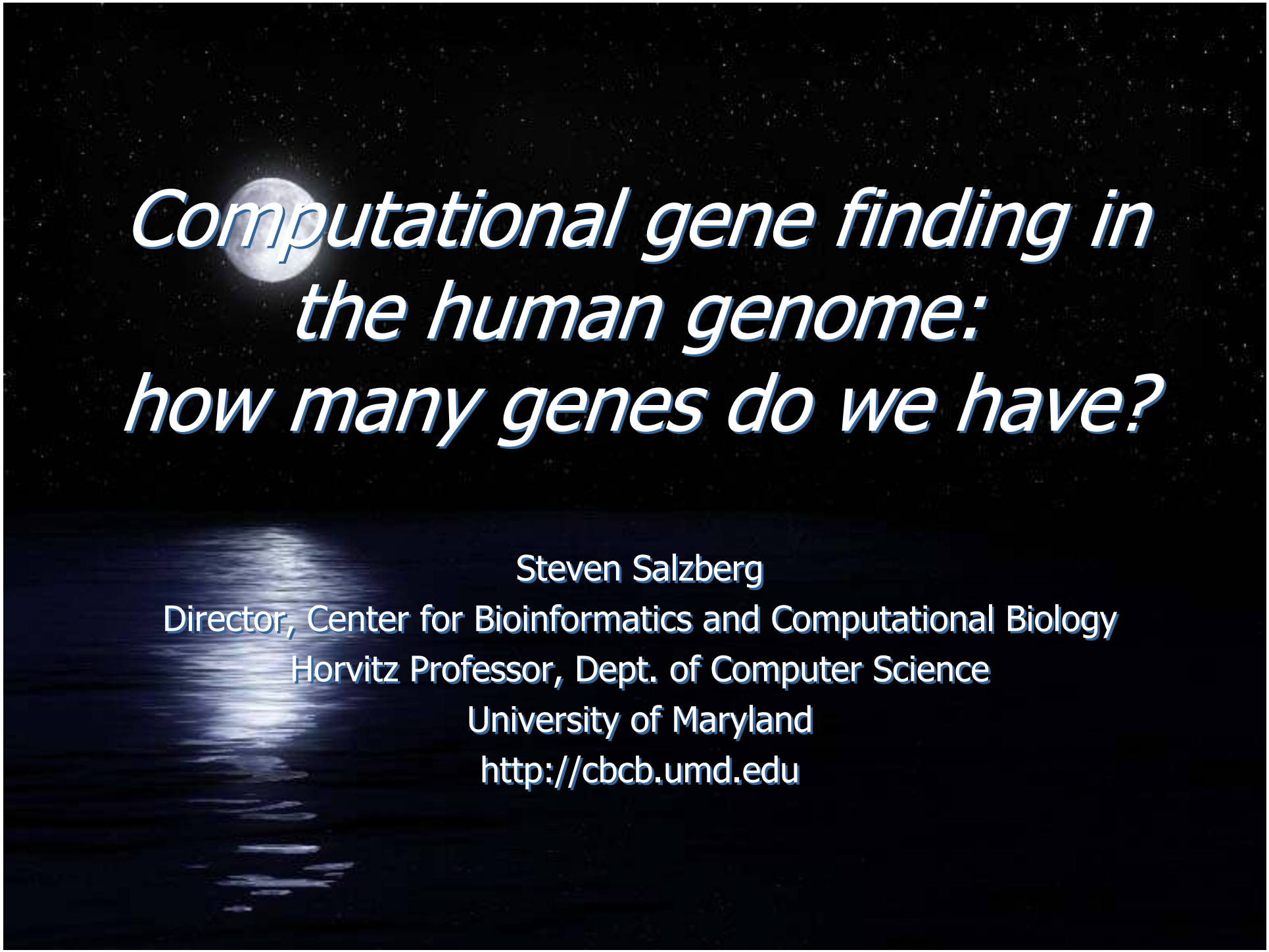

Computational gene finding in the human genome: how many genes do we have?
Steven Salzberg
Director, Center for Bioinformatics and Computational Biology
Horvitz Professor, Dept. of Computer Science
University of Maryland
http://cbcb.umd.edu

We've been trying for a long time to determine how many genes we have

letters to nature

Nature **201**, 847 (22 February 1964); doi:10.1038/201847a0

A Preliminary Estimate of the Number of Human Genes

F. VOGEL

Institut für Anthropologie und Humangenetik, University of Heidelberg, Germany.

RECENT results of molecular genetics enable us to estimate the number of human genes, if certain assumptions are made. The following data are available: (1) The α-chain of human hæmoglobin contains 141, the β-chain contains 146 amino-acids, corresponding to a molecular weight of about 17,000 each[1]. Assuming a triplet code[2,3] this means that the α- and β-chains are determined by 423 and 438 nucleotide pairs, respectively. According to 'Svedberg's law'4, many proteins consist of sub-units of the same order of magnitude (molecular weight of about 17,500). Hence, the assumption seems to be warranted that one average structural geno might have a length of about 450 nucleotide pairs. (2) The weight of one haploid human chromosome set in human spermatozoa is about 2.72×10^{-12} g. Granulocytes contain about 6.23×10^{-12} g; lymphocytes contain about $5.84 \times ^{-12}$ g (ref. 5). Extensive examinations have shown that the DKA. content is constant in all resting cells of one species, which have the same number of chromosome sets, and depends on the degree of polyploidy[5,6]. The assumption seems to be justified that most of the DNA works as genetic material, even if in some cells minor fractions with other functions might possibly be present[7]. In the following calculations the total amount of DNA in a haploid human chromosome set is estimated to be about 3×10^{-12} g. (3) Usually the genetic variants of human haemoglobins differ in one amino-acid substitution only[1,8]. One structural gene can only produce one single type of genetically determined polypeptide chain. As much as we know, this applies for other genetically determined proteins as well. This means that the genetic information for these structural genes can only be present once. Any degree of polyteny for these loci in the germ cells is highly unlikely. As has been mentioned, however, the DNA content of diploid cells is about twice the content of (haploid) spermatozoa. We assume that the total genetic information is only present once.

1,000,000 genes? 100,000 genes?

Science 25 October 1996:
Vol. 274. no. 5287, pp. 540 – 546
DOI: 10.1126/science.274.5287.540

‹ Prev | Table of Contents | Next ›

ARTICLES

A Gene Map of the Human Genome

G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. R. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W.-L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. A. R. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. R. Hudson Jr., S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James, D. Bentley, P. Deloukas, E. S. Lander, T. J. Hudson

The human genome is thought to harbor 50,000 to 100,000 genes, of which about half have been sampled to date in the form of expressed sequence tags. An

CORRESPONDENCE

Nature Genetics **8**, 114 (1994)
doi:10.1038/ng1094-114a

Predicting the total number of human genes

Francisco Antequera[1] & Adrian Bird[1]

[1]Institute of Cell and Molecular Biology, Darwin Building, University of Edinburgh, Scotland EH9 3JR, UK

NEWS AND VIEWS

Nature Genetics **7**, 345 - 346 (1994)
doi:10.1038/ng0794-345

How many genes in the human genome?

Chris Fields[1], Mark D. Adams[1], Owen White[1] & J. Craig Venter[1]

[1]The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, Maryland 20878, USA

Proceedings of the National Academy of Sciences, Vol 90, 11995-11999, Copyright © 1993 by National Academy of Sciences

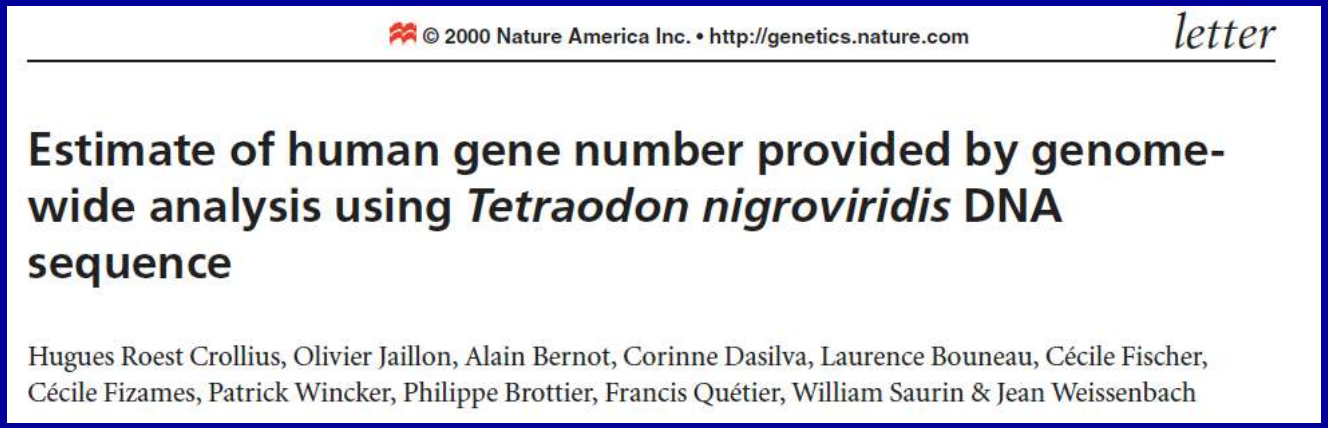
ARTICLE

Number of CpG Islands and Genes in Human and Mouse

F Antequera and A Bird

Estimation of gene number in mammals is difficult due to the high proportion of noncoding DNA within the nucleus. In this study, we provide a direct measurement of the number of genes in human and mouse. We have taken advantage of the fact that many mammalian genes are associated with CpG islands whose distinctive properties allow their physical separation from bulk DNA. Our results suggest that there are ≈45,000 CpG islands per haploid genome in humans and 37,000 in the mouse. Sequence comparison confirms that about 20% of the human CpG islands are absent from the homologous mouse genes. Analysis of a selection of genes suggests that both human and mouse are losing CpG islands over evolutionary time due to de novo methylation in the germ line followed by CpG loss through mutation. This process appears to be more rapid in rodents. Combining the number of CpG islands with the proportion of island-associated genes, we estimate that the total number of genes per haploid genome is ≈80,000 in both organisms.

Estimates in *Nature Genetics* (2000)



letter

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

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

- 28,000 - 34,000
- Based on alignments to pufferfish (*Tetraodon nigroviridis*)

Estimates in *Nature Genetics* (2000)

letter

Analysis of expressed sequence tags indicates 35,000 human genes

Brent Ewing & Phil Green

- Based on expressed sequence tag (EST) alignments to human chromosome 22

Estimates in *Nature Genetics* (2000)

letter

Gene Index analysis of the human genome estimates approximately 120,000 genes

Feng Liang, Ingeborg Holt, Geo Pertea, Svetlana Karamycheva, Steven L. Salzberg & John Quackenbush

- Based on assemblies of ESTs
- Extrapolated to whole human genome

Estimates in *Nature Genetics* (2000)

corrections

Gene Index analysis estimates the human genome contains 120,000 genes

F. Liang *et al.*

Nature Genet. 25, 239–240 (2000).

- Correction to avoid over-counting immunoglobulin ESTs
- 81,000 genes, based on assemblies of ESTs
- 57,000 genes, extrapolating from Chr 21 and 22

The gene count guessing game

Science 19 May 2000:
Vol. 288. no. 5469, pp. 1146 - 1147
DOI: 10.1126/science.288.5469.1146

NEWS OF THE WEEK

HUMAN GENOME PROJECT:
And the Gene Number Is ...?

Elizabeth Pennisi

COLD SPRING HARBOR, NEW YORK--Even though a draft sequence of the human genome is nearing completion, biologists still don't know how many genes it contains. Indeed, the range of estimates seems to be growing rather than shrinking. The question lies at the core of our understanding of genetic complexity. If genomes are the books of life, then genes are the words that tell the story of each organism. Biologists have long assumed that microorganisms are short stories and complex organisms such as humans, great tomes.

Place your bet. Uncertainty over the number of human genes has sparked a debate--and a betting pool.

Human genome paper I: *Nature* 409(15 Feb 2001), 860-921

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

- 30,000 - 40,000 genes
- Large degree of uncertainty about total
- Number of distinct transcripts and proteins even less certain

Human genome paper II: *Science* 291(16 Feb 2001), 1304-51

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-Rui 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]

- 26,588 genes
- 12,000 additional "likely" genes based on similarity to mouse or other evidence

Steven Salzberg,[12]

Human genome version 2.0: *Nature* 431 (21 October 2004)

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

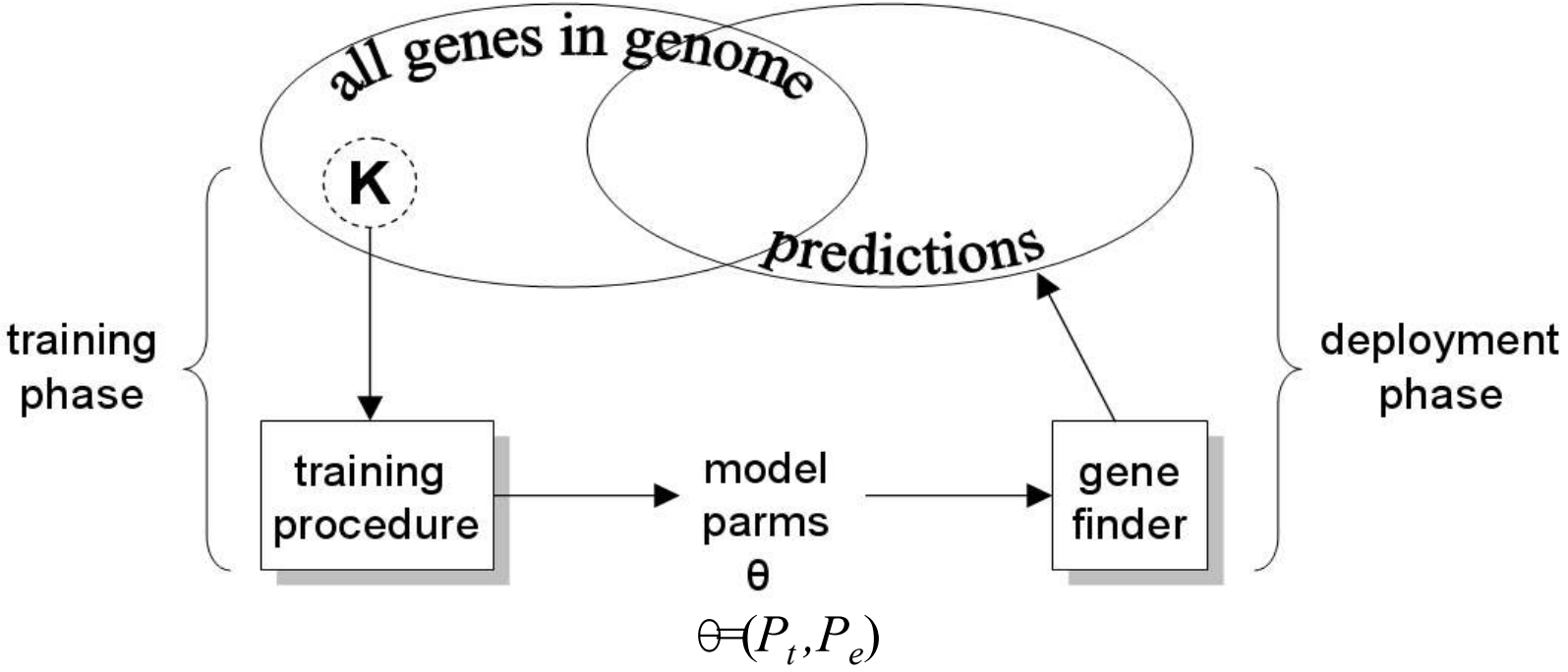
The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

- 20,000 - 25,000 genes

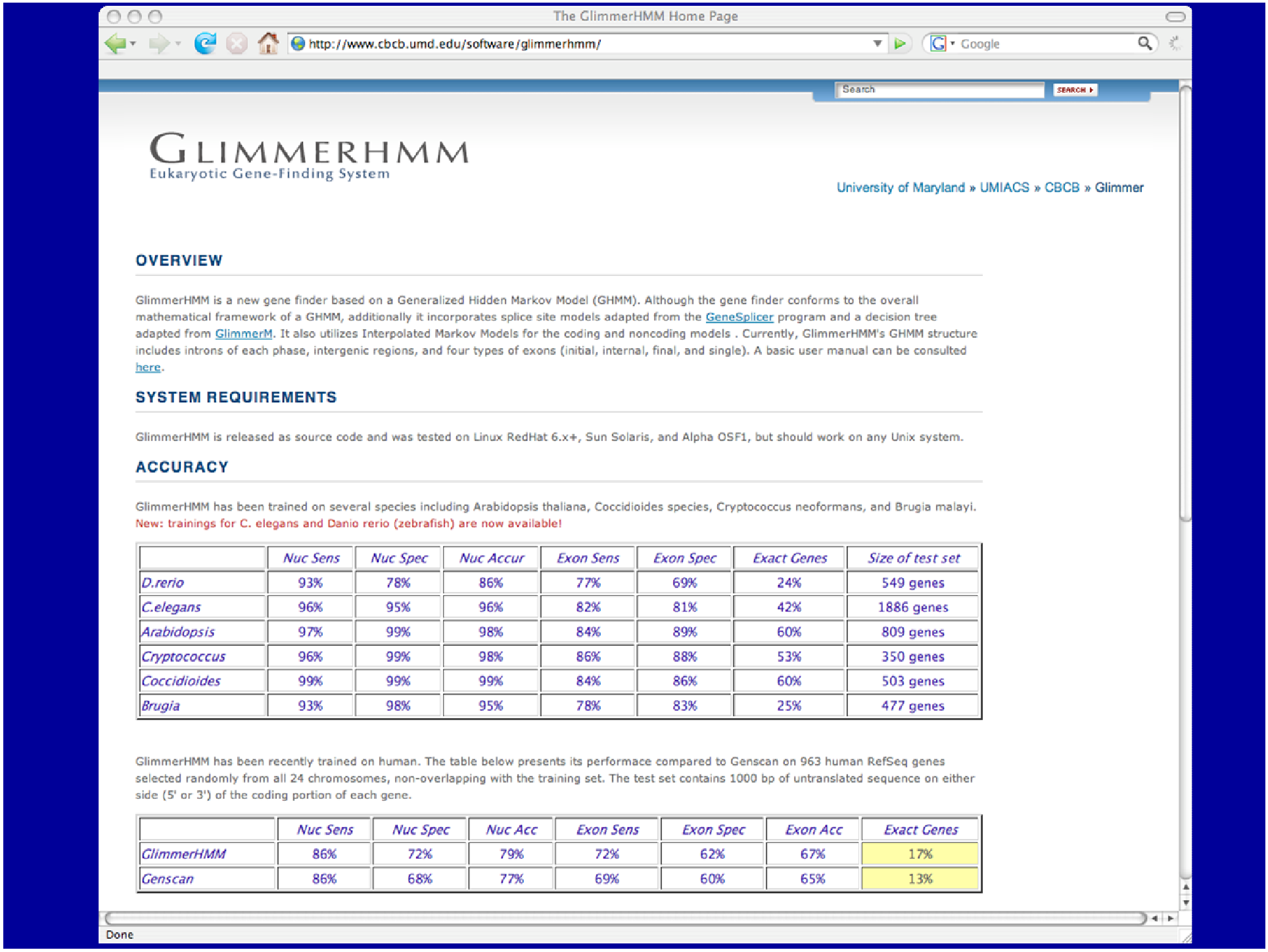
How do we find genes?

- *Ab initio* gene finding
- Expressed sequence tags (ESTs)
- Full-length cDNA sequencing
- Alignment of protein sequences to genomic DNA
- Combining all the evidence together

Training a Gene Finder



$$\theta_{\max} = \underset{\theta}{\operatorname{argmax}} \sum_{(S,\phi)\in K} P(\phi \mid S,\theta) = \underset{\theta}{\operatorname{argmax}} \frac{\sum_{(S,\phi)\in K} P_t(q_0,q_L)\prod_{i=1}^{L} P_e(x_i \mid q_i)P_t(q_i \mid q_{i-1})}{P(S \mid \theta)}$$


The GlimmerHMM Home Page
http://www.cbcb.umd.edu/software/glimmerhmm/
Google
Search
SEARCH
GLIMMERHMM
Eukaryotic Gene-Finding System
University of Maryland » UMIACS » CBCB » Glimmer
OVERVIEW
GlimmerHMM is a new gene finder based on a Generalized Hidden Markov Model (GHMM). Although the gene finder conforms to the overall mathematical framework of a GHMM, additionally it incorporates splice site models adapted from the GeneSplicer program and a decision tree adapted from GlimmerM. It also utilizes Interpolated Markov Models for the coding and noncoding models . Currently, GlimmerHMM's GHMM structure includes introns of each phase, intergenic regions, and four types of exons (initial, internal, final, and single). A basic user manual can be consulted here.
SYSTEM REQUIREMENTS
GlimmerHMM is released as source code and was tested on Linux RedHat 6.x+, Sun Solaris, and Alpha OSF1, but should work on any Unix system.
ACCURACY
GlimmerHMM has been trained on several species including Arabidopsis thaliana, Coccidioides species, Cryptococcus neoformans, and Brugia malayi.
New: trainings for C. elegans and Danio rerio (zebrafish) are now available!
Nuc Sens Nuc Spec Nuc Accur Exon Sens Exon Spec Exact Genes Size of test set
D.rerio 93% 78% 86% 77% 69% 24% 549 genes
C.elegans 96% 95% 96% 82% 81% 42% 1886 genes
Arabidopsis 97% 99% 98% 84% 89% 60% 809 genes
Cryptococcus 96% 99% 98% 86% 88% 53% 350 genes
Coccidioides 99% 99% 99% 84% 86% 60% 503 genes
Brugia 93% 98% 95% 78% 83% 25% 477 genes
GlimmerHMM has been recently trained on human. The table below presents its performace compared to Genscan on 963 human RefSeq genes selected randomly from all 24 chromosomes, non-overlapping with the training set. The test set contains 1000 bp of untranslated sequence on either side (5' or 3') of the coding portion of each gene.
Nuc Sens Nuc Spec Nuc Acc Exon Sens Exon Spec Exon Acc Exact Genes
GlimmerHMM 86% 72% 79% 72% 62% 67% 17%
Genscan 86% 68% 77% 69% 60% 65% 13%
Done

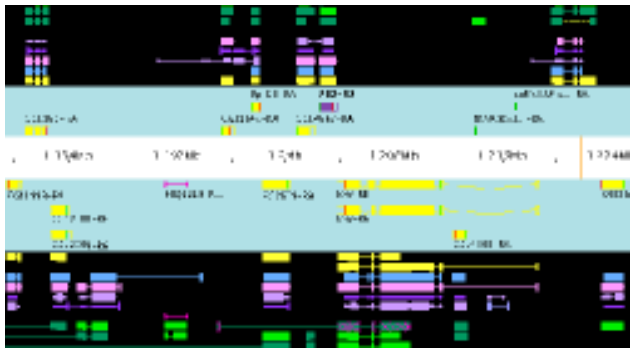
How do we find genes?

- *Ab initio* gene finding
- Expressed sequence tags (ESTs)
- Full-length cDNA sequencing
- Alignment of protein sequences to genomic DNA
- Combining all the evidence together

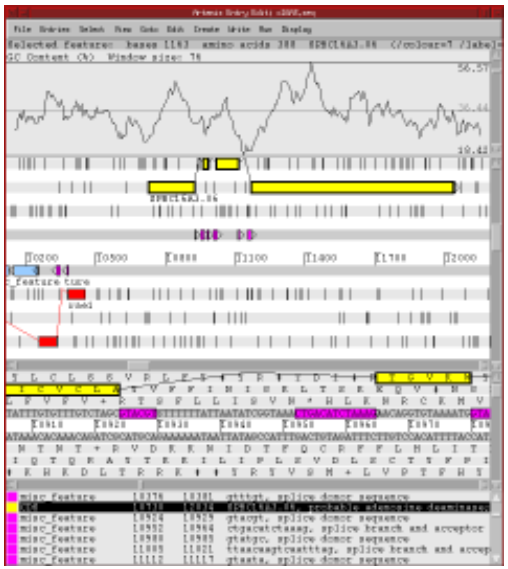
Putting it all together manually

- View *ab initio* predictions and sequence alignments within a genome editor:

Apollo

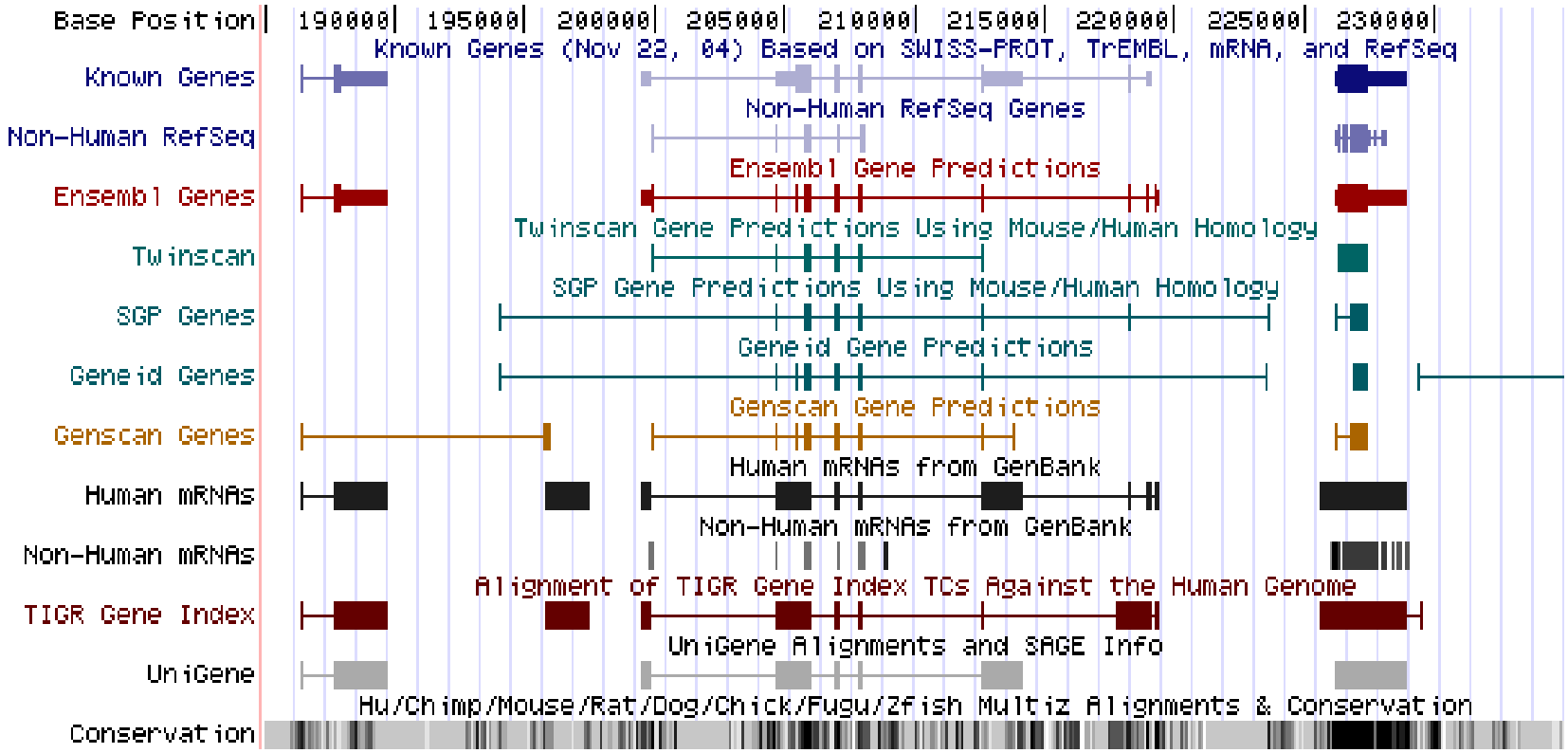


Artemis



Annotation Station

Putting it all together automatically with JIGSAW



Evaluating methods on 1% of the human genome: ENCODE

SPECIAL SECTION

GENES IN ACTION

VIEWPOINT

The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*†

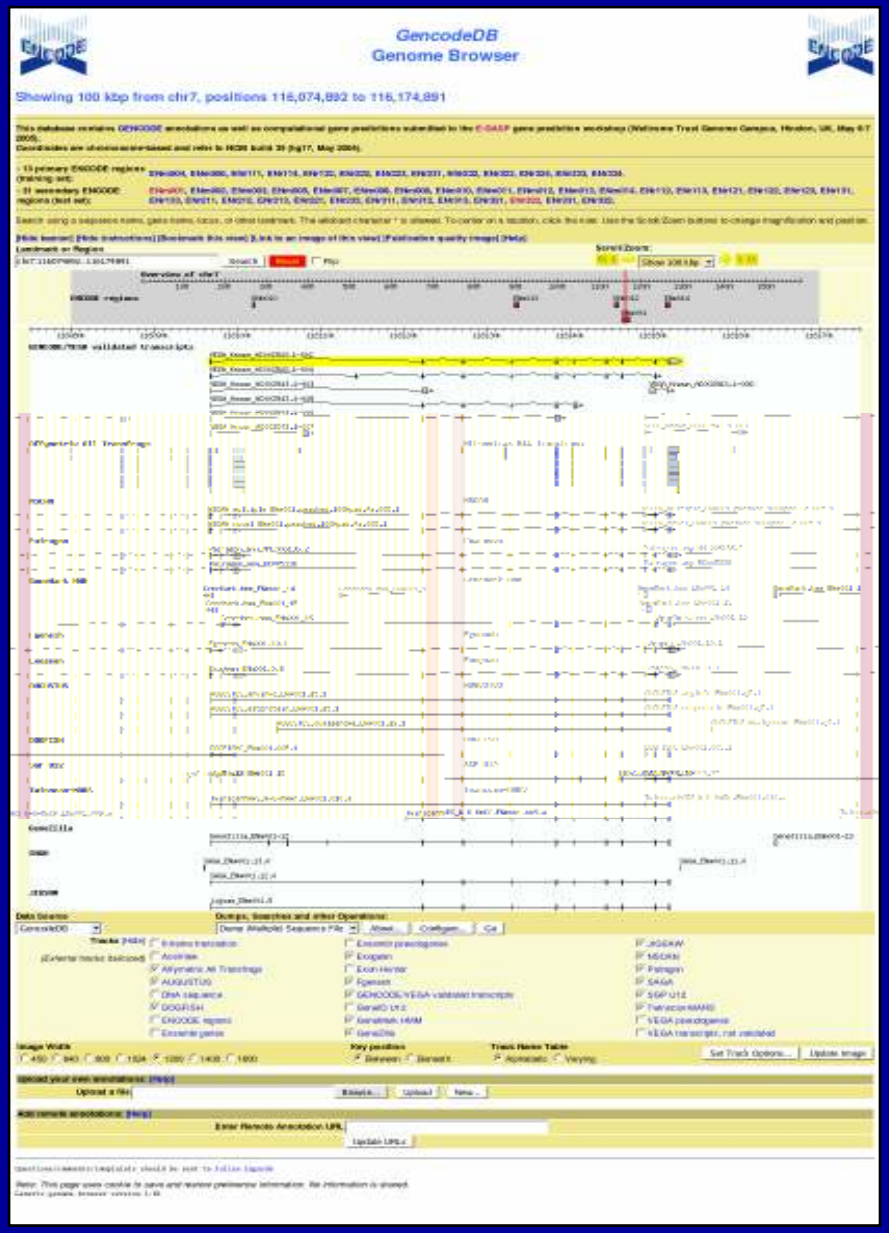
The ENCyclopedia Of DNA Elements (ENCODE) Project aims to identify all functional elements in the human genome sequence. The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.

With the complete human genome sequence now in hand (*1*–*3*), we face the enormous challenge of interpreting it and learning how to use that information to understand the biology of human health and disease. The ENCyclopedia Of DNA Elements (ENCODE) Project is predicated on the belief that a comprehensive catalog of the structural and functional components encoded in the human genome sequence will be critical for understanding human biology well enough to address those fundamental aims of biomedical research. Such a complete catalog, or "parts list," would include protein-coding

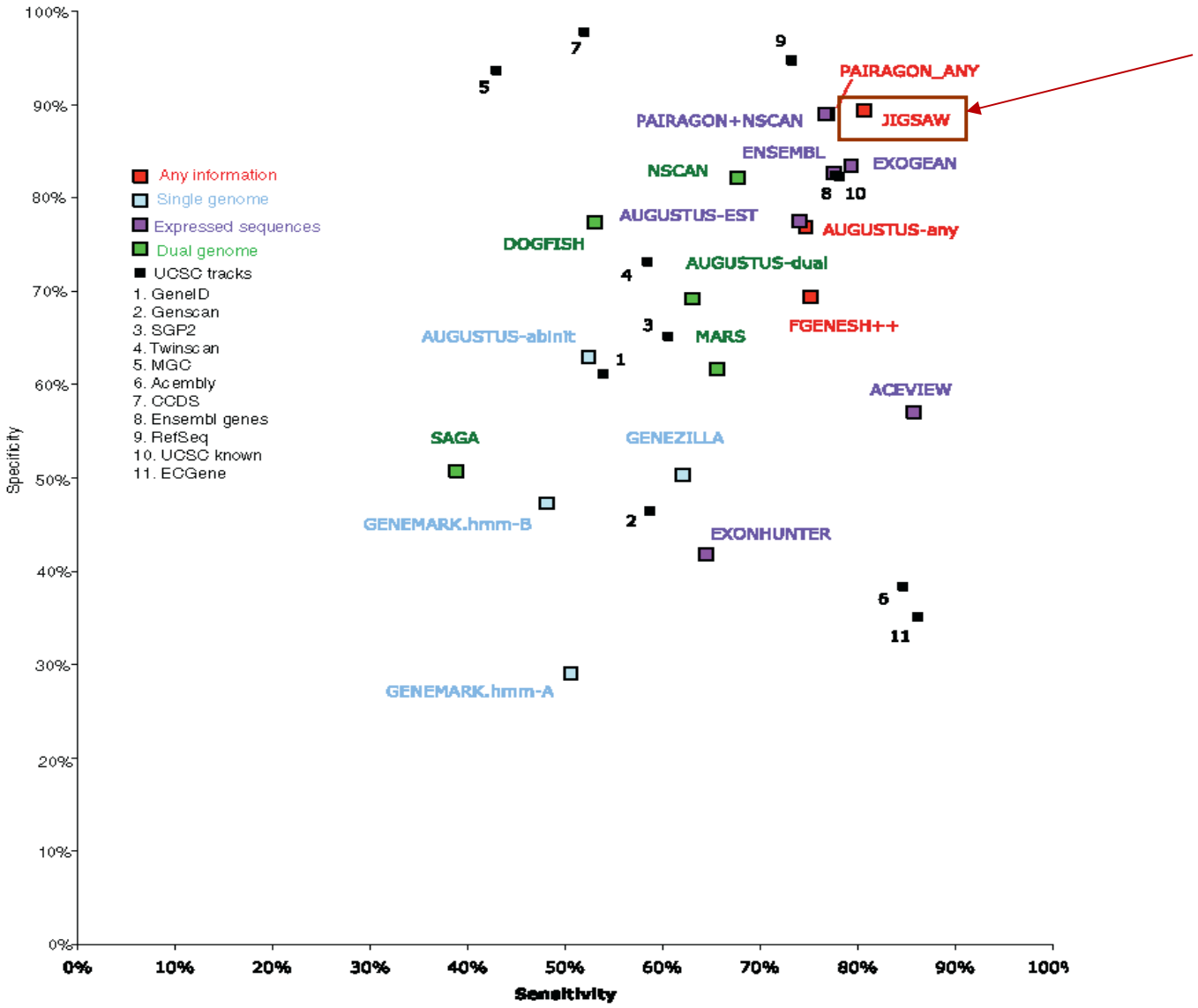
approaches, such as cDNA-cloning efforts (*4*, *5*) and chip-based transcriptome analyses (*6*, *7*), have revealed the existence of many transcribed sequences of unknown function. As a reflection of this complexity, about 5% of the human genome is evolutionarily conserved with respect to rodent genomic

Science 306 (2004), 636-640.

ENCODE Gene finding ASsessment Project (EGASP)



The GencodeDB Genome Browser. Fig. 2 from Guigo *et al.*, *Genome Biology* 2006, 7(Suppl 1):S2.



EGASP results: Exon prediction accuracy among 28 different methods

From Guigo *et al*., *Genome Biology* 2006, 7(Suppl 1):S2

EGASP results: overall gene accuracy

Genefinding Method	Sens	Spec
AUGUSTUS-any	47.9	35.5
FGENESH++	69.9	42.0
JIGSAW	72.6	65.9
PAIRAGON-any	69.5	61.3
AUGUSTUS-abinit	24.3	17.2
GENEMARK.hmm-A	15.2	3.2
GENEMARK.hmm-B	16.8	7.9
GENEZILLA	19.5	8.8
ACEVIEW	63.5	48.6
AUGUSTUS-EST	47.6	37.0
ENSEMBL	71.6	67.3
EXOGEAN	63.1	80.8
EXONHUNTER	21.9	6.3
PAIRAGON+NSCAN_EST	69.5	61.7
AUGUSTUS-dual	26.0	18.6
DOGFISH	10.8	14.6
MARS	33.4	24.9
NSCAN	35.4	36.7
SAGA	4.3	3.4
GENSCAN	15.5	10.1
KNOWNgene	77.0	72.7
TWINSCAN	22.3	20.2

So, where are we now?

Ensembl genes
www.ensembl.org

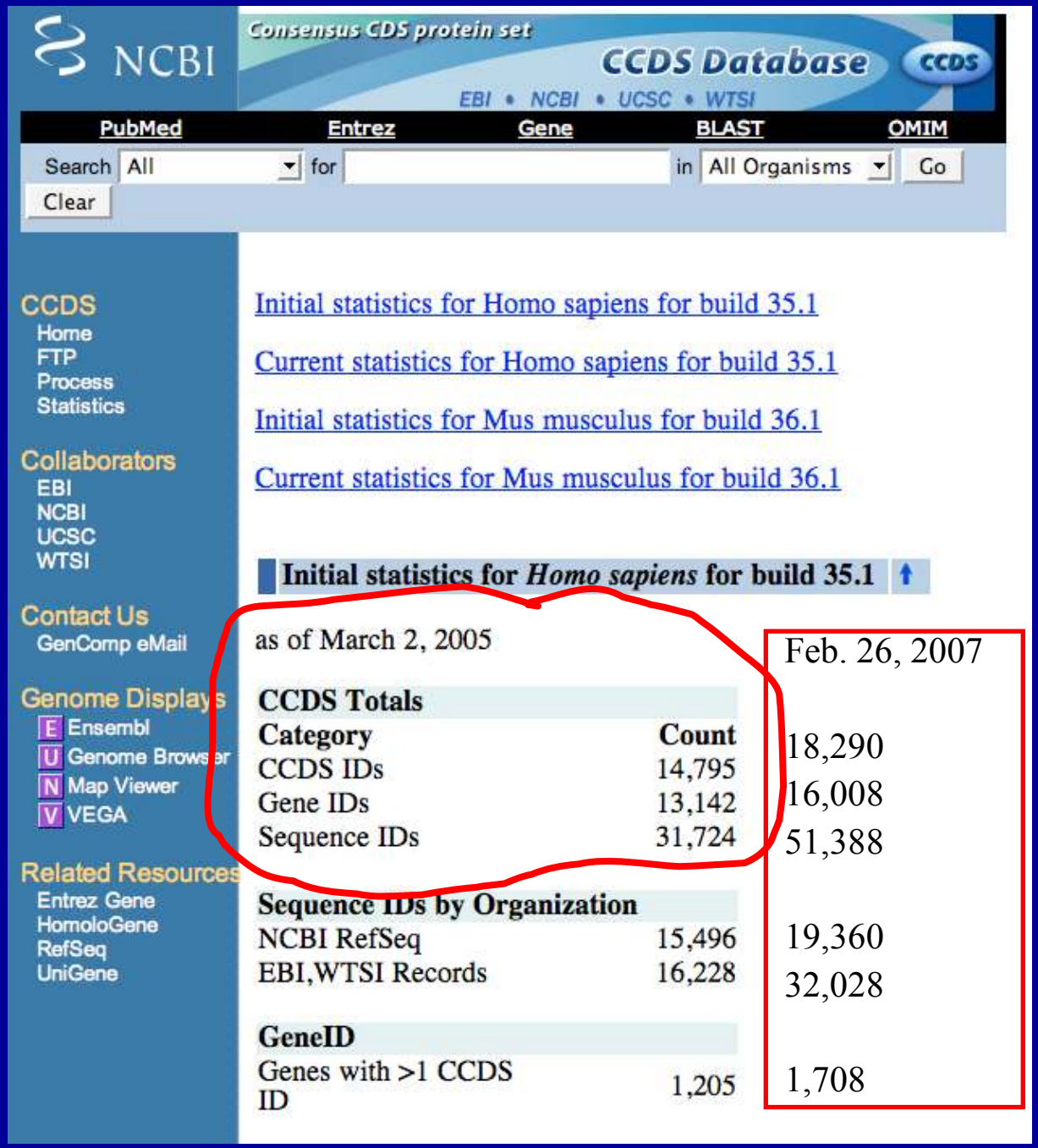
- 21,774 “Known genes”
- 1,036 Novel genes
- 3,994 RNA genes
- 69,185 Genscan gene predictions

Let’s not forget pseudogenes

- 27,130 pseudogenes
- www.pseudogene.org

Current NCBI gene counts www.ncbi.nih.gov

- Entrez Gene:
  - 38,621 genes
  - But this includes pseudogenes
- RefSeq:
  - 28,961 genes
  - 31,784 transcripts



Lesson:

Science isn't decided by voting

So, we still don't have a gene count

…and for many genes, we aren't yet sure of their exon-intron structure…

…and there are > 1000 other genomes already complete or under way…

…so we aren't giving up!

Thanks…

Mihaela Pertea

Jonathan Allen
(LLNL)

Bill Majoros
(Duke Univ.)

Art Delcher

Brian Haas
(Broad Institute)

Evaluating Gene Predictions

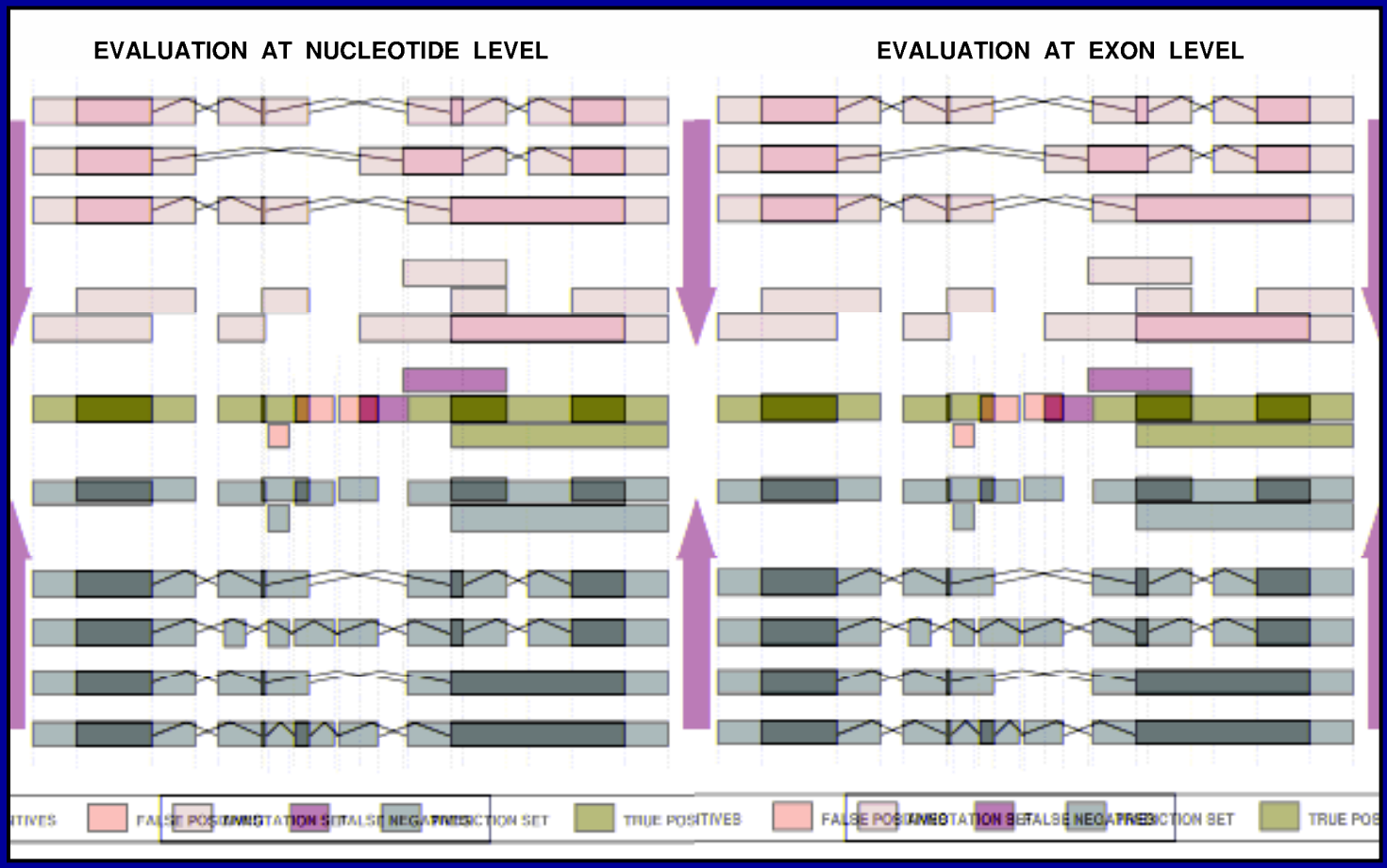


Fig. 3 from Guigo *et al.*, *Genome Biology* 2006, 7(Suppl 1):S2.