacleb,[2] Michael Palazzolo,[2] Gjange S. Pittman,[1] Sue Pan,[1] John Pollard,[1] Vinita Puri,[1] Martin G. Reese,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Sidén-Kiamos,[11] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[1] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Aihui H. Wang,[1] Xin Wang,[1] Zhen-Yuan Wang,[1] David A. Wassarman,[33] George M. Weinstock,[3] Jean Weissenbach,[14] Sherita M. Williams,[1] Trevor Woodage,[1] Kim C. Worley,[3] David Wu,[1] Song Yang,[2] Q. Alison Yao,[1] Jane Ye,[1] Ru-Fang Yeh,[19] Jayshree S. Zaveri,[1] Ming Zhan,[1] Guangren Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Xiangqun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[3] Shiaoping Zhu,[1] Xiaohong Zhu,[1] Hamilton O. Smith,[1] Richard A. Gibbs,[3] Eugene W. Myers,[1] Gerald M. Rubin,[34] J. Craig Venter[1]

The fly *Drosophila melanogaster* is one of the most intensively studied organisms in biology and serves as a model system for the investigation of many developmental and cellular processes common to higher eukaryotes, including humans. We have determined the nucleotide sequence of nearly all of the ~120-megabase euchromatic portion of the *Drosophila* genome using a whole-genome shotgun sequencing strategy supported by extensive clone-based sequence and a high-quality bacterial artificial chromosome physical map. Efforts are under way to close the remaining gaps; however, the sequence is of sufficient accuracy and contiguity to be declared substantially complete and to support an initial analysis of genome structure and preliminary gene annotation and interpretation. The genome encodes ~13,600 genes, somewhat fewer than the smaller *Caenorhabditis elegans* genome, but with comparable functional diversity.

The annotated genome sequence of *Drosophila melanogaster*, together with its associated biology, will provide the foundation for a new era of sophisticated functional studies (*1*–*3*). Because of its historical importance, large research community, and powerful research tools, as well as its modest genome size, *Drosophila* was chosen as a test system to explore the applicability of whole-genome shotgun (WGS) sequencing for large and complex eukaryotic genomes (*4*). The groundwork for this project was laid over many years by the fly research community, which has molecularly characterized ~2500 genes; this work in turn has been supported by nearly a century of genetics (*5*). Since *Drosophila* was chosen in 1990 as one of the model organisms to be studied under the auspices of the federally funded Human Genome Project, genome projects in the United States, Europe, and Canada have produced a battery of genome-wide resources (Table 1). The Berkeley and European *Drosophila* Genome Projects (BDGP and EDGP) initiated genomic sequencing (Tables 1 to 3) and finished 29 Mb. The bacterial artificial chromosome (BAC) map and other genomic resources available for *Drosophila* serve both as an independent confirmation of the assembly of data from the shotgun strategy and as a set of resources for further biological analysis of the genome.

The *Drosophila* genome is ~180 Mb in size, a third of which is centric heterochromatin (Fig. 1). The 120 Mb of euchromatin is on two large autosomes and the X chromosome; the small fourth chromosome contains only ~1 Mb of euchromatin. The heterochromatin consists mainly of short, simple sequence elements repeated for many megabases, occasionally interrupted by inserted transposable elements, and tandem arrays of ribosomal RNA genes. It is known that there are small islands of unique sequence embedded within heterochromatin—for example, the mitogen-activated protein kinase gene *rolled* on chromosome 2, which is flanked on each side by at least 3 Mb of heterochromatin. Unlike the *C. elegans* genome, which can be completely cloned in yeast artificial chromosomes (YACs), the simple sequence repeats are not stable in YACs (*6*) or other large-insert cloning systems. This has led to a functional definition of the euchromatic genome as that portion of the genome that can be cloned stably in BACs. The euchromatic portion of the genome is the subject of both the federally funded *Drosophila* sequencing project and the work presented here. We began WGS sequencing of *Drosophila* less than 1 year ago, with two major goals: (i) to test the strategy on a large and complex eukaryotic genome as a prelude to sequencing the human genome, and (ii) to provide a complete, high-quality genomic sequence to the *Drosophila* research community so as to advance research in this important model organism.

WGS sequencing is an effective and efficient way to sequence the genomes of prokaryotes, which are generally between 0.5 and 6 Mb in size (*7*). In this strategy, all the DNA of an organism is sheared into segments a few thousand base pairs (bp) in length and cloned directly into a plasmid vector suitable for DNA sequencing. Sufficient DNA sequencing is performed so that each base pair is covered numerous times, in fragments of ~500 bp. After sequencing, the fragments are assembled in overlapping segments to reconstruct the complete genome sequence.

In addition to their much larger size, eukaryotic genomes often contain substantial amounts of repetitive sequence that have the potential to interfere with correct sequence assembly. Weber and Myers (*8*) presented a theoretical analysis of WGS sequencing in which they examined the impact of repetitive sequences, discussed experimental strategies to mitigate their effect on sequence assembly, and suggested that the WGS method could be applied effectively to large eukaryotic genomes. A key component of the strategy is obtaining sequence data from each end of the cloned DNA inserts; the juxtaposition of these end-sequences ("mate pairs") is a critical element in producing a correct assembly.

Genomic Structure

WGS libraries were prepared with three different insert sizes of cloned DNA: 2 kb, 10 kb, and 130 kb. The 10-kb clones are large enough to span the most common repetitive sequence elements in *Drosophila*, the retrotransposons. End-sequence from the BACs provided long-range linking information that was used to confirm the overall structure of the assembly (*9*). More than 3 million sequence reads were obtained from whole-genome libraries (Fig. 2 and Table 2). Only ~2% of the sequence reads contained heterochromatic simple sequence repeats, indicating that the heterochromatic DNA is not stably cloned in the small-insert vectors used for the WGS libraries. A BAC-based physical map spanning >95% of the euchromatic portion of the genome was constructed by screening a BAC library with sequence-tagged site (STS) markers (*10*). More than 29 Mb of high-quality finished sequence has been completed from BAC, P1, and cosmid clones, and draft sequence data (~1.5× average coverage) were obtained from an additional 825 BAC and P1 clones spanning in total >90% of the genome (Table 3). The clone-based draft sequence served two purposes: It improved the likelihood of accurate assembly, and it allowed the identification of templates and primers for filling gaps that remain after assembly. An initial assembly was performed using the WGS data and BAC end-sequence [WGS-only assembly (*4*)]; subsequent assemblies included the clone-based draft sequence data (joint assembly). Figure 3 and Table 3 illustrate the status of the euchromatic sequence resulting from each of these assemblies and the current status following the directed gap closure completed to date. The sequence assembly process is described in detail in an accompanying paper (*11*).

Assembly resulted in a set of "scaffolds." Each scaffold is a set of contiguous sequences (contigs), ordered and oriented with respect to one another by mate-pairs such that the gaps between adjacent contigs are of known size and are spanned by clones with end-sequences flanking the gap. Gaps within scaffolds are called sequence gaps; gaps between scaffolds are called "physical gaps" because there are no clones identified spanning the gap. Two methods were used to map the scaffolds to chromosomes: (i) cross-referencing between STS markers present in the assembled sequence and the BAC-based STS content map, and (ii) cross-referencing between assembled sequence and shotgun sequence data obtained from individual tiling-path clones selected from the BAC physical map. The mapped scaffolds from the joint assembly, totaling 116.2 Mb after initial

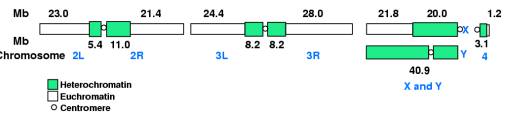

Fig. 1. Mitotic chromosomes of *D. melanogaster*, showing euchromatic regions, heterochromatic regions, and centromeres. Arms of the autosomes are designated 2L, 2R, 3L, 3R, and 4. The euchromatic length in megabases is derived from the sequence analysis. The heterochromatic lengths are estimated from direct measurements of mitotic chromosome lengths (*67*). The heterochromatic block of the X chromosome is polymorphic among stocks and varies from one-third to one-half of the length of the mitotic chromosome. The Y chromosome is nearly entirely heterochromatic.

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]Berkeley *Drosophila* Genome Project (BDGP), Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [4]BDGP, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [5]European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [6]Neomorphic Inc., 2612 Eighth Street, Berkeley, CA 94710, USA. [7]GenetixXpress Pty. Ltd., 78 Pacific Road, Palm Beach, Sydney, NSW 2108, Australia. [8]Department of Medical Informatics, IMIM–UPF C/Dr. Aiguader 80, 08003 Barcelona, Spain. [9]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. [10]Department of Genetics, Box 8232, Washington University Medical School, 4566 Scott Avenue, St. Louis, MO 63110, USA. [11]Institute of Molecular Biology and Biotechnology, Forth, Heraklion, Greece. [12]European *Drosophila* Genome Project (EDGP), EMBL, Heidelberg, Germany. [13]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94710, USA. [14]Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France. [15]Section of Molecular and Cellular Biology, University of California, Davis, CA 95618, USA. [16]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK. [17]EDGP, Rennes University Medical School, UPR 41 CNRS Recombinaisons Genetiques, Faculte de Medicine, 2 av. du Pr. Leon Bernard, 35043 Rennes Cedex, France. [18]Department of Genetics, Stanford University, Palo Alto, CA 94305, USA. [19]Department of Statistics, University of California, Berkeley, CA 94720, USA. [20]EDGP, Centro de Biología Molecular Severo Ochoa, CSIC, Universidad Autónoma de Madrid, 28049 Madrid, Spain. [21]MBVL, Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA. [22]Department of Biochemistry and Center for Insect Science, University of Arizona, Tucson, AZ 85721, USA. [23]EDGP, Montpellier University Medical School, Institut de Genetique Humaine, CNRS (CRBM), 114 rue de la Cardonille, 34396 Montpellier Cedex 5, France. [24]Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health (NIH), Bethesda, MD 20892, USA. [25]Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montreal, Quebec, Canada. [26]The Institute for Genomic Research, Rockville, MD 20850, USA. [27]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA. [28]EDGP, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [29]Department of Biochemistry, University of Tennessee, Memphis, TN 38163, USA. [30]EDGP, Department of Anatomy and Physiology, University of Dundee, Dundee DD1 4HN, UK, and Department of Biological Sciences, Open University, Milton Keynes MK7 6AA, UK. [31]HHMI/Embryology, Carnegie Institution of Washington, Baltimore, MD 21210, USA. [32]Motorola BioChip Systems, Tempe, AZ 85284, USA. [33]Cell Biology and Metabolism Branch, National Institute of Child Health and Human Development, NIH, Bethesda, MD 20892, USA. [34]Howard Hughes Medical Institute, BDGP, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed.

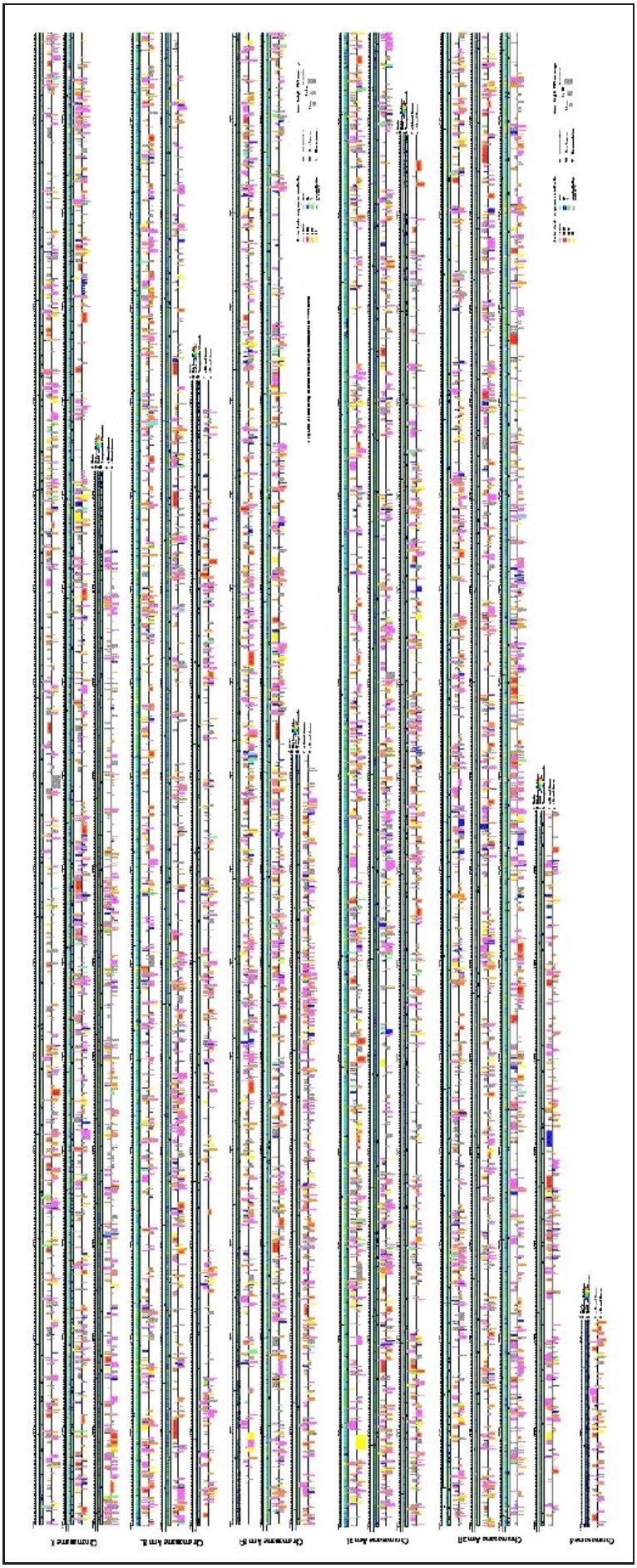

Figura 2: “Coding Content of the *Drosophila* Genome”.
La mida real d’aquesta figura és de dues pàgines de 108x28cm.

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

THE HUMAN GENOME

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

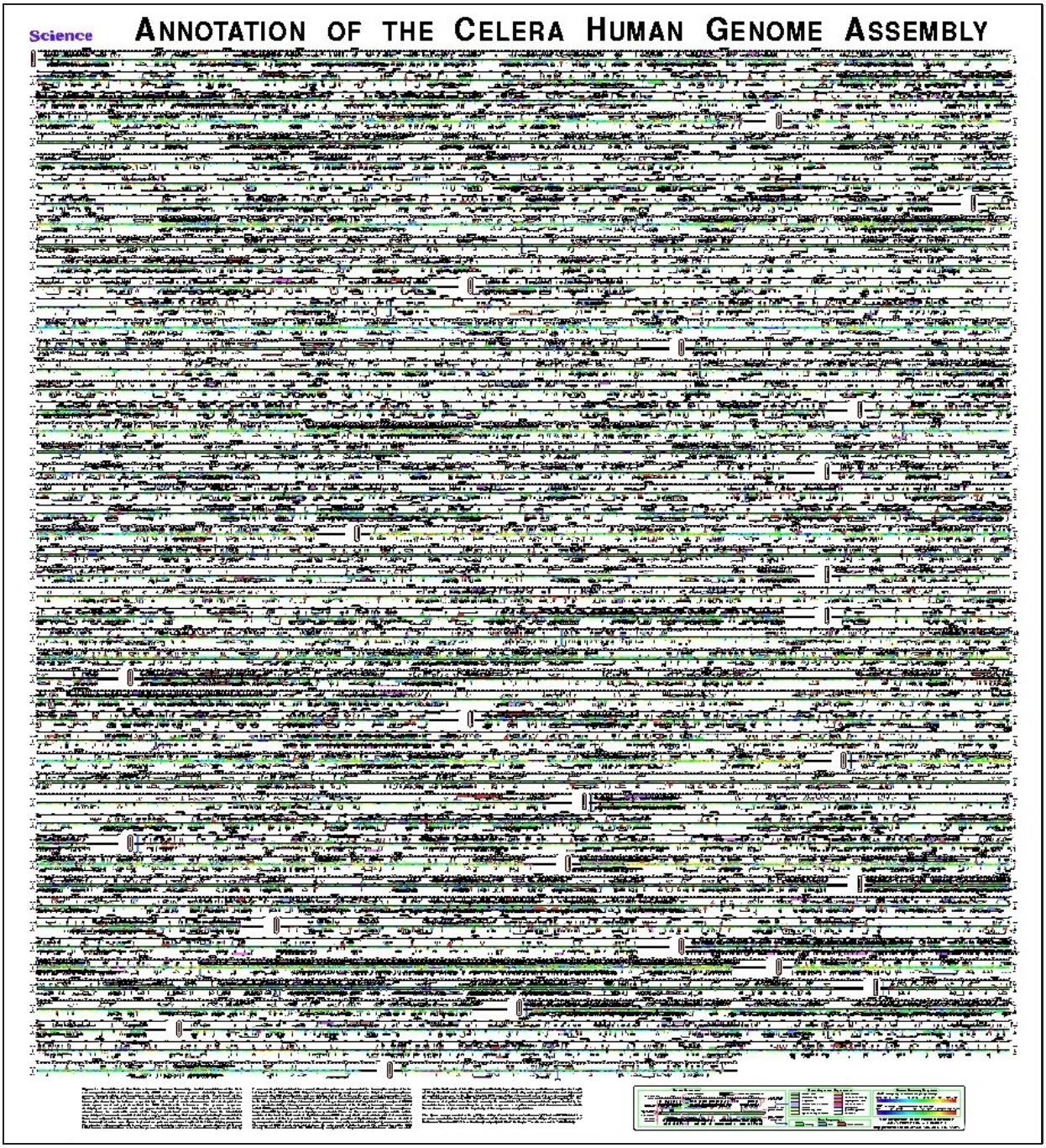

Figura 3: “Annotation of the Celera Human Genome Assembly”.
La mida real d’aquesta figura és de 106x150cm.

An Assessment of Gene Prediction Accuracy in Large DNA Sequences

Roderic Guigó,[1,3] Pankaj Agarwal,[2] Josep F. Abril,[1] Moisés Burset,[1] and James W. Fickett[2]

[1]*Grup de Recerca en Informática Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, E-08003 Barcelona, Spain;* [2]*Department of Bioinformatics, SmithKline Beecham Pharmaceuticals Research and Development, King of Prussia, Pennsylvania 19406, USA*

One of the first useful products from the human genome will be a set of predicted genes. Besides its intrinsic scientific interest, the accuracy and completeness of this data set is of considerable importance for human health and medicine. Though progress has been made on computational gene identification in terms of both methods and accuracy evaluation measures, most of the sequence sets in which the programs are tested are short genomic sequences, and there is concern that these accuracy measures may not extrapolate well to larger, more challenging data sets. Given the absence of experimentally verified large genomic data sets, we constructed a semiartificial test set comprising a number of short single-gene genomic sequences with randomly generated intergenic regions. This test set, which should still present an easier problem than real human genomic sequence, mimics the ~200kb long BACs being sequenced. In our experiments with these longer genomic sequences, the accuracy of GENSCAN, one of the most accurate ab initio gene prediction programs, dropped significantly, although its sensitivity remained high. Conversely, the accuracy of similarity-based programs, such as GENEWISE, PROCRUSTES, and BLASTX, was not affected significantly by the presence of random intergenic sequence, but depended on the strength of the similarity to the protein homolog. As expected, the accuracy dropped if the models were built using more distant homologs, and we were able to quantitatively estimate this decline. However, the specificities of these techniques are still rather good even when the similarity is weak, which is a desirable characteristic for driving expensive follow-up experiments. Our experiments suggest that though gene prediction will improve with every new protein that is discovered and through improvements in the current set of tools, we still have a long way to go before we can decipher the precise exonic structure of every gene in the human genome using purely computational methodology.

The nucleotide genomic sequence is the primary product of the Human Genome Project, but a major short- and mid-term interest will be the amino acid sequences of the proteins encoded in the genome. Thus, methods that reliably predict the genes encoded in genomic sequence are essential, and computational gene identification continues to be an active field of research (for reviews, see Fickett 1996; Claverie 1997; Guigó 1997a; Burge and Karlin 1998; Haussler 1998). A new generation of gene prediction programs based on Hidden Markov Models (Burge and Karlin 1997) have shown significantly greater accuracy than previous programs based on other methodologies (Burset and Guigó 1996). Conversely, as the databases of known coding sequences increase in size, gene prediction methods based on sequence similarity to coding sequences, mainly proteins and ESTs, are becoming increasingly useful and are routinely used to identify putative genes in genomic sequences (The *C. elegans* Sequencing Consortium 1998). We have recently published an evaluation of sequence similarity-based gene prediction methods, in particular of EST-based gene prediction (Guigó et al. 2000). The accuracy of gene identification programs, however, has usually been estimated on controlled data sets made of short genomic sequences encoding a single and complete gene with a simple structure. Moreover, these data sets are often similar if not overlapping, to the sets of sequences on which the programs have been trained. Thus, these data sets are not representative of the sequences being produced at the genome centers, which are mostly large sequences of low coding density, encoding several genes or incomplete genes with complex gene structure. It is thus difficult to know how well the figures of accuracy estimated in the controlled benchmark data sets extrapolate to actual genomic sequences. Furthermore, programs that combine both sequence similarity and ab initio gene finding approaches, and those that predict genes by producing a splicing alignment between a genomic sequence and a candidate amino acid sequence have become recently available, such as PROCRUSTES (Gelfand et al. 1996) and GENEWISE (Birney and Durbin 1997), (http://www.sanger.ac.uk/Software/Wise2/). Programs that align genomic sequences with EST sequences, such as EST_GENOME (Mott 1997), could also be included in this category. These programs promise highly accurate predictions, but at the cost of greater computational time. However, this increase in accuracy has not been well-quantified on challenging data sets. The effects of the degree of similarity between the candidate homolog and the genomic sequence also deserve careful evaluation.

We believe a more realistic evaluation of the currently available gene prediction tools on challenging data sets would be useful. Ideally, one would like to benchmark the computational gene identification programs in real genomic sequences. The main problem is that most real sequences the structure of the genes has not been verified exhaustively by experimental means, and thus it is impossible to calibrate the accuracy of the predictions. Only recently, extensively annotated large genomic sequences from higher eukaryotic organisms have become available from the human genome (http://www.hgmp.mrc.ac.uk/Genesafe) and from the fly genome (http://www.fruitfly.org/GASP1/). In spite of the experimental analysis, the possibility of undetected genes in the sequence cannot be easily ruled out, which makes accuracy difficult to measure. Here, we attempt to overcome the lack of well-annotated large genomic sequences by constructing semiartificial ones. In these semiartificial sequences, known genomic sequences have been embedded in simulated intergenic DNA, and therefore, the location of all coding exons is known. Although the approach may seem unrealistic, we believe that the results obtained are instructive with regard to the accuracy of currently available gene identification tools.

We evaluate the accuracy of representatives of a wide variety of computational gene identification approaches: GENSCAN (Burge and Karlin 1997), an ab initio genefinder; BLASTX (Altschul et al. 1990; Gish and States 1993), a genefinding-oriented similarity search program; and PROCRUSTES (Gelfand et al. 1996) and GENEWISE (Birney and Durbin 1997), genefinders based on aligning a genomic DNA sequence fragment to a homologous protein sequence. We evaluate these programs on two benchmark data sets: A set of well-annotated single-gene DNA sequences, and a set of semiartificial genomic (SAG) sequences created by embedding the single-gene sequences from the first data set in simulated intergenic DNA.

RESULTS

We investigated the accuracy of the gene prediction tools (GENSCAN, PROCRUSTES, GENEWISE, BLASTX) described in Methods on two benchmark sets. In all cases, sequences were masked previously for repeated regions using REPEATMASKER (A. Smit and P. Green, unpubl.). The gene predictions obtained using the different tools were compared with the actual gene annotations using the accuracy measures described Methods.

Accuracy in Single Gene Sequences

Table 1 shows the accuracy of the different gene prediction tools on h178, the set of single gene sequences.

GENSCAN's accuracy is comparable to that reported earlier (Burge and Karlin 1997). On average, 90% of the coding nucleotides and 70% of the exons are predicted correctly by GENSCAN. Only 7% of the actual exons are missed completely, and only 9% of the predicted exons are wrong. We believe this is close to the maximum accuracy that can be achieved using currently available ab initio gene prediction programs.

The quality of the gene models inferred from BLASTX searches depends on the strategy used. Default usage of BLASTX produced poorer predictions than more sophisticated strategies. (Results for BLASTX default correspond to those published in Guigó et al. 2000.) Discrepancies between numbers in Table 1 and those reported in Guiqoet al. (2000) are due to the differences in the way the accuracy measures are summarized. In Guigó et al. 2000, we computed the accuracy measures on each test sequence, and averaged all of them. Here, we compute the accuracy measures globally from the total number of prediction successes and failures (at the base or exon level) on all sequences. The default BLASTX strategy produces reasonably high sensitivity (0.91) by projecting all HSPs over a given threshold along the query DNA sequence, but the sensitivity rises to an amazing 0.97, if the topcomboN fea-

Table 1. Accuracy of Gene Prediction Tools in the Set of Single Gene Sequences (h178)

		Nucleotide			Exon				
Program	No.	Sn	Sp	CC	Sn	Sp	$\frac{Sn+Sp}{2}$	ME	WE
GenScan	177	0.93	0.90	0.90	0.78	0.75	0.76	0.08	0.10
Blastx default	175	0.91	0.79	0.82	0.04	0.04	0.04	0.12	0.05
Blastx topcomboN	174	0.97	0.80	0.86	0.04	0.04	0.04	0.08	0.05
Blastx 2 stages	175	0.90	0.92	0.90	0.10	0.12	0.11	0.19	0.02
GeneWise	177	0.98	0.98	0.97	0.88	0.91	0.89	0.06	0.02
Procrustes	177	0.93	0.95	0.93	0.76	0.82	0.79	0.11	0.04

[3]**Correspondlng author.**
E-MAIL rguigo@lmlm.es; FAX 3493-221-3237.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.122800.