Gene prediction in animal genomes

Shigehiro Kuraku

University of Konstanz

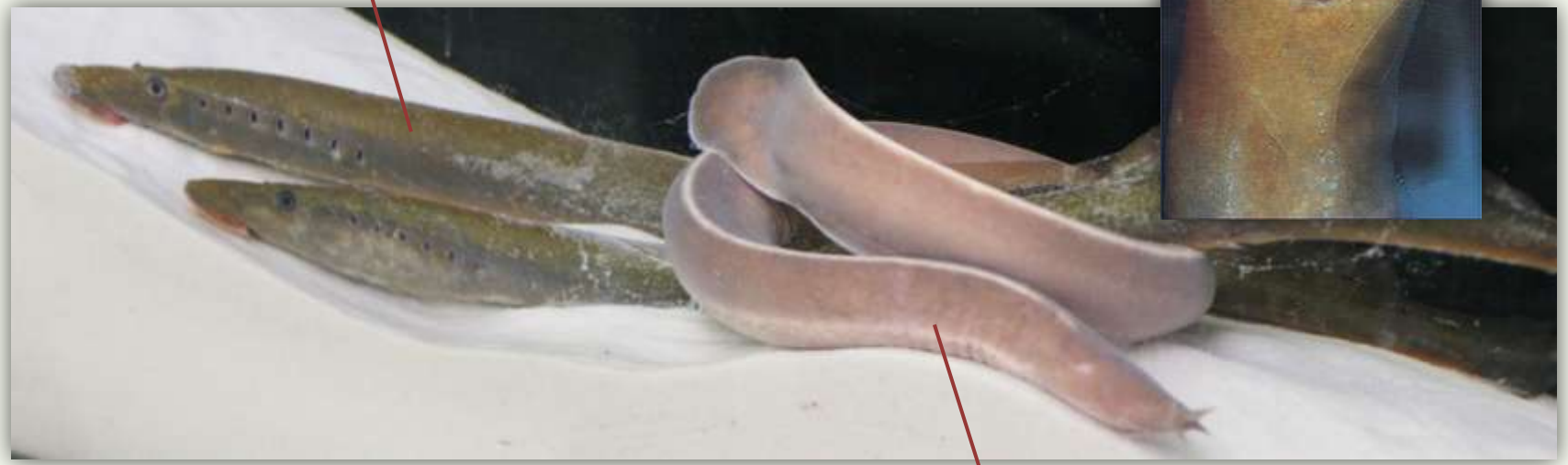
May 16, 2011. ETH Zurich

Gene prediction in animal genomes

Cyclostomata (hagfishes and lampreys)

sea lamprey
Petromyzon marinus

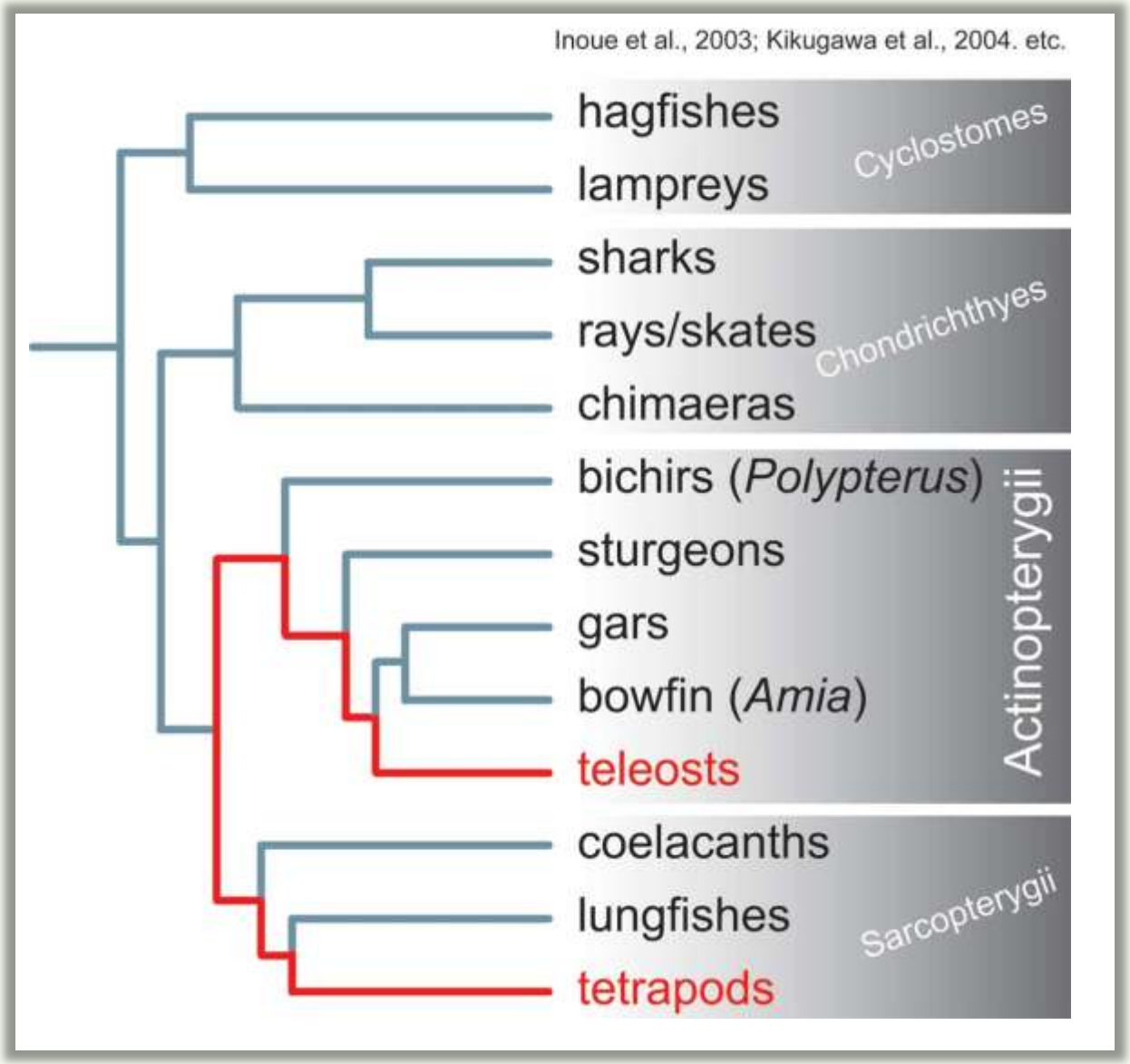
Japanese lamprey
Lethenteron japonicum



By K.G. Ota, S. Kuraku and S. Kuratani

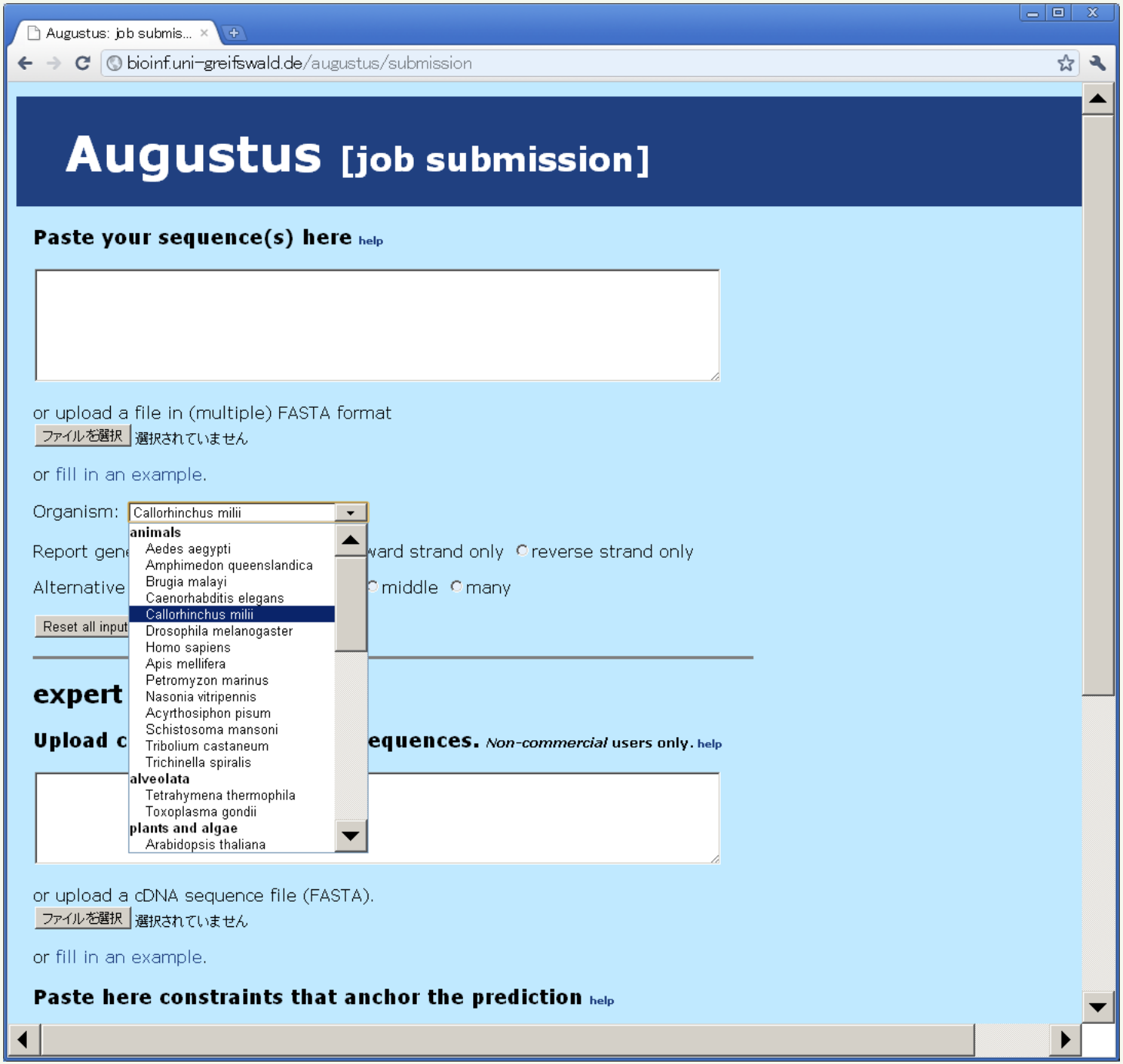
Inshore hagfish
Eptatretus burgeri

Highly biased distribution of sequenced genomes



5.9 x

1.4 x


Augustus: job submis...
bioinf.uni-greifswald.de/augustus/submission
Augustus [job submission]
Paste your sequence(s) here help
or upload a file in (multiple) FASTA format
ファイルを選択 選択されていません
or fill in an example.
Organism: Callorhinchus milii
animals
Aedes aegypti
Amphimedon queenslandica
Brugia malayi
Caenorhabditis elegans
Callorhinchus milii
Drosophila melanogaster
Homo sapiens
Apis mellifera
Petromyzon marinus
Nasonia vitripennis
Acyrthosiphon pisum
Schistosoma mansoni
Tribolium castaneum
Trichinella spiralis
alveolata
Tetrahymena thermophila
Toxoplasma gondii
plants and algae
Arabidopsis thaliana
ward strand only
reverse strand only
middle
many
Reset all input
expert
equences. Non-commercial users only. help
or upload a cDNA sequence file (FASTA).
ファイルを選択 選択されていません
or fill in an example.
Paste here constraints that anchor the prediction help

Gene prediction in animal genomes

What is a ‘gene’ ?

‘one-gene-one-enzyme hypothesis’

Beadle and Tatum, 1941.

***Proc Natl Acad Sci USA*, 27: 499-**

*GENETIC CONTROL OF BIOCHEMICAL REACTIONS IN NEUROSPORA**

BY G. W. BEADLE AND E. L. TATUM

BIOLOGICAL DEPARTMENT, STANFORD UNIVERSITY

Communicated October 8, 1941

From the standpoint of physiological genetics the development and functioning of an organism consist essentially of an integrated system of chemical reactions controlled in some manner by genes. It is entirely tenable to suppose that these genes which are themselves a part of the system, control or regulate specific reactions in the system either by acting directly as enzymes or by determining the specificities of enzymes.[1] Since the components of such a system are likely to be interrelated in complex ways, and since the synthesis of the parts of individual genes are presumably dependent on the functioning of other genes, it would appear that there must exist orders of directness of gene control ranging from simple one-to-one relations to relations of great complexity. In investigating the rôles of genes, the physiological geneticist usually attempts to determine the physiological and biochemical bases of already known hereditary traits. This approach, as made in the study of anthocyanin pigments in plants,[2] the fermentation of sugars by yeasts[3] and a number of other instances,[4] has established that many biochemical reactions are in fact controlled in specific ways by specific genes. Furthermore, investigations of this type tend to support the assumption that gene and enzyme

Pertea and Salzberg *Genome Biology* 2010, 11:206
http://genomebiology.com/2010/11/5/206

Genome Biology

REVIEW

Between a chicken and a grape: estimating the number of human genes

Mihaela Pertea and Steven L. Salzberg*

We still do not know how many genes there are in our genomes

In this lecture, a 'gene' means a 'protein-coding' gene

Alternative splicing and non-coding RNA confuse gene prediction

There are two major types of gene predictors
- a) *ab initio* (intrinsic) – gHMM (e.g., GenScan)
- b) evidence-based (extrinsic) – EST or/and homology (e.g., Contrast, Jigsaw)

The most likely gene count in the human genome is 22,333

Gene prediction in animal genomes

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes

Mark Yandell[1,2,3¤a*], Chris J. Mungall[1,2], Chris Smith[3], Simon Prochnik[3¤b], Joshua Kaminker[3¤c], George Hartzell[3], Suzanna Lewis[3], Gerald M. Rubin[1,2,3]

1 Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, California, United States of America, **2** Howard Hughes Medical Institute, University of California Berkeley, Berkeley, California, United States of America, **3** Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

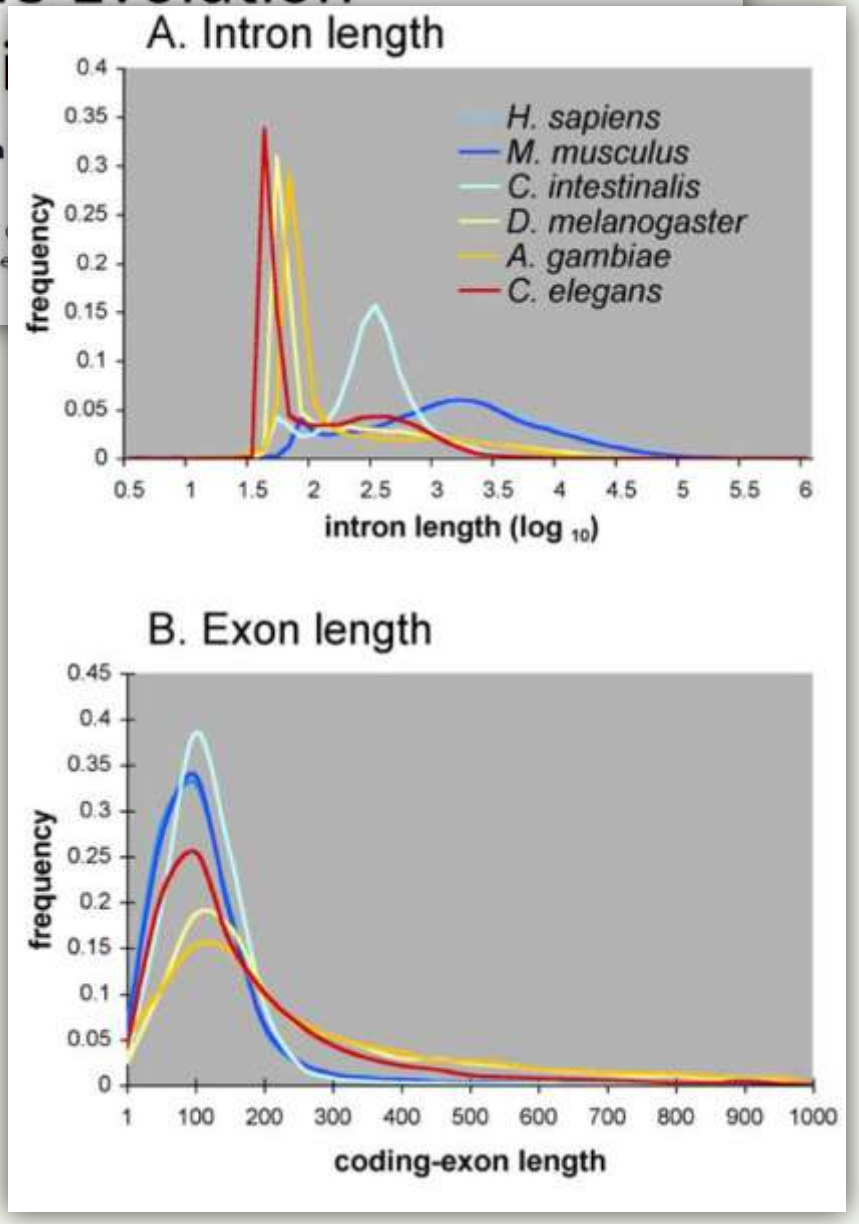
OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Large-Scale Trends in the Evolution of Gene Structures with

Mark Yandell[1,2,3¤a*], **Chris J. Mungall**[1,2], **Chris Smith**[3], **Simon**
Suzanna Lewis[3], **Gerald M. Rubin**[1,2,3]

1 Department of Molecular and Cell Biology, University of California Berkeley, Berkeley,
California Berkeley, Berkeley, California, United States of America, 3 Department of Ge
States of America



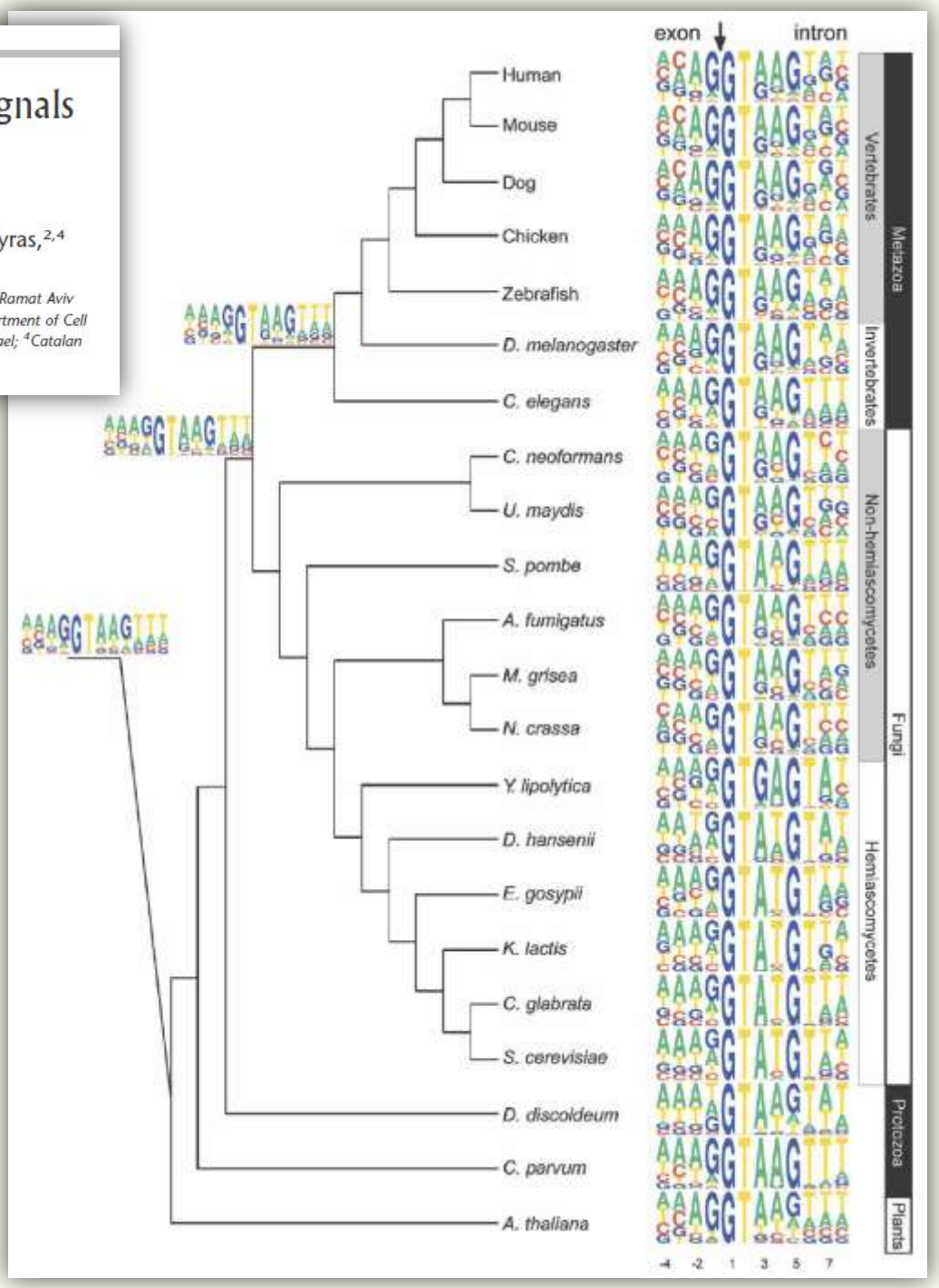
Letter

Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes

Schraga H. Schwartz,[1] João Silva,[2] David Burstein,[3] Tal Pupko,[3] Eduardo Eyras,[2,4] and Gil Ast[1,5]

[1]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel-Aviv University, Ramat Aviv 69978, Israel; [2]Biomedical Informatics Unit, Pompeu Fabra University, PRBB E08003, Barcelona, Spain; [3]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel; [4]Catalan Institution for Research and Advanced Studies (ICREA), E08010, Barcelona, Spain

Genome Res. 2008. 18:88-



>99% of introns begin with ‘GT’

_computational BIOLOGY

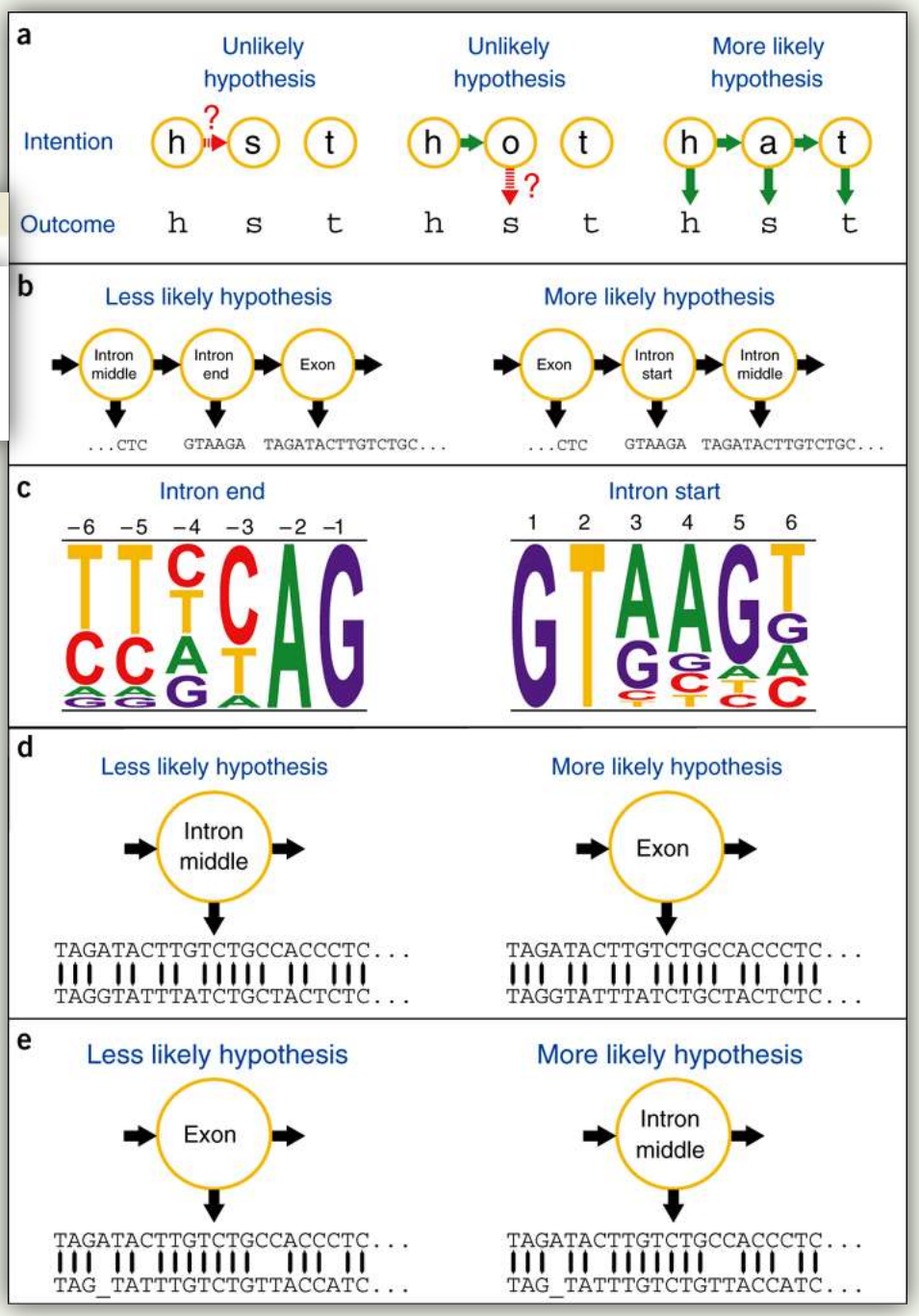
PRIMER

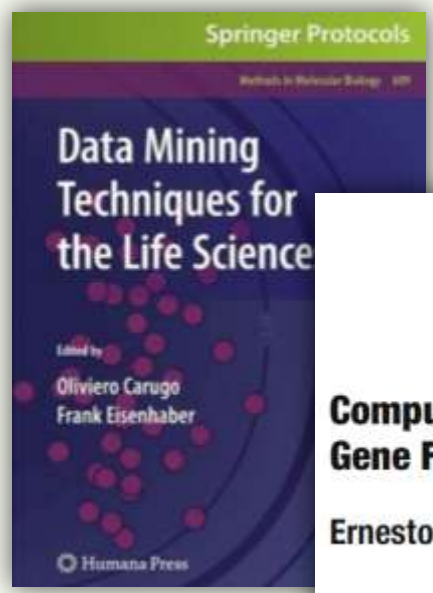
How does eukaryotic gene prediction work?

Michael R Brent

Computational prediction of gene structure is crucial for interpreting genomic sequences. But how do the algorithms involved work and how accurate are they?

NATURE BIOTECHNOLOGY VOLUME 25 NUMBER 8 AUGUST 2007





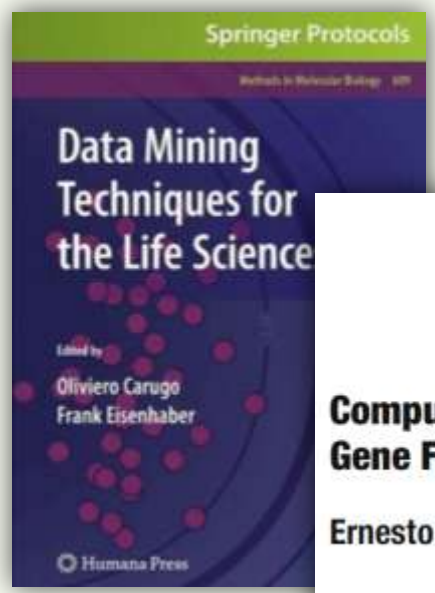
Chapter 16

Computational Methods for Ab Initio and Comparative Gene Finding

Ernesto Picardi and Graziano Pesole

Abstract

High-throughput DNA sequencing is increasing the amount of public complete genomes even though a precise gene catalogue for each organism is not yet available. In this context, computational gene finders play a key role in producing a first and cost-effective annotation. Nowadays a compilation of gene prediction tools has been made available to the scientific community and, despite the high number, they can be divided into two main categories: (1) ab initio and (2) evidence based. In the following, we will provide an overview of main methodologies to predict correct exon–intron structures of eukaryotic genes falling in such categories. We will take into account also new strategies that commonly refine ab initio predictions employing comparative genomics or other evidence such as expression data. Finally, we will briefly introduce metrics to in house evaluation of gene predictions in terms of sensitivity and specificity at nucleotide, exon, and gene levels as well.

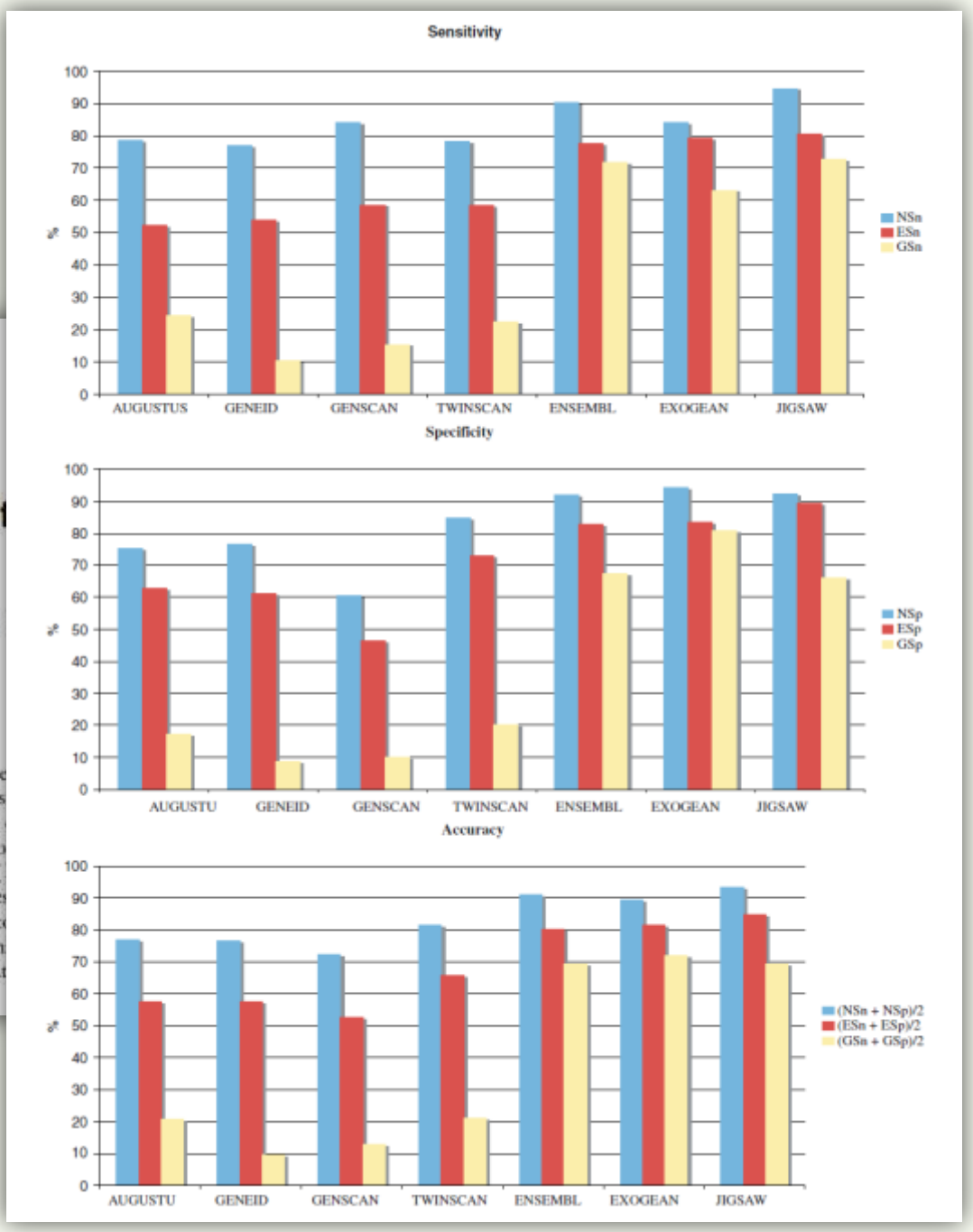


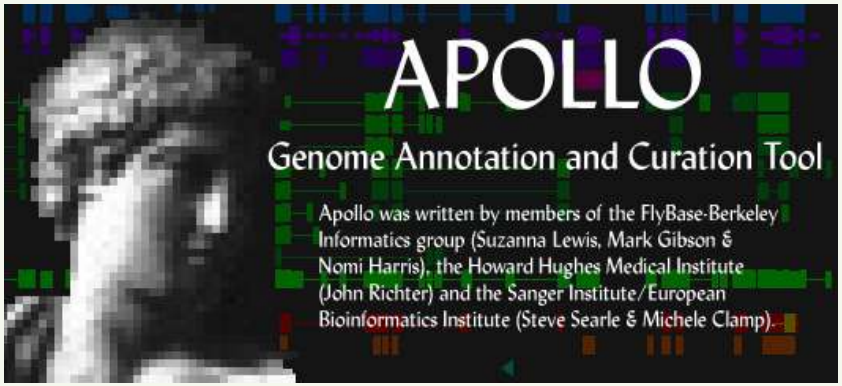
Computational Methods f
Gene Finding

Ernesto Picardi and Graziano

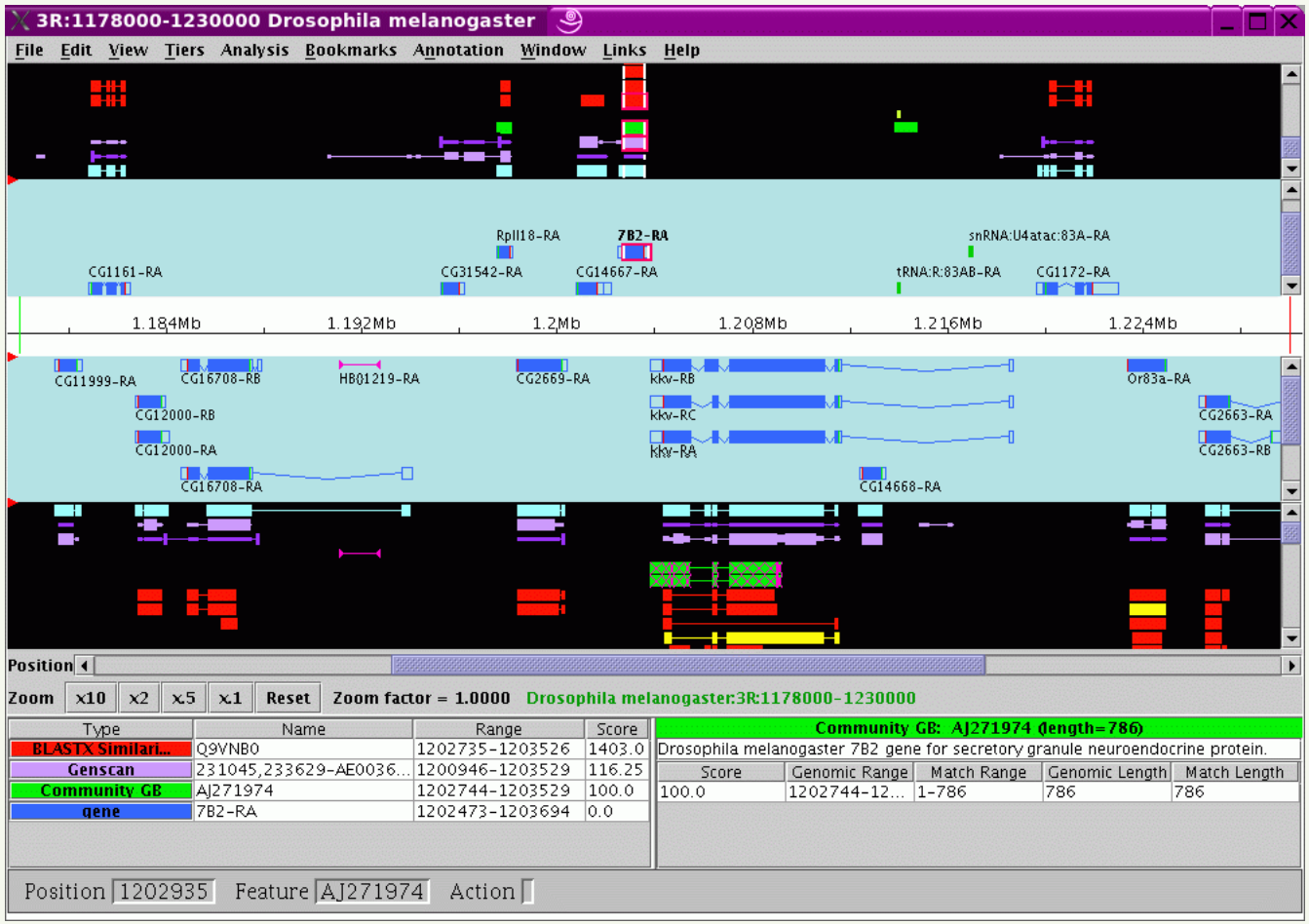
Abstract

High-throughput DNA sequencing is incre
precise gene catalogue for each organism is
play a key role in producing a first and
prediction tools has been made available to
can be divided into two main categories: (
provide an overview of main methodologies
falling in such categories. We will take int
predictions employing comparative genom
briefly introduce metrics to in house evaluat
nucleotide, exon, and gene levels as well.





http://apollo.berkeleybop.org/current/userguide.html




nature
Tetraodon to human
nature
THE OPOSSUM GENOME
nature
nature
THE AMPHIOXUS GENOME
nature
the human genome
The mouse genome
Science
The rat genome
Insights into mammalian evolution

nature Vol 464|25 March 2010|doi:10.1038/nature08830

LETTERS

The dynamic genome of *Hydra*

Jarrod A. Chapman[1]*, Ewen F. Kirkness[2]*, Oleg Simakov[3,4]*, Steven E. Hampson[5]‡, Therese Mitros[4], Thomas Weinmaier[6], Thomas Rattei[6], Prakash G. Balasubramanian[3], Jon Borman[2], Dana Busam[2], Kathryn Disbennett[2], Cynthia Pfannkoch[2], Nadezhda Sumin[2], Granger G. Sutton[2], Lakshmi Devi Viswanathan[2], Brian Walenz[2], David M. Goodstein[1], Uffe Hellsten[1], Takeshi Kawashima[4], Simon E. Prochnik[1], Nicholas H. Putnam[1,4]†, Shengquiang Shu[1], Bruce Blumberg[7,8], Catherine E. Dana[8,9], Lydia Gee[7,8], Dennis F. Kibler[5], Lee Law[7,8], Dirk Lindgens[7,8], Daniel E. Martinez[10], Jisong Peng[7,8], Philip A. Wigge[11]†, Bianca Bertulat[3], Corina Guder[3], Yukio Nakamura[3], Suat Ozbek[3], Hiroshi Watanabe[3], Konstantin Khalturin[12], Georg Hemmrich[12], André Franke[12], René Augustin[12], Sebastian Fraune[12], Eisuke Hayakawa[13], Shiho Hayakawa[13], Mamiko Hirose[13], Jung Shan Hwang[13], Kazuho Ikeo[13], Chiemi Nishimiya-Fujisawa[13], Atshushi Ogura[13]†, Toshio Takahashi[14], Patrick R. H. Steinmetz[15], Xiaoming Zhang[16], Roland Aufschnaiter[17], Marie-Kristin Eder[17], Anne-Kathrin Gorny[17]†, Willi Salvenmoser[17], Alysha M. Heimberg[18], Benjamin M. Wheeler[19], Kevin J. Peterson[18], Angelika Böttger[20], Patrick Tischler[6], Alexander Wolf[20], Takashi Gojobori[13], Karin A. Remington[2]†, Robert L. Strausberg[2], J. Craig Venter[2], Ulrich Technau[15], Bert Hobmayer[17], Thomas C. G. Bosch[12], Thomas W. Holstein[3], Toshitaka Fujisawa[13], Hans R. Bode[7,8], Charles N. David[20], Daniel S. Rokhsar[1,4] & Robert E. Steele[8,9]

S6. Construction of gene models

Homology based gene modeling was done with GenomeScan[24] using the RP genome assembly. Putative loci were found by blastx of soft-masked genomic scaffolds versus the proteomes of *Nematostella vectensis*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Additionally, *Hydra* EST assemblies were aligned to the genome with BLAT[18]. The BLAT output was processed so that the best hit to the genome, as well as any other hit within 97% coverage of the best hit were considered matches. Possible pseudogene matches were filtered out by disallowing secondary matches with only one exon if the best hit has multiple exons. Putative loci were defined by these peptide and EST hits and joined if overlapping. Each region with flanking sequence was submitted with its best template from each organism to GenomeScan and the resulting models were run through PASA[25], which verified/improved some of the models.

AUGUSTUS 2.0.3[26] was trained on 1061 EST assemblies suggested by PASA to be full-length and good training set candidates. These candidate genes were filtered so that they encoded at least 100 amino acids and more than one exon. The gene models were created by running AUGUSTUS on an unmasked version of the genome, incorporating *Hydra* and *Clytia* EST evidence. The majority of the produced models corresponded to transposable element proteins. The models were therefore filtered to remove genes that have more than 50% of their exonic length overlapping with an annotated transposable element (see Supplementary Information Section S9). The final gene model set was produced by running PASA on the filtered AUGUSTUS and the original GenomeScan predictions. About 9,000 models could be verified and/or improved by PASA. PASA-unverified models with no homology to any known protein in the GenBank nr database and no EST support (from *Hydra* or *Clytia*) were removed. If a locus had models from both predictors (GenomeScan and AUGUSTUS), the best model was selected based on its hit e-value to the GenBank nr database and EST support.

Chapman et al., 2010. Nature. 464: 592-

doi: 10.1038/nature08830

SUPPLEMENTARY INFORMATION

The final gene model set contains 31,452 genes. In addition to our annotation of the RP assembly, the CA assembly was annotated by NCBI using the Gnomon gene prediction pipeline (http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml). The 17,835 predicted protein sequences from this annotation have been deposited in GenBank.

The 31,452 predicted protein-coding loci from the RP assembly are an overestimate, as they may include predictions that are not *bona fide* protein-coding genes. Such spurious predictions could arise from unrecognized repetitive elements. Splitting of genes between scaffolds would also contribute to an over-estimate of gene number. Clustering of the 31,452 predicted *Hydra* genes with those of other metazoans identifies 22,083 *Hydra* gene models that have homology to other metazoan genes or are members of paralogous groups in the *Hydra* genome. Of these 22,083 models, 2,117 were found, after manual review, to be from transposons and other repetitive elements. Thus there are 19,966 predicted genes from the RP assembly that meet one of the following three criteria: (1) homology to other metazoan genes; (2) membership in a *Hydra*-specific paralogous group; (3) not from a transposon or repetitive element. This analysis does not capture *Hydra*-specific genes that are not members of paralogous groups.

While the remaining 9,369 predictions (31,452 minus 22,083) may contain additional *bona fide* *Hydra*-specific genes that are not part of gene families, most are likely missed repetitive sequences, artifacts of gene prediction algorithms, or pseudogenes. Only ~30% (2,892) have some (partial) support from ESTs, and of these many (902) are single or two-exon genes that appear to be enriched in pseudogenes. Taken together, we estimate that the *Hydra* genome encodes ~20,000 genes, although this is necessarily a rough estimate based on the above considerations.

Chapman et al., 2010. Nature. 464: 592-

Reality

- Even well-known databases contain ……. (e.g., *Pax4*, *Bmp16*)

- Most users are genome annotators in sequencing centers

- Decisions in big/expensive projects are usually highly political

- Most other users still work on genome annotation

- Few users are laboratory-oriented and naïve with *in silico* tools

- Evaluation of gene prediction is difficult (gene/exon/nucleotide)

- Always better to increase accuracy ? (= [sensitivity + specificity]/2)

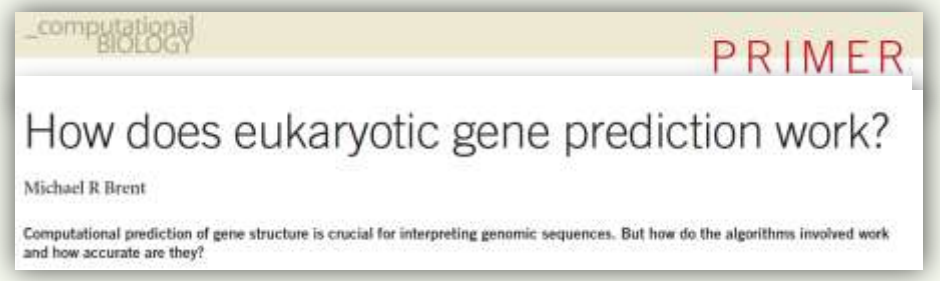
- Technologies and forms of genome seq. projects are evolving

Aims

Accomplish the first review from the users' viewpoint

1. Cover all **existing review papers** on gene prediction
2. List **which gene predictors** annotated which genomes
3. Who are **current and potential users** of gene predictors?
4. Consider **typical research contexts** in need of gene prediction
5. Outline challenges based on **difficult cases**
6. Speculate challenges related to **next generation sequencing (NGS)**
7. Discuss how we can **assess quality** of gene prediction
8. Any other clue for improvement in quality and user-friendliness?

1. Existing review papers



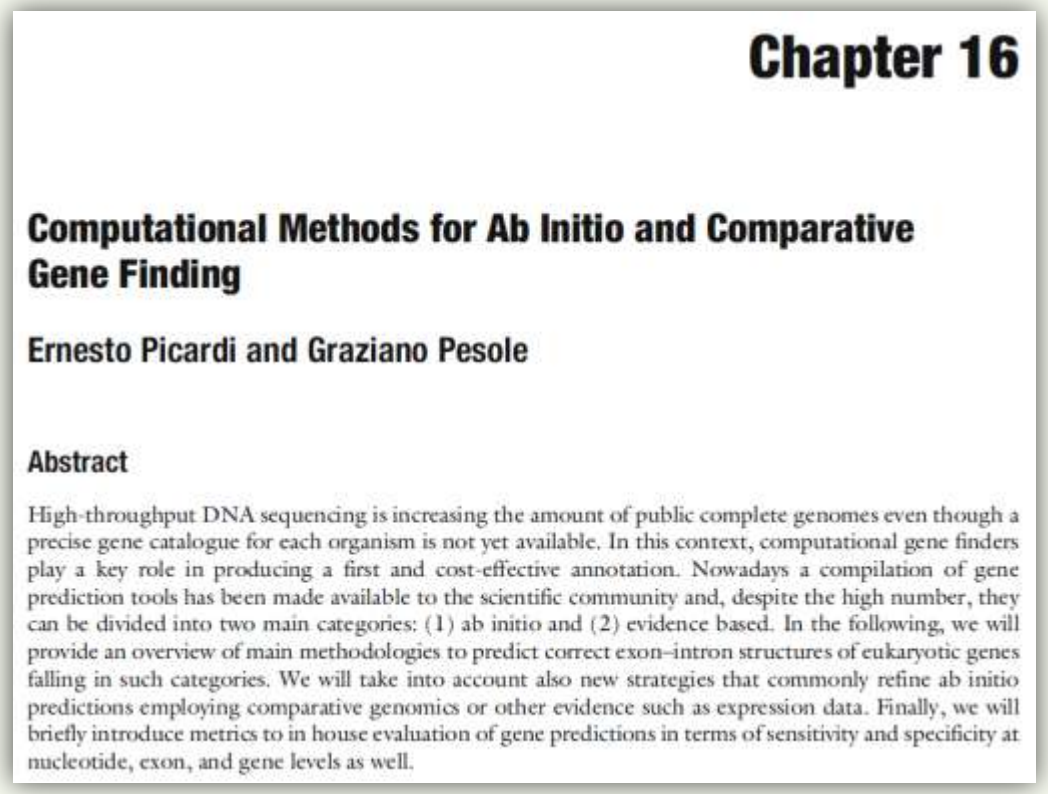
_computational BIOLOGY

PRIMER

How does eukaryotic gene prediction work?

Michael R Brent

Computational prediction of gene structure is crucial for interpreting genomic sequences. But how do the algorithms involved work and how accurate are they?



Chapter 16

Computational Methods for Ab Initio and Comparative Gene Finding

Ernesto Picardi and Graziano Pesole

Abstract

High-throughput DNA sequencing is increasing the amount of public complete genomes even though a precise gene catalogue for each organism is not yet available. In this context, computational gene finders play a key role in producing a first and cost-effective annotation. Nowadays a compilation of gene prediction tools has been made available to the scientific community and, despite the high number, they can be divided into two main categories: (1) ab initio and (2) evidence based. In the following, we will provide an overview of main methodologies to predict correct exon–intron structures of eukaryotic genes falling in such categories. We will take into account also new strategies that commonly refine ab initio predictions employing comparative genomics or other evidence such as expression data. Finally, we will briefly introduce metrics to in house evaluation of gene predictions in terms of sensitivity and specificity at nucleotide, exon, and gene levels as well.

2. Which gene predictors annotated which genomes?

3. Current and potential users

4. Typical research contexts requiring gene prediction

- Bulky genome projects
- Compact genome projects
- Other projects (lab approach-based)

Experiment: Search in ISI for papers citing GenScan/Augustus

Positional cloning of the APECED gene

Kentaro Nagamine[1,2*], Pärt Peterson[3*], Hamish S. Scott[4*], Jun Kudoh[1], Shinsei Minoshima[1], Maarit Heino[3], Kai J. E. Krohn[3], Maria D. Lalioti[4], Primus E. Mullis[5], Stylianos E. Antonarakis[4], Kazuhiko Kawasaki[1], Shuichi Asakawa[1], Fumiaki Ito[2] & Nobuyoshi Shimizu[1]

Nat. Genet. 1997. 17:393

APECED = autoimmune polyendocrinopathy-candidiasis-ectodermal-dystrophy (Polyendokrine Autoimmunerkrankungen)

*"We used genomic sequence information to determine that the distance between D21S1912 and PFKL is approxitmately **140 kb** (Fig. 1). Using computer programs such as **GRAIL** and **GENSCAN**, we performed gene screening in the sequencing data within this regions. **GENSCAN** predicted **several genes** between D21S1912 and PFKL. One of these genes, **AIRE**, is located just proximal to the PFKL gene and contained the previously trapped exons HC21EXc33 (ref.9) and MDL04M06 (ref.19)."*

Positional cloning of the APECED gene

Kentaro Nagamine[1,2*], Pärt Peterson[3*], Hamish S. Scott[4*], Jun Kudoh[1], Shinsei Minoshima[1], Maarit Heino[3], Kai J. E. Krohn[3], Maria D. Lalioti[4], Primus E. Mullis[5], Stylianos E. Antonarakis[4], Kazuhiko Kawasaki[1], Shuichi Asakawa[1], Fumiaki Ito[2] & Nobuyoshi Shimizu[1]

Nat. Genet. 1997. 17:393

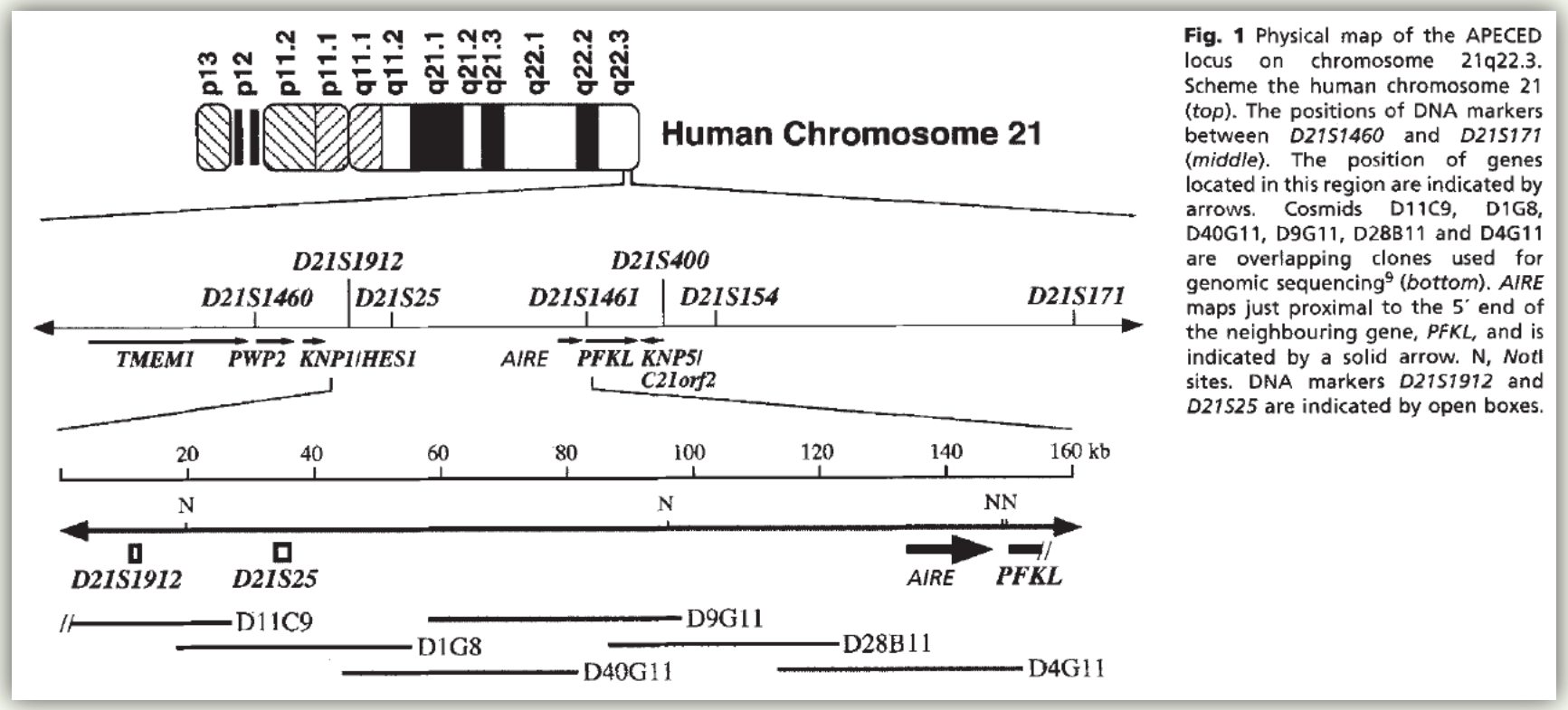


Fig. 1 Physical map of the APECED locus on chromosome 21q22.3. Scheme the human chromosome 21 (*top*). The positions of DNA markers between *D21S1460* and *D21S171* (*middle*). The position of genes located in this region are indicated by arrows. Cosmids D11C9, D1G8, D40G11, D9G11, D28B11 and D4G11 are overlapping clones used for genomic sequencing[9] (*bottom*). *AIRE* maps just proximal to the 5′ end of the neighbouring gene, *PFKL*, and is indicated by a solid arrow. N, *Not*I sites. DNA markers *D21S1912* and *D21S25* are indicated by open boxes.

3. Current and potential users

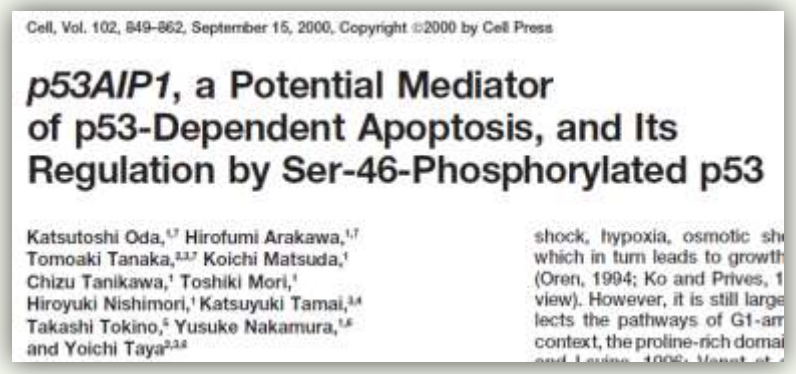
4. Typical research contexts requiring gene prediction

- Bulky genome projects
- Compact genome projects
- Other projects (lab approach-based)

Positional cloning of the APECED gene

Kentaro Nagamine[1,2*], Pärt Peterson[3*], Hamish S. Scott[4*], Jun Kudoh[1], Shinsei Minoshima[1], Maarit Heino[3], Kai J. E. Krohn[3], Maria D. Lalioti[4], Primus E. Mullis[5], Stylianos E. Antonarakis[4], Kazuhiko Kawasaki[1], Shuichi Asakawa[1], Fumiaki Ito[2] & Nobuyoshi Shimizu[1]

Nat. Genet. 1997. 17:393

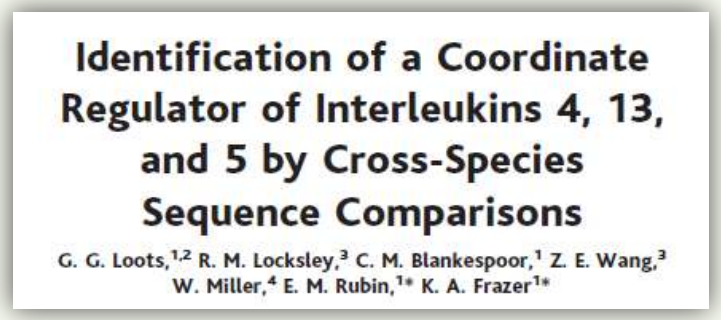


Cell, Vol. 102, 849–862, September 15, 2000, Copyright ©2000 by Cell Press

***p53AIP1*, a Potential Mediator of p53-Dependent Apoptosis, and Its Regulation by Ser-46-Phosphorylated p53**

Katsutoshi Oda,[1,7] Hirofumi Arakawa,[1,7] Tomoaki Tanaka,[2,3,7] Koichi Matsuda,[1] Chizu Tanikawa,[1] Toshiki Mori,[1] Hiroyuki Nishimori,[1] Katsuyuki Tamai,[3,4] Takashi Tokino,[5] Yusuke Nakamura,[1,6] and Yoichi Taya[2,3,6]

shock, hypoxia, osmotic sh
which in turn leads to growth
(Oren, 1994; Ko and Prives, 1
view). However, it is still large
lects the pathways of G1-ar
context, the proline-rich domai

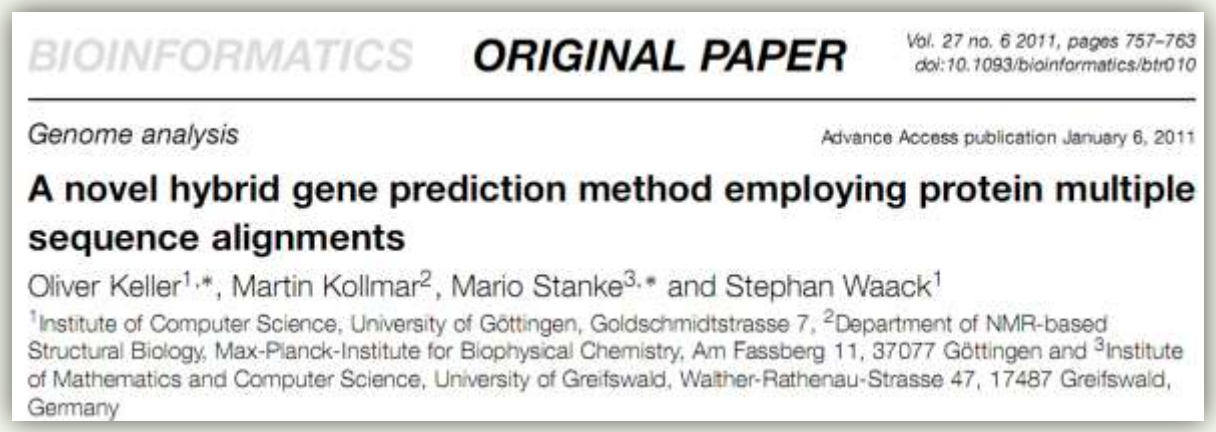


Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons

G. G. Loots,[1,2] R. M. Locksley,[3] C. M. Blankespoor,[1] Z. E. Wang,[3] W. Miller,[4] E. M. Rubin,[1*] K. A. Frazer[1*]

Science 2000. 288: 136

Find the right tool for your demand*!*



BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 6 2011, pages 757–763
doi:10.1093/bioinformatics/btr010

Genome analysis
Advance Access publication January 6, 2011

A novel hybrid gene prediction method employing protein multiple sequence alignments

Oliver Keller[1,*], Martin Kollmar[2], Mario Stanke[3,*] and Stephan Waack[1]

[1]Institute of Computer Science, University of Göttingen, Goldschmidtstrasse 7, [2]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen and [3]Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Strasse 47, 17487 Greifswald, Germany

*"…, manually assembled and verified **dynein heavy chain (DHC)** sequences are regarded as almost correct predictions and taken as reference sequences for the test runs. …… The task of recognizing a gene as a member of the DHC protein family was accomplished by **AUGUSTUS**-PPX in almost all cases. Four sequences were not identified as DHCs, three from less similar subfamilies and one case with an incomplete genomic reference sequence."*

What would you do if you already know the target proteins?

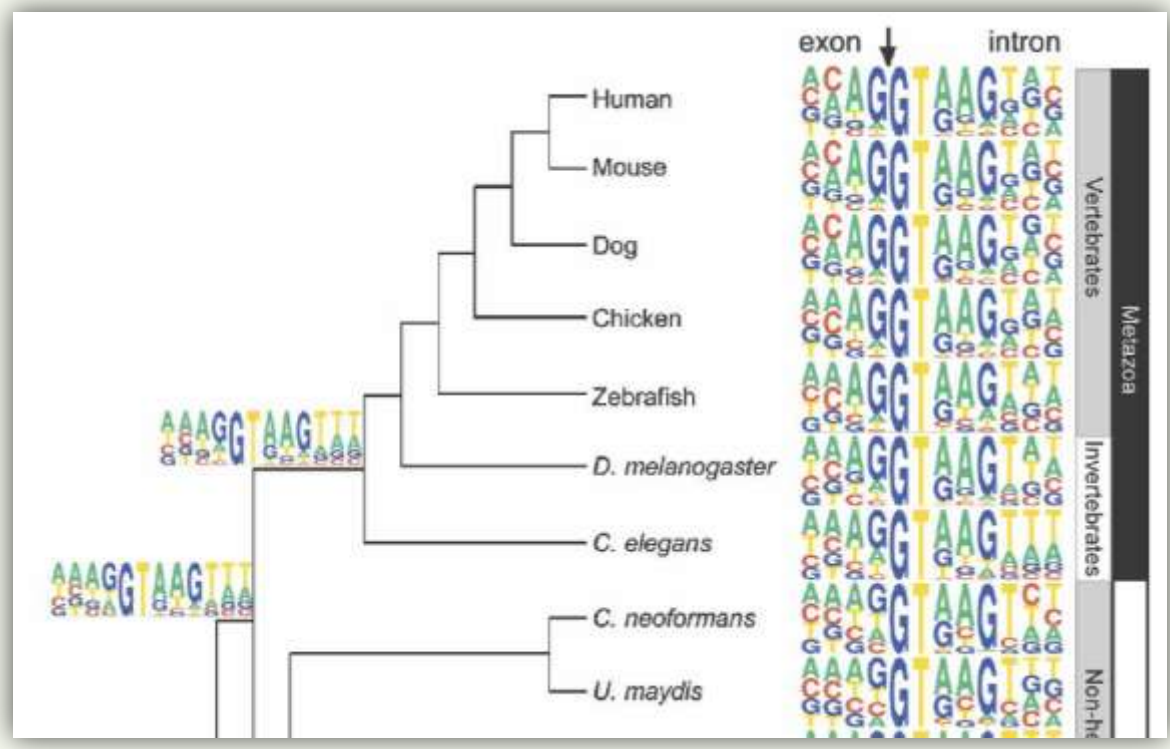
peptide query (aa) >>>>> **tblastn** >>>>> genome assembly (nt)

5. ‘Difficult’ cases

A. Divergent genome

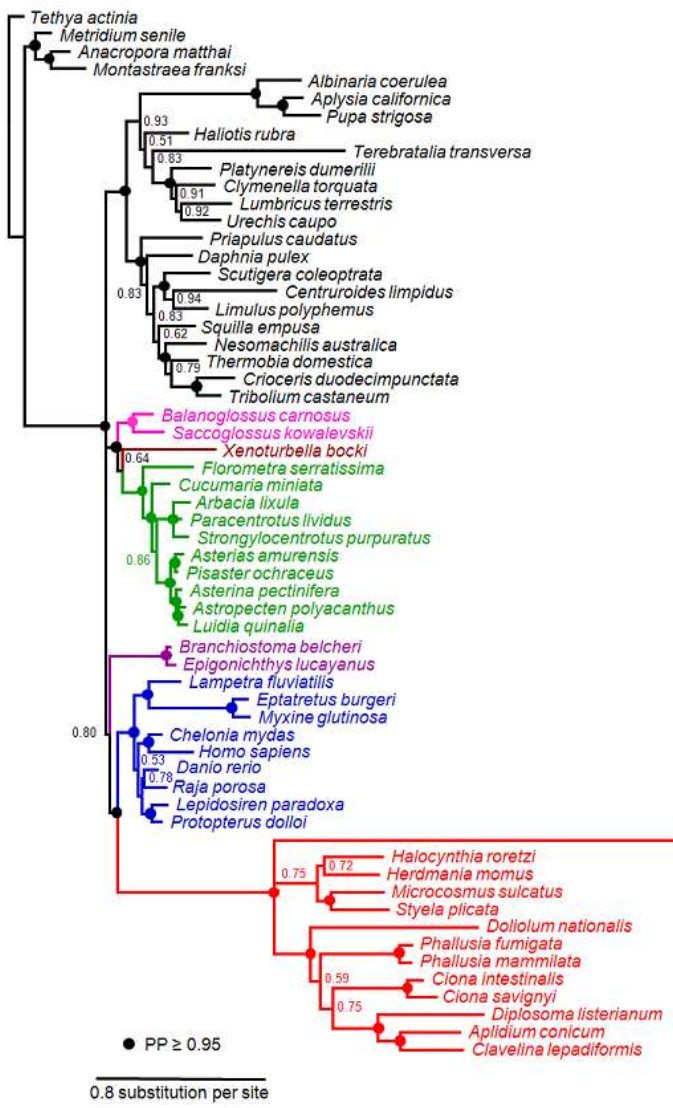
- What if the ‘typical’ gene structure is different ?
  e.g., frequent non-canonical splice sites

Typical problem
in *ab initio* algorithms
(=> false-negative)



Schwartz et al., 2008. Genome Res. 18:88-

A. Divergent genome



Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate

France Denoeud,[1,2,3] Simon Henriet,[4]* Sutada Mungpakdee,[4]* Jean-Marc Aury,[1,2,3]* Corinne Da Silva,[1,2,3]* Henner Brinkmann,[5] Jana Mikhaleva,[4] Lisbeth Charlotte Olsen,[4] Claire Jubin,[1,2,3] Cristian Cañestro,[6,24] Jean-Marie Bouquet,[4] Gemma Danks,[4,7] Julie Poulain,[1,2,3] Coen Campsteijn,[4] Marcin Adamski,[4] Ismael Cross,[8] Fekadu Yadetie,[4] Matthieu Muffato,[9] Alexandra Louis,[9] Stephen Butcher,[10] Georgia Tsagkogeorga,[11] Anke Konrad,[22] Sarabdeep Singh,[12] Marit Flo Jensen,[4] Evelyne Huynh Cong,[4] Helen Eikeseth-Otteraa,[4] Benjamin Noel,[1,2,3] Véronique Anthouard,[1,2,3] Betina M. Porcel,[1,2,3] Rym Kachouri-Lafond,[13] Atsuo Nishino,[14] Matteo Ugolini,[4] Pascal Chourrout,[15] Hiroki Nishida,[14] Rein Aasland,[16] Snehalata Huzurbazar,[12] Eric Westhof,[13] Frédéric Delsuc,[11] Hans Lehrach,[17] Richard Reinhardt,[17] Jean Weissenbach,[1,2,3] Scott W. Roy,[18] François Artiguenave,[1,2,3] John H. Postlethwait,[6] J. Robert Manak,[10] Eric M. Thompson,[4,19] Olivier Jaillon,[1,2,3] Louis Du Pasquier,[20] Pierre Boudinot,[21] David A. Liberles,[22] Jean-Nicolas Volff,[23] Hervé Philippe,[5] Boris Lenhard,[4,7,19] Hugues Roest Crollius,[9] Patrick Wincker,[1,2,3]† Daniel Chourrout[4]†

www.sciencemag.org SCIENCE VOL 330 3 DECEMBER 2010

Oikopleura dioica
Wikipedia

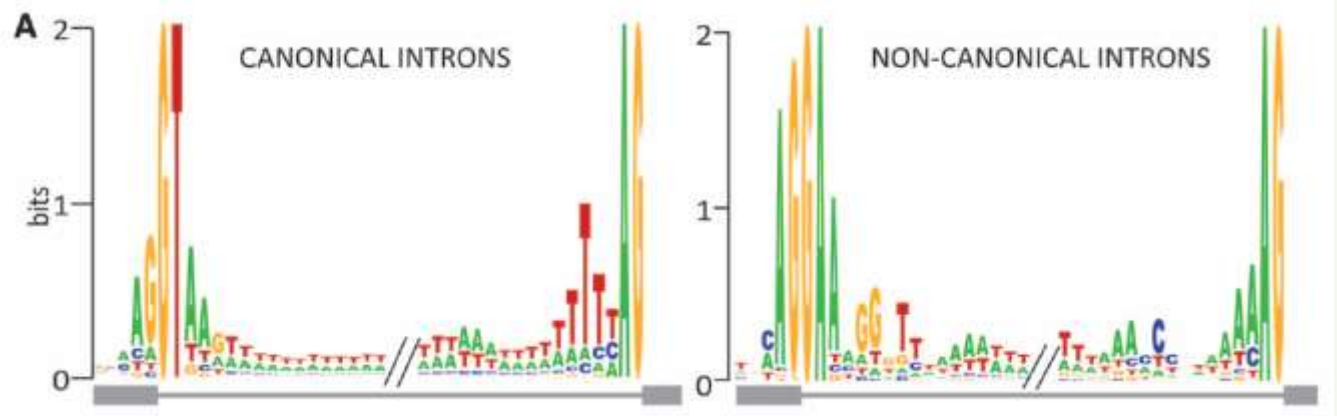
Oikopleura dioica

"Oikopleura has a rather common number of introns per gene (4.1), but the turnover of its introns has been extraordinarily high: Of 5589 introns mapped by interspecies protein alignments, ***76% had positions unique to Oikopleura*** *(newly acquired introns), 17% were at ancestral positions (old introns), and 7% could not be classified (fig. S21) (3). Noncanonical introns, mostly* ***GA-AG*** *and with a very specific acceptor site, are unusually frequent (****12%****) (Fig. 2A and figs. S22 to S25) (3). They show several peculiarities (tables S11 and S12), including preferential insertion in phase 1, which is compatible with the current codon usage, as would be expected for the most recently gained introns (3, 7). The most distinctive feature of newly acquired introns (figs. S26 and S27 and tables S13 to S15) is that they are* ***more often noncanonical than old introns*** *(8.4 versus 2.6%) (3)."*

Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate

France Denoeud,[1,2,3] Simon Henriet,[4]* Sutada Mungpakdee,[4]* Jean-Marc Aury,[1,2,3]* Corinne Da Silva,[1,2,3]* Henner Brinkmann,[5] Jana Mikhaleva,[4] Lisbeth Charlotte Olsen,[4] Claire Jubin,[1,2,3] Cristian Cañestro,[6,24] Jean-Marie Bouquet,[4] Gemma Danks,[4,7] Julie Poulain,[1,2,3] Coen Campsteijn,[4] Marcin Adamski,[4] Ismael Cross,[8] Fekadu Yadetie,[4] Matthieu Muffato,[9] Alexandra Louis,[9] Stephen Butcher,[10] Georgia Tsagkogeorga,[11] Anke Konrad,[22] Sarabdeep Singh,[12] Marit Flo Jensen,[4] Evelyne Huynh Cong,[4] Helen Eikeseth-Otteraa,[4] Benjamin Noel,[1,2,3] Véronique Anthouard,[1,2,3] Betina M. Porcel,[1,2,3] Rym Kachouri-Lafond,[13] Atsuo Nishino,[14] Matteo Ugolini,[4] Pascal Chourrout,[15] Hiroki Nishida,[14] Rein Aasland,[16] Snehalata Huzurbazar,[12] Eric Westhof,[13] Frédéric Delsuc,[11] Hans Lehrach,[17] Richard Reinhardt,[17] Jean Weissenbach,[1,2,3] Scott W. Roy,[18] François Artiguenave,[1,2,3] John H. Postlethwait,[6] J. Robert Manak,[10] Eric M. Thompson,[4,19] Olivier Jaillon,[1,2,3] Louis Du Pasquier,[20] Pierre Boudinot,[21] David A. Liberles,[22] Jean-Nicolas Volff,[23] Hervé Philippe,[5] Boris Lenhard,[4,7,19] Hugues Roest Crollius,[9] Patrick Wincker,[1,2,3]† Daniel Chourrout[4]†

www.sciencemag.org SCIENCE VOL 330 3 DECEMBER 2010



Training should help adapt to species-specific genic characters

5. ‘Difficult’ cases

Considering gene size …

Nat. Rev. Genet. 2003. 4:741

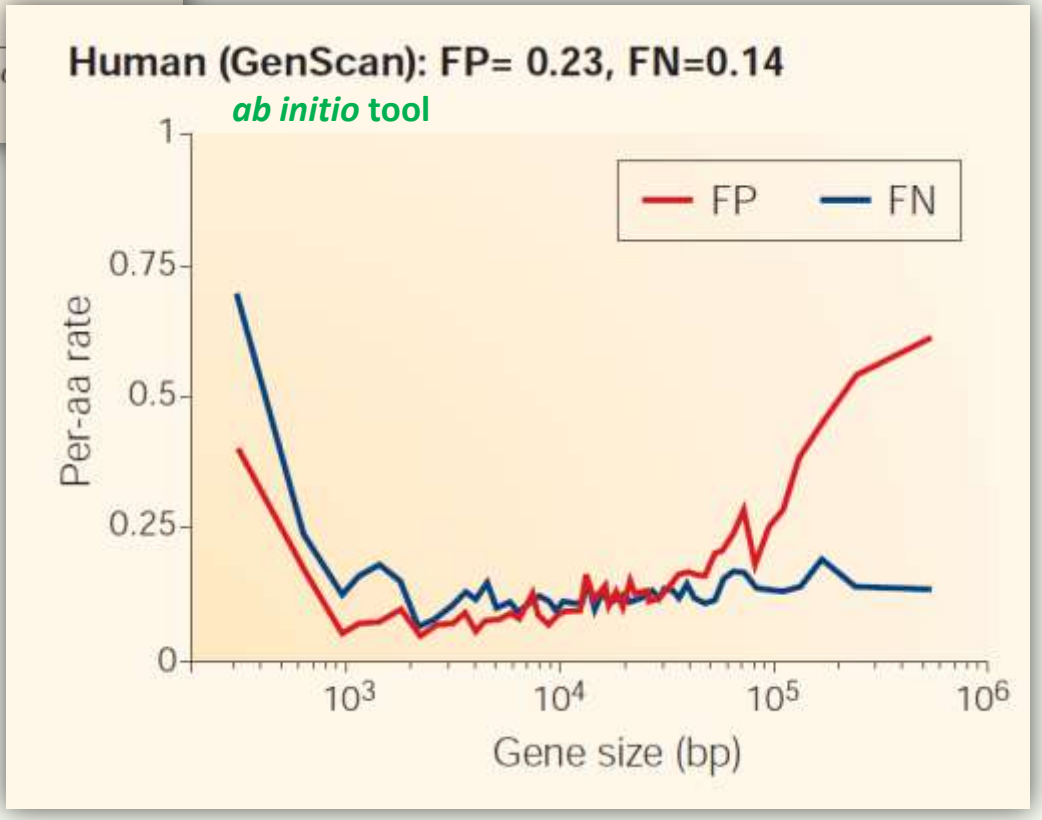
OPINION

Vertebrate gene predictions and the problem of large genes

Jun Wang, ShengTing Li, Yong Zhang, HongKun Zheng, Zha
Jun Yu and Gane Ka-Shu Wong

*“As **gene size increases**, **FP** (false-positive) rate goes up,but **FN** (false-negative) rate does not.”*

*“**Single-exon genes** that are smaller than 1 kb present a different problem, as **FP** and **FN** rate both increase within the limits of small genes. It might be thought that these would be the easiest genes to predict ……. . In fact, small genes are intrinsically difficult to detect, partly because of the lack of splicing signals on either side of the single exon, but mostly because of the decreasing signal-to-noise ratios as the size of the coding region decreases.”*



5. ‘Difficult’ cases

B. Large genes

Nat. Rev. Genet. 2003. 4:741

OPINION

Vertebrate gene predictions and the problem of large genes

Jun Wang, ShengTing Li, Yong Zhang, HongKun Zheng, Zhao Xu, Jia Ye, Jun Yu and Gane Ka-Shu Wong

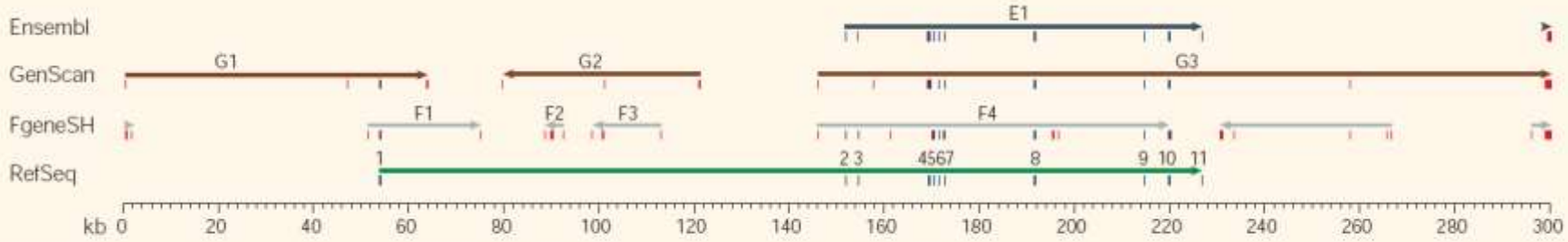
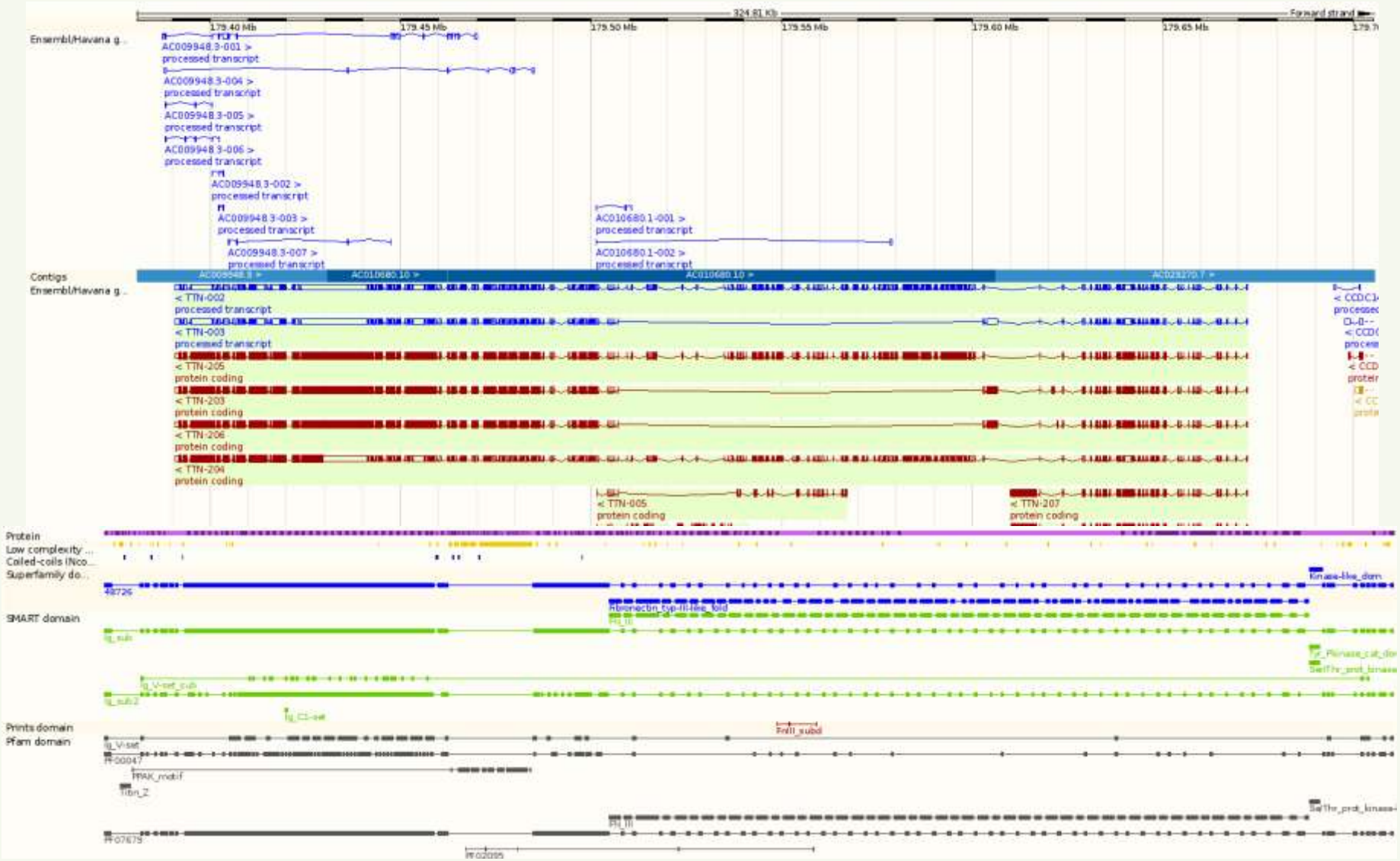


Figure 1 | **Actual versus predicted exons in a known gene: TEA domain family member 1 (SV40 transcriptional enhancer factor on human chromosome 11).** Correctly predicted exons are coloured blue, whereas incorrectly predicted exons are coloured red. Arrows indicate the direction of transcription for each predicted gene. RefSeq indicates that there are 11 exons, labelled 1–11, which span a genomic region of 173 kb. FgeneSH has four predictions that overlap with this gene, labelled F1–F4. GenScan has three predictions that overlap with this gene, labelled G1–G3. Note that F2, F3 and G2 are false-positive exons, in the large 98 kb first intron of this gene. OVER-PREDICTION is another problem, as exemplified by G1 and G3, which did not terminate correctly at the start and stop codons. Most of these problems are fixed in the Ensembl prediction, labelled E1, but even so, it failed to identify the first exon. Approximately half of this genomic region is incorrectly annotated as a ‘gene desert’, because of one large intron.

Large gene: Titin

Molecular weight: 3,713,426.90
Number of residues: 33,421
# of Exons : 313



5. 'Difficult' cases

C. Short genes

- 11 aa-long fruitfly TAL (tarsal-less) gene involved in development (Galindo et al., 2007)

Method

Discovery and annotation of small proteins using genomics, proteomics, and computational approaches

Xiaohan Yang,[1,2,6] Timothy J. Tschaplinski,[1,2] Gregory B. Hurst,[3] Sara Jawdy,[1,2] Paul E. Abraham,[2,4] Patricia K. Lankford,[1] Rachel M. Adams,[2,4] Manesh B. Shah,[1] Robert L. Hettich,[2,3] Erika Lindquist,[5] Udaya C. Kalluri,[1,2] Lee E. Gunter,[1,2] Christa Pennacchio,[5] and Gerald A. Tuskan[1,2,5,6]

[1]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA;* [2]*BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA;* [3]*Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA;* [4]*Graduate School of Genome Science and Technology, University of Tennessee–Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA;* [5]*DOE Joint Genome Institute, Walnut Creek, California 94598, USA*

Genome Res. 2011. 21: 634-

Poplus deltoides
Wikipedia

"Ab initio discovery of small proteins has been relatively overlooked"
"an arbitrary minimum ORF cutoff (e.g., 100aa) is applied in gene annotation algorithms …."

- 11 aa-long fruitfly TAL (tarsal-l

Method

Discovery and annotation of genomics, proteomics, and c

Xiaohan Yang,[1,2,6] Timothy J. Tschaplinski,[1,2] Paul E. Abraham,[2,4] Patricia K. Lankford,[1] Rach Robert L. Hettich,[2,3] Erika Lindquist,[5] Udaya C Christa Pennacchio,[5] and Gerald A. Tuskan[1,2,

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Ter National Laboratory, Oak Ridge, Tennessee 37831, USA; [3]Chemical S Tennessee 37831, USA; [4]Graduate School of Genome Science and Te Laboratory, Oak Ridge, Tennessee 37830, USA; [5]DOE Joint Genome I

Genome Res. 2011. 21: 634-

"Ab initio discovery of small
"an arbitrary minimum ORF
algorithms …."

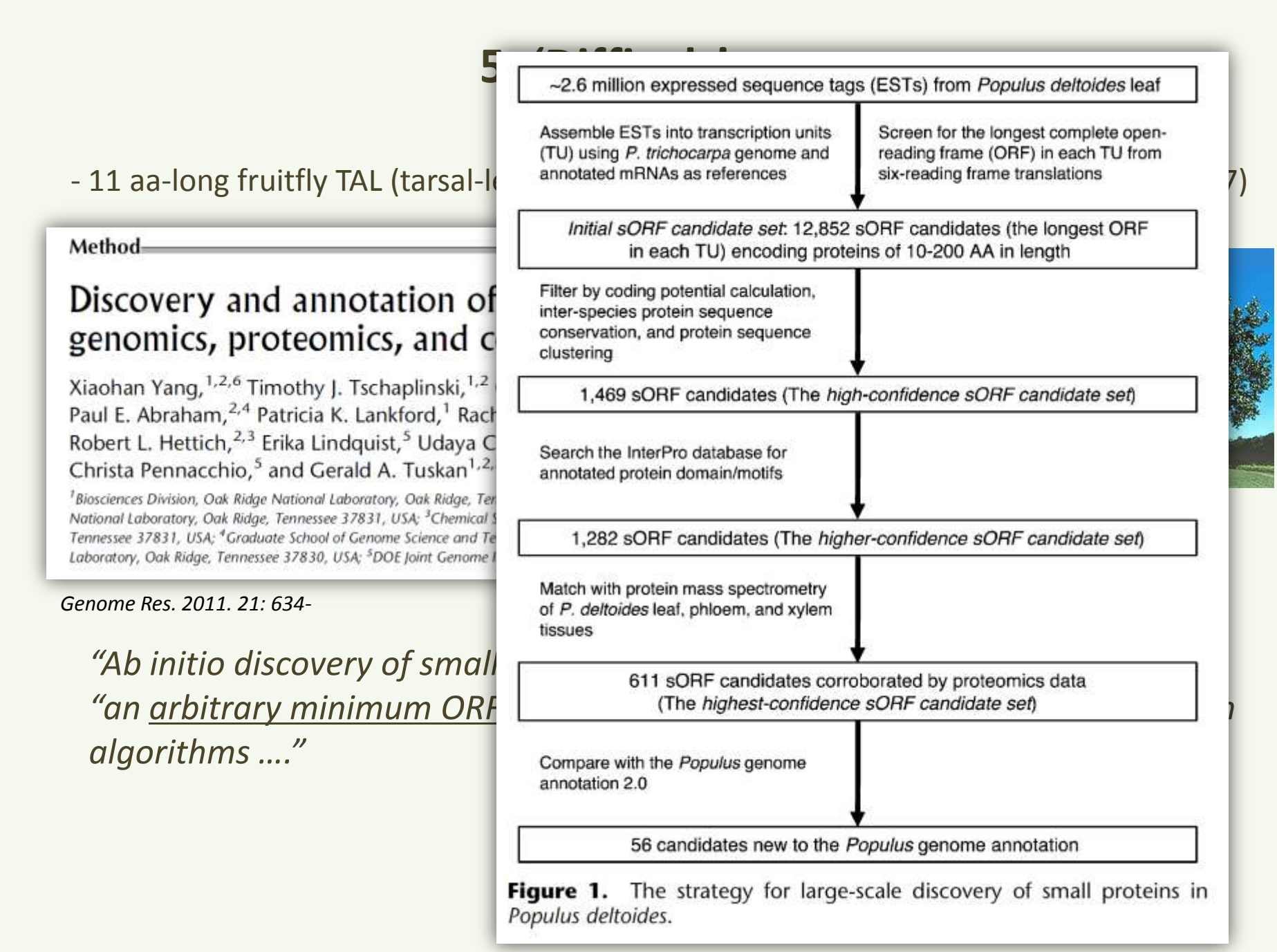


Figure 1. The strategy for large-scale discovery of small proteins in *Populus deltoides*.

5. 'Difficult' cases

C. Short genes

- 11 aa-long fruitfly TAL (tarsal-less) gene involved in development (Galindo et al., 2007)

Method

Discovery and annotation of small proteins using genomics, proteomics, and computational approaches

Xiaohan Yang,[1,2,6] Timothy J. Tschaplinski,[1,2] Gregory B. Hurst,[3] Sara Jawdy,[1,2] Paul E. Abraham,[2,4] Patricia K. Lankford,[1] Rachel M. Adams,[2,4] Manesh B. Shah,[1] Robert L. Hettich,[2,3] Erika Lindquist,[5] Udaya C. Kalluri,[1,2] Lee E. Gunter,[1,2] Christa Pennacchio,[5] and Gerald A. Tuskan[1,2,5,6]

[1]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA;* [2]*BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA;* [3]*Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA;* [4]*Graduate School of Genome Science and Technology, University of Tennessee–Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA;* [5]*DOE Joint Genome Institute, Walnut Creek, California 94598, USA*

Genome Res. 2011. 21: 634-

Poplus deltoides
Wikipedia

"Ab initio discovery of small proteins has been relatively overlooked"
"an arbitrary minimum ORF cutoff (e.g., 100aa) is applied in gene annotation algorithms …."

Results: 1282 genes (encoding proteins of **10-200 aa** in length)

Out of 1282**, 611** genes supported by proteomics data

Out of 611, **56** were new to the current genome annotation

5. ‘Difficult’ cases

D. Species (or lineage) specific genes

(‘orphan genes’, ‘new genes’)


Review

More than just orphans: are taxonomically-restricted genes important in evolution?

Konstantin Khalturin, Georg Hemmrich, Sebastian Fraune, René Augustin and Thomas C.G. Bosch

Zoological Institute, Christian-Albrechts-University Kiel, Olshausenstrasse 40, 24098 Kiel, Germany

Trends Genet. 2009. 25: 404-

Examples)

Experimentally identified genes in **hydra** (Milde et al., ’09; Khalturin et al., ’08)

1761 genes specific to **Brassicaceae** identified *in silico* (Donoghue et al., 2011)

One **mouse**-specific gene involved in spermatogenesis (Heinen et al., 2009)

Typical problem in cross-species information-based algorithms (=> false-negative)

5. ‘Difficult’ cases

D. Others

Genes with:

- Non-canonical splice sites

- Selenocysteine codons

- Regulated frame-shifting

- Atypical codon usage (‘new genes’)

- Elevated evolutionary rate

Non-protein-coding transcripts

6. Challenges related to NGS

Application of NGS technology => low-quality assembly (Alkan et al., 2011; Birney, 2011)

Challenge: Detection of incomplete ORFs

(Missing repeats is not an issue?)

Clue 1: Extrinsic **scaffolding** of split genes
based on ESTs, peptides and cross-species comparison (e.g., *ESPRIT*)

Clue 2: Tools for **metagenomics** (where assembly is 'forbidden'*!!*)

Nucleic Acids Research, 2009, 1–5
doi:10.1093/nar/gkp327

Orphelia: predicting genes in metagenomic sequencing reads

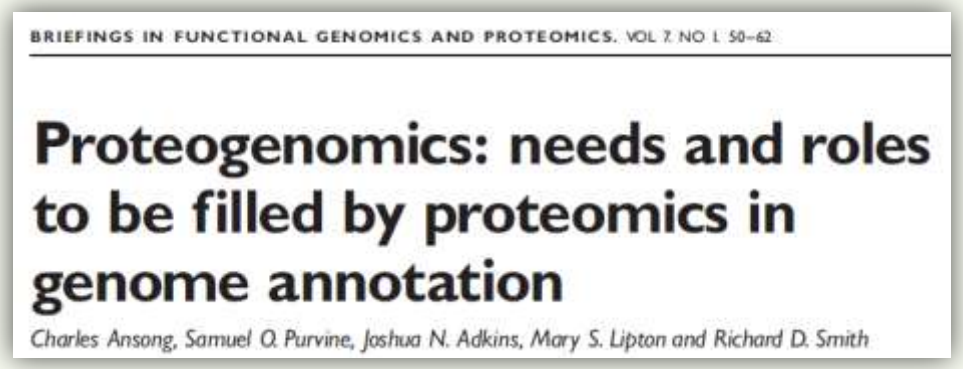
Katharina J. Hoff[1,*], Thomas Lingner[1,2], Peter Meinicke[1] and Maike Tech[1]

[1]Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany and [2]Center for Genomic Regulation, Comparative Bioinformatics Research Group, Biomedical Research Park, c/Dr. Aiguader 88, 08003 Barcelona, Spain

*"**Orphelia** is a program for predicting genes in **short DNA sequences** (<300bp) that is available through a web server application (http://orphelia.gobics.de)."*

7. Quality assessment

Really transcribed? Really translated?



BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 7. NO 1. 50–62

Proteogenomics: needs and roles to be filled by proteomics in genome annotation

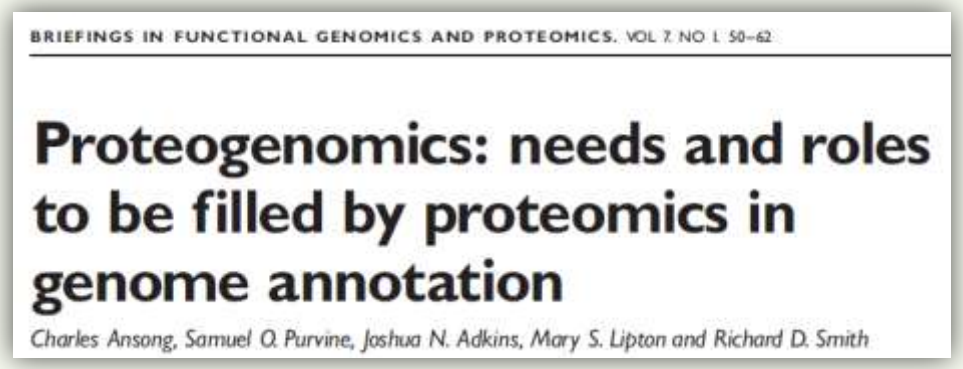
Charles Ansong, Samuel O. Purvine, Joshua N. Adkins, Mary S. Lipton and Richard D. Smith

Peptide sequencing
using Tandem Mass Spectrometry (MS/MS)

Identifies ‘novel’ and ‘correct’ proteins

7. Quality assessment

Really transcribed? Really translated?



BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 7. NO 1. 50–62

Proteogenomics: needs and roles to be filled by proteomics in genome annotation

Charles Ansong, Samuel O. Purvine, Joshua N. Adkins, Mary S. Lipton and Richard D. Smith

Peptide sequencing
using Tandem Mass Spectrometry (MS/MS)

Identifies 'novel' and 'correct' proteins

In **human**, translation of
224 hypothetical proteins confirmed

Methods

Improving gene annotation using peptide mass spectrometry

Stephen Tanner,[1,6] Zhouxin Shen,[2] Julio Ng,[1] Liliana Florea,[3] Roderic Guigó,[4] Steven P. Briggs,[2] and Vineet Bafna[5]

[1]Bioinformatics Program, University of California, San Diego, La Jolla, California 92093-0419, USA; [2]Department of Biology, University of California, San Diego, La Jolla, California 92093-0346, USA; [3]Department of Computer Science, George Washington University, Washington, DC 20052, USA; [4]Centre de Regualció Genòmica, 08003 Barcelona, Spain; [5]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093-0404, USA

Genome Res. 2007. 17:231-

In ***Arabidopsis***, translation of
778 hypothetical proteins confirmed

Discovery and revision of *Arabidopsis* genes by proteogenomics

Natalie E. Castellana[a,1], Samuel H. Payne[b,1], Zhouxin Shen[c,1], Mario Stanke[d], Vineet Bafna[a,2], and Steven P. Briggs[c,2]

[a]Department of Computer Science and Engineering, [b]Bioinformatics Program, [c]Division of Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093; and [d]Institute for Microbiology and Genetics, Goldschmidtstrasse 1, 37077 Göttingen, Germany

21034–21038 | PNAS | December 30, 2008 | vol. 105 | no. 52

8. Perspectives

- Cases when training does not help
- Taxon-specific gene predictors (based on taxon-specific training) ?
- Detecting incomplete ORFs (e.g., GenScan vs Augustus)
- Alternative splicing
- User-friendliness
- Promising solutions for ongoing genome projects

Aims

Accomplish the first review from the users' viewpoint

1. Cover all **existing review papers** on gene prediction
2. List **which gene predictors** annotated which genomes
3. Who are **current and potential users** of gene predictors?
4. Consider **typical research contexts** in need of gene prediction
5. Outline challenges based on **difficult cases**
6. Speculate challenges related to **next generation sequencing (NGS)**
7. Discuss how we can **assess quality** of gene prediction
8. Any other clue for improvement in quality and user-friendliness?