# Gene Finding

**BCH394P/374C Systems Biology / Bioinformatics**

**Edward Marcotte, Univ of Texas at Austin**

The Telegraph

Home | News | World | Sport | Finance | Comment | Culture | Travel | Life | Women | Fashion

Politics | Investigations | Obits | Education | Earth | Science | Defence | Health | Scotland | Roya

Science News | Space | Night Sky | Roger Highfield | Dinosaurs | Evolution | Steve Jones | Scien

HOME » SCIENCE » SCIENCE NEWS

World's largest genome belongs to slow-growing mountain flower

An unremarkable and slow-growing plant has stunned scientists after they found it had the world's largest genome – 50 times bigger than that of our own species.

The DNA contained within Paris japonica dwarves all other plant and animal genomes that have been analysed so far Photo: CLIVE NICHOLS

Print this article

Share 304

Facebook 248

Twitter 56

Email

LinkedIn 0

g+1 0

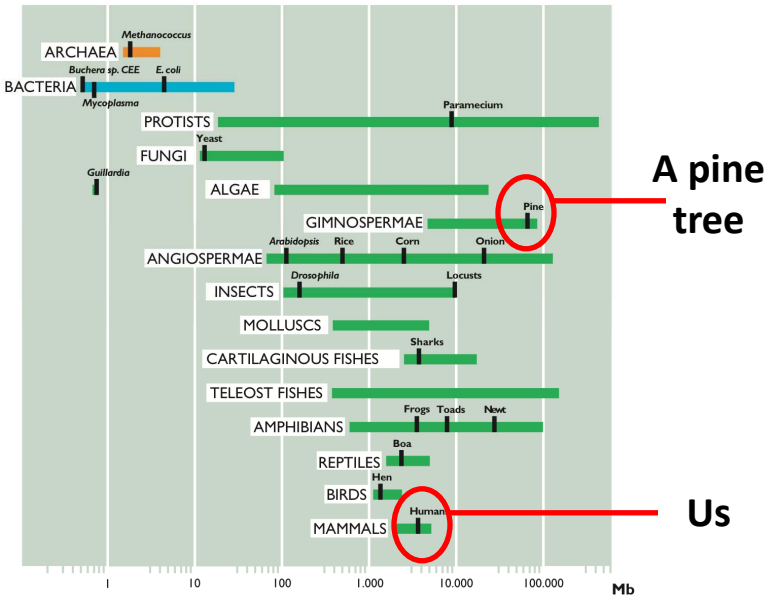
Science News

News » UK News »

Science »

Earth News »

# Genome size ranges vary widely across organisms

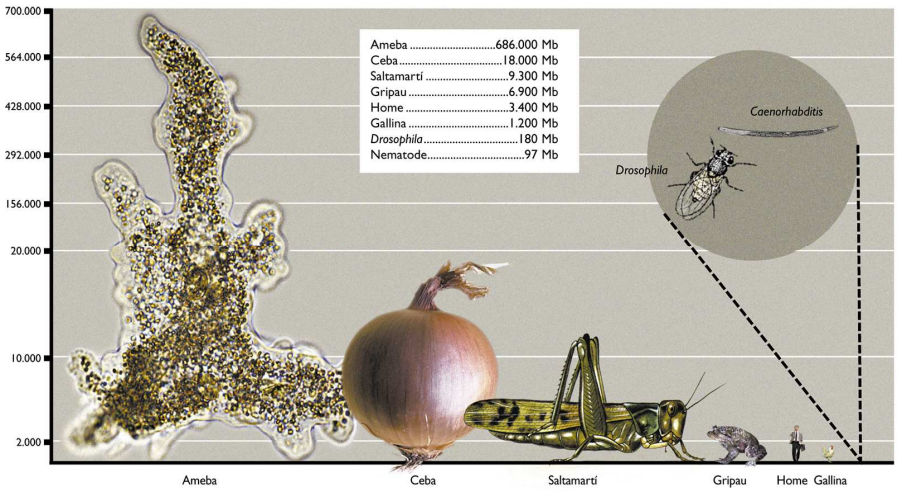



A pine tree

Us

https://metode.org/issues/monographs/the-size-of-the-genome-and-the-complexity-of-living-beings.html

# Genome size ranges vary widely across organisms




Here, the height (i.e. vertical axis, not area) indicates genome size

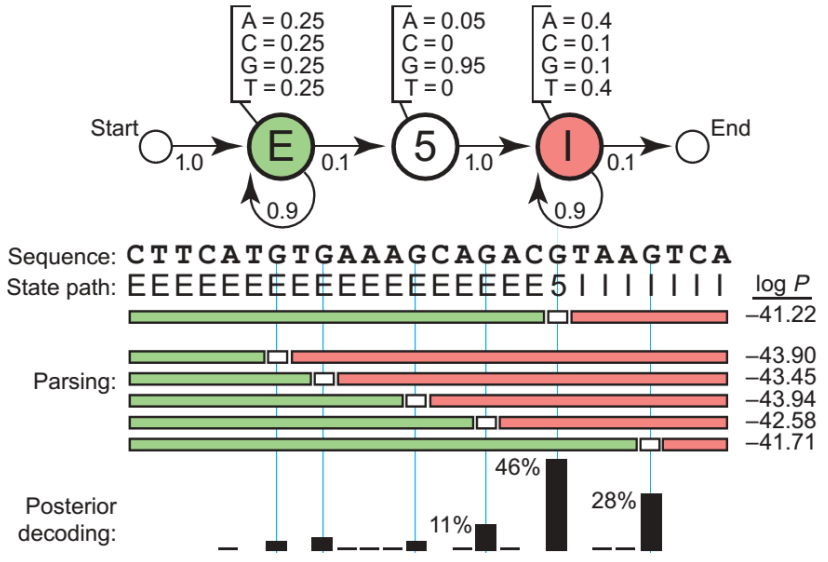
https://metode.org/issues/monographs/the-size-of-the-genome-and-the-complexity-of-living-beings.html

## Where are the genes? How can we find them?

GATCACTTGATAAATGGGCTGAAGTAACTCGCCCAGATGAGGAGTGTGCTGCCTCCAGAAT
CCAAACAGGCCCACTAGGCCCGAGACACCTTGTCTCAGATGAAACTTTGGACTCGGAATT
TTGAGTTAATGCCGGAATGAGTTCAGACTTTGGGGGACTGTTGGGAAGGCATGATTGGTT
TCAAAATGTGAGAAGGACATGAGATTTGGGAGGGGCTGGGGGCAGAATGATATAGTTTG
GCTCTGCGTCCCCACCCAATCTCATGTCAAATTGTAATCCTCATGTGTCAGGGGAGAGGCCT
GGTGGGATGTGATTGGATCATGGGAGTGGATTTCCCTCTTGCAGTTCTCGTGATAGTGAGT
GAGTTCTCACGAGATCTGGTTGTTTGAAAGTGTGCAGCTCCTCCCCCTTCGCGCTCTCTCTC
TCCCCTGCTCCACCATGGTGAGACGTGCTTGCGTCCCCTTTGCCTTCTGCCATGATTGTAAG
CTTCCTCAGGCGTCCTAGCCACGCTTCCTGTACAGCCTGAGGAACTGGGAGTCAATGAAA
CCTCTTCTCTTCATAAATTACCCAGTTTCAGGTAGTTCTTTCTAGCAGTGTGATAATGGACGA
TACAAGTAGAGACTGAGATCAATAGCATTTGCACTGGGCCTGGAACACACTGTTAAGAAC
GTAAGAGCTATTGCTGTCATTAGTAATATTCTGTATTATTGGCAACATCATCACAATACACTGC
TGTGGGAGGGTCTGAGATACTTCTTTGCAGACTCCAATATTTGTCAAAACATAAAATCAGG
AGCCTCATGAATAGTGTTTAAATTTTTACATAATAATACATTGCACCATTTGGTATATGAGTCT
TTTTGAAATGGTATATGCAGGACGGTTTCCTAATATACAGAATCAGGTACACCTCCTCTTCCA
TCAGTGCGTGAGTGTGAGGGATTGAATTCCTCTGGTTAGGAGTTAGCTGGCTGGGGGTTC
TACTGCTGTTGTTACCCACAGTGCACCTCAGACTCACGTTTCTCCAGCAATGAGCTCCTGTT
CCCTGCACTTAGAGAAGTCAGCCCGGGGACCAGACGGTTCTCTCCTCTTGCCTGCTCCAG
CCTTGGCCTTCAGCAGTCTGGATGCCTATGACACAGAGGGCATCCTCCCCAAGCCCTGGTC
CTTCTGTGAGTGGTGAGTTGCTGTTAATCCAAAAGGACAGGTGAAAACATGAAAGCC...

## A toy HMM for 5′ splice site recognition (from linked on the course web page)

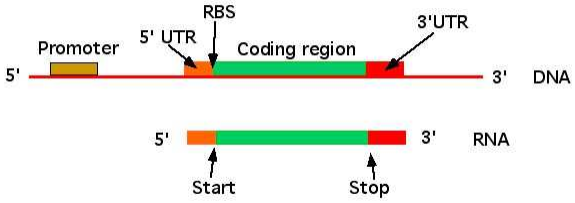
**Remember this?**



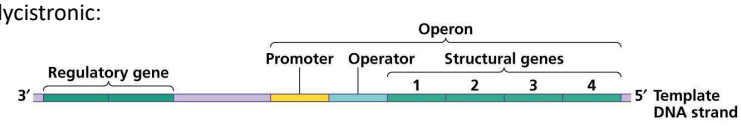
# Let's start with prokaryotic genes

**What elements should we build into an HMM to find bacterial genes?**

# Let's start with prokaryotic genes




Can be polycistronic:



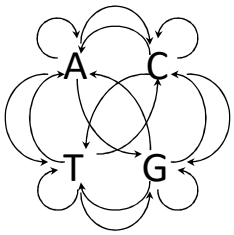



http://nitro.biosci.arizona.edu/courses/EEB600A-2003/lectures/lecture24/lecture24.html

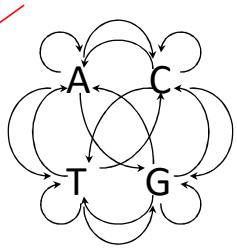
***A CpG island model might look like:***

**Remember this?**

( of course, need the parameters, but maybe these are the most important….)




**p(C→G) is higher**



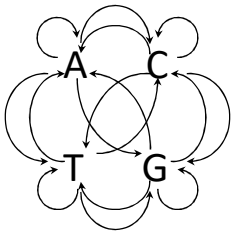

**p(C→G) is lower**

CpG island model

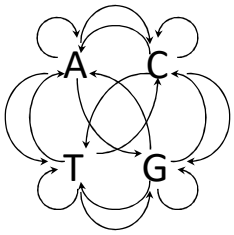
Not CpG island model

Could calculate $\frac{P(X \mid \text{CpG island})}{P(X \mid \text{not CpG island})}$ (or log ratio) along a sliding window, just like the fair/biased coin test

***One way to build a minimal gene finding Markov model***




**Transition probabilities reflect codons**




**Transition probabilities reflect intergenic DNA**

Coding DNA model

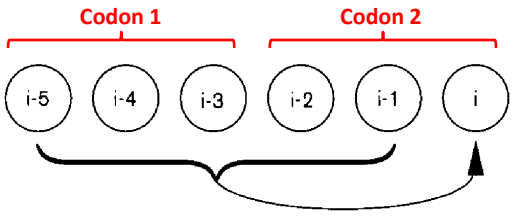
Intergenic DNA model

Could calculate $\frac{P(X \mid \text{coding})}{P(X \mid \text{not coding})}$ (or log ratio) along a sliding window, just like the fair/biased coin test

**Really, we'll want to detect codons.**
**The usual trick is to use a *higher-order Markov process*.**

**A standard Markov process only considers the current position in calculating transition probabilities.**

**An *n*$^{th}$*-order Markov process* takes into account the past *n* nucleotides, e.g. as for a 5$^{th}$ order:**



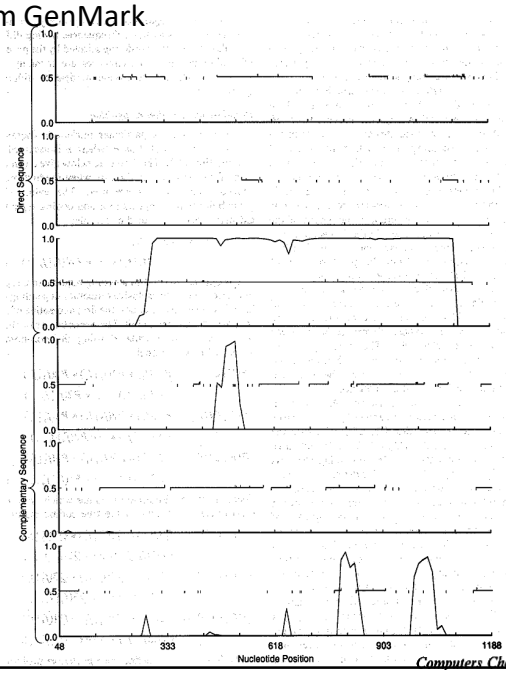

But we need to learn $4^{(n+1)}$ transition probabilities!
That's 4096 entries for a 5th-order model.

Image from *Curr Op Struct Biol* 8:346-354 (1998)

5$^{th}$ order Markov chain, using models of coding vs. non-coding using the classic algorithm GenMark

Direct strand

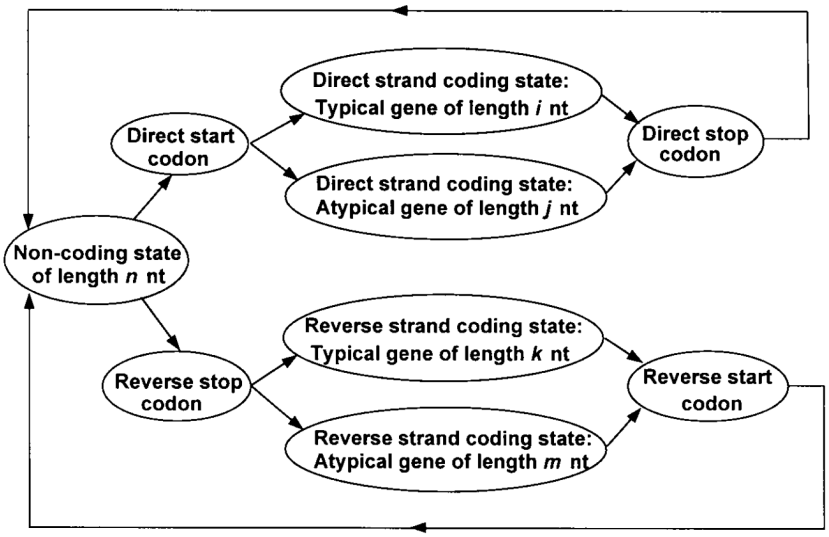
Complementary (reverse) strand




1$^{st}$ reading frame
2$^{nd}$ reading frame
3$^{rd}$ reading frame
1$^{st}$ reading frame
2$^{nd}$ reading frame
3$^{rd}$ reading frame

*Computers Chem.* Vol. 17, No. 2, pp. 123–133, 1993

An HMM version of GenMark




GeneMark.hmm: new solutions for gene finding
Alexander V. Lukashin and Mark Borodovsky[1,*]
*Nucleic Acids Research, 1998, Vol. 26, No. 4* **1107–1115**

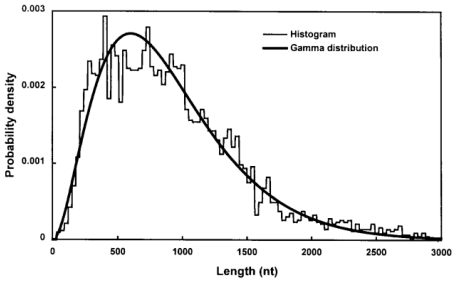
For example, accounting for variation in start codons…

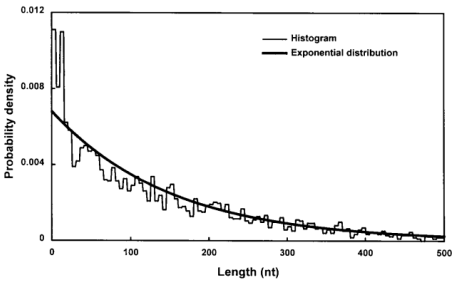
The probabilities of the start codons were defined in agreement with the *E.coli* genome statistics: $P(\text{ATG}) = 0.905$, $P(\text{GTG}) = 0.090$, $P(\text{TTG}) = 0.005$. The probability of transition from a non-coding state to a Typical (Atypical) coding state was set to 0.85 (0.15).

GeneMark.hmm: new solutions for gene finding
Alexander V. Lukashin and Mark Borodovsky[1,*]
*Nucleic Acids Research, 1998, Vol. 26, No. 4* **1107–1115**

… and variation in gene lengths

## Length distributions (in # of nucleotides)

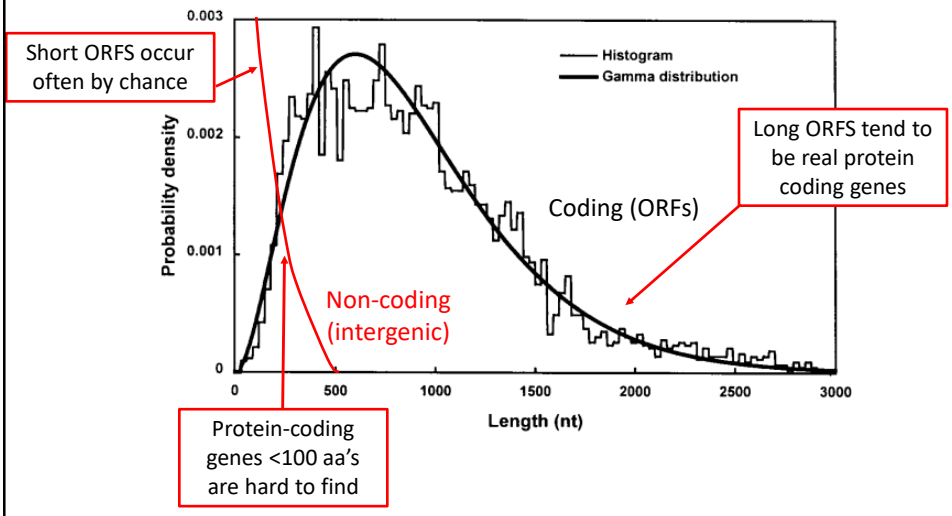



Coding (ORFs)  Non-coding (intergenic)

**GeneMark.hmm: new solutions for gene finding**
Alexander V. Lukashin and Mark Borodovsky[1,*]
*Nucleic Acids Research, 1998, Vol. 26, No. 4* **1107–1115**

(Placing these curves on top of each other)




**GeneMark.hmm: new solutions for gene finding**
Alexander V. Lukashin and Mark Borodovsky[1,*]
*Nucleic Acids Research, 1998, Vol. 26, No. 4* **1107–1115**

## Model for a ribosome binding site (based on ~300 known RBS's)

| Nucleotide | Position 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| T | 0.161 | 0.050 | 0.012 | 0.071 | 0.115 |
| C | 0.077 | 0.037 | 0.012 | 0.025 | 0.046 |
| A | **0.681** | 0.105 | 0.015 | **0.861** | 0.164 |
| G | 0.077 | **0.808** | **0.960** | 0.043 | **0.659** |

**GeneMark.hmm: new solutions for gene finding**

Alexander V. Lukashin and Mark Borodovsky[1,*]

*Nucleic Acids Research, 1998, Vol. 26, No. 4* ***1107–1115***

## How well does it do on well-characterized genomes?

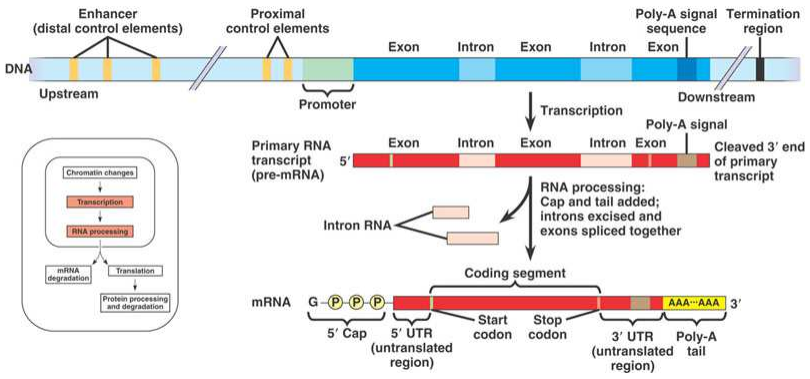
| Genome | Genes annotated | Genes predicted | Exact prediction (%) | Missing genes (%) | Wrong genes (%) |
|---|---|---|---|---|---|
| *A.fulgidus* | 2407 | 2530 | 73.1 | 10.8 (2.0) | 15.1 |
| *B.subtilis* | 4101 | 4384 | 77.5 | 3.6 (2.8) | 9.8 |
| *E.coli* | 4288 | 4440 | 75.4 | 5.0 (2.7) | 8.2 |
| *H.influenzae* | 1718 | 1840 | 86.7 | 3.8 (3.2) | 10.2 |
| *H.pylori* | 1566 | 1612 | 79.7 | 6.0 (4.4) | 8.7 |
| *M.genitalium* | 467 | 509 | 78.4 | 9.9 (1.7) | 17.3 |
| *M.jannaschii* | 1680 | 1841 | 72.7 | 4.6 (0.8) | 12.9 |
| *M.pneumoniae* | 678 | 734 | 70.1 | 7.8 (4.1) | 13.6 |
| *M.thermoauthotrophicum* | 1869 | 1944 | 70.9 | 5.0 (3.5) | 8.6 |
| *Synechocystis* | 3169 | 3360 | 89.6 | 4.0 (1.5) | 9.4 |
| Averaged | 21 943 | 23 194 | 78.1 | 5.4 (2.7) | 10.4 |

But this was a long time ago!

# Eukaryotic genes

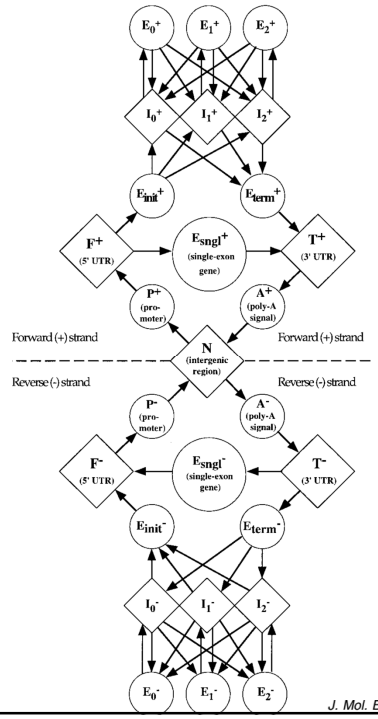
**What elements should we build into an HMM to find eukaryotic genes?**

# Eukaryotic genes




http://greatneck.k12.ny.us/GNPS/SHS/dept/science/krauz/bio_h/Biology_Handouts_Diagrams_Videos.htm

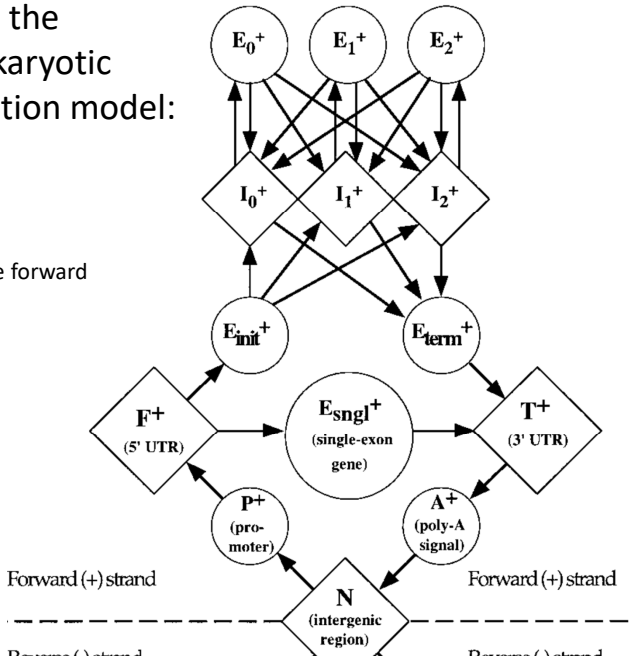
We'll look at the GenScan eukaryotic gene annotation model:




*J. Mol. Biol.* (1997) **268**, 78–94

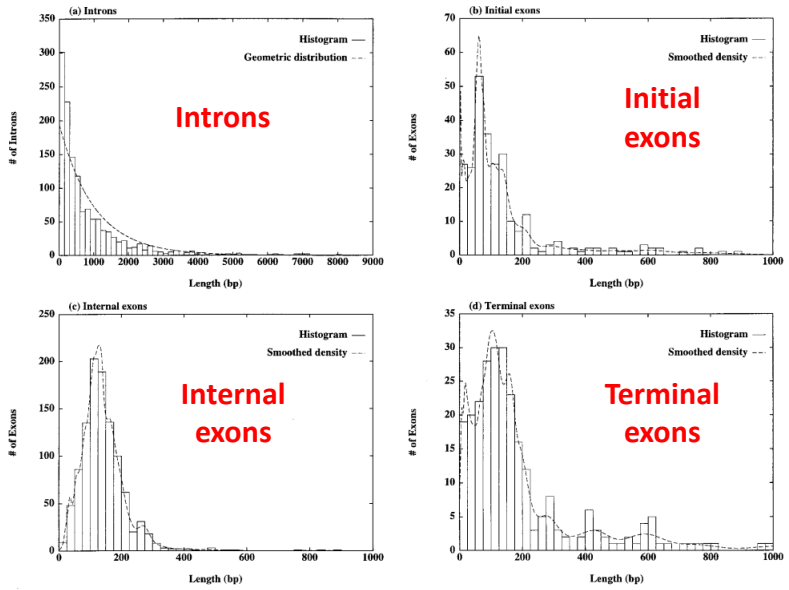
We'll look at the GenScan eukaryotic gene annotation model:

Zoomed in on the forward strand model…




*J. Mol. Biol.* (1997) **268**, 78–94

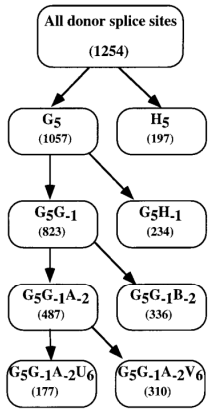
## Introns and different flavors of exons all have different typical lengths




*J. Mol. Biol.* (1997) **268**, 78–94

## Taking into account donor splice sites

| Pos | A% | C% | G% | U% |
|---|---|---|---|---|
| -3 | 33 | 36 | 19 | 13 |
| -2 | 56 | 15 | 15 | 15 |
| -1 | 9 | 4 | 78 | 9 |
| +3 | 44 | 3 | 51 | 3 |
| +4 | 75 | 4 | 13 | 9 |
| +6 | 14 | 18 | 19 | 49 |
| -3 | 34 | 37 | 18 | 11 |
| -2 | 59 | 10 | 15 | 16 |
| +3 | 40 | 4 | 53 | 3 |
| +4 | 70 | 4 | 16 | 10 |
| +6 | 17 | 21 | 21 | 42 |
| -3 | 37 | 42 | 18 | 3 |
| +3 | 39 | 5 | 51 | 5 |
| +4 | 62 | 5 | 22 | 11 |
| +6 | 19 | 20 | 25 | 36 |
| -3 | 32 | 40 | 23 | 5 |
| +3 | 27 | 4 | 59 | 10 |
| +4 | 51 | 5 | 25 | 19 |




| Pos | A% | C% | G% | U% |
|---|---|---|---|---|
| -3 | 35 | 44 | 16 | 6 |
| -2 | 85 | 4 | 7 | 5 |
| -1 | 2 | 1 | 97 | 0 |
| +3 | 81 | 3 | 15 | 2 |
| +4 | 51 | 28 | 9 | 12 |
| +6 | 22 | 20 | 30 | 28 |
| -3 | 29 | 31 | 21 | 18 |
| -2 | 43 | 30 | 17 | 11 |
| +3 | 56 | 0 | 43 | 0 |
| +4 | 93 | 2 | 3 | 3 |
| +6 | 5 | 10 | 10 | 76 |
| -3 | 29 | 30 | 18 | 23 |
| +3 | 42 | 1 | 56 | 1 |
| +4 | 80 | 4 | 8 | 8 |
| +6 | 14 | 21 | 16 | 49 |
| -3 | 39 | 43 | 15 | 2 |
| +3 | 46 | 6 | 46 | 3 |
| +4 | 69 | 5 | 20 | 7 |

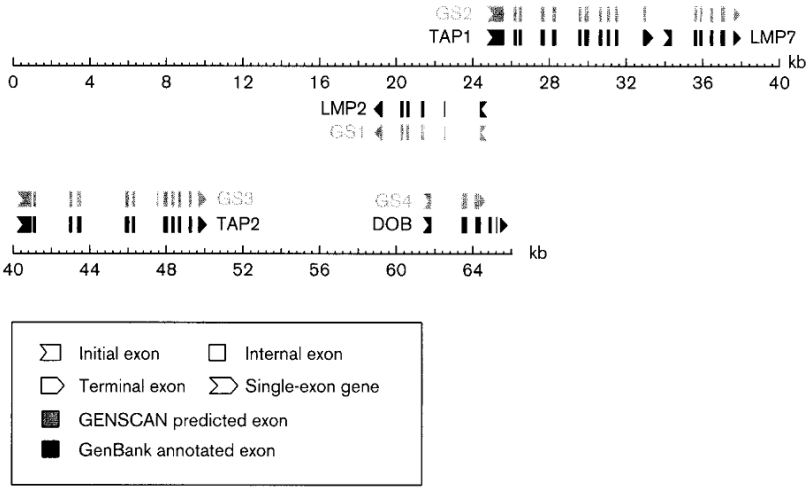
**All sites:** Position

| Base | -3 | -2 | -1 | +1 | +2 | +3 | +4 | +5 | +6 |
|---|---|---|---|---|---|---|---|---|---|
| A% | **33** | **60** | 8 | 0 | 0 | **49** | **71** | 6 | 15 |
| C% | **37** | 13 | 4 | 0 | 0 | 3 | 7 | 5 | 19 |
| G% | 18 | 14 | **81** | **100** | 0 | **45** | 12 | **84** | 20 |
| U% | 12 | 13 | 7 | 0 | **100** | 3 | 9 | 5 | **46** |

| U1 snRNA: | 3' | G | U | C | C | A | U | U | C | A | 5' |
|---|---|---|---|---|---|---|---|---|---|---|---|

*J. Mol. Biol.* (1997) **268**, 78–94

An example of an annotated gene…




Current Opinion in Structural Biology 1998, 8:346–354

How well do these programs work?
We can measure how well an algorithm works using these:

**True answer:**

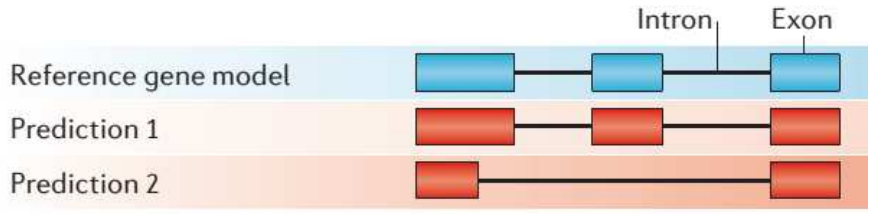
| **Algorithm predicts:** | Positive | Negative |
|---|---|---|
| Positive | True positive | False positive |
| Negative | False negative | True negative |

Specificity = TP / (TP + FP)

Sensitivity = TP / (TP + FN)

*Nature Reviews Genetics* 13:329-342 (2012)

How well do these programs work?
How good <u>are</u> our current gene models?




| SN | SP |
|---|---|
| 1 (1) | 1 (1) |
| 0.63 (0.33) | 1 (0.5) |

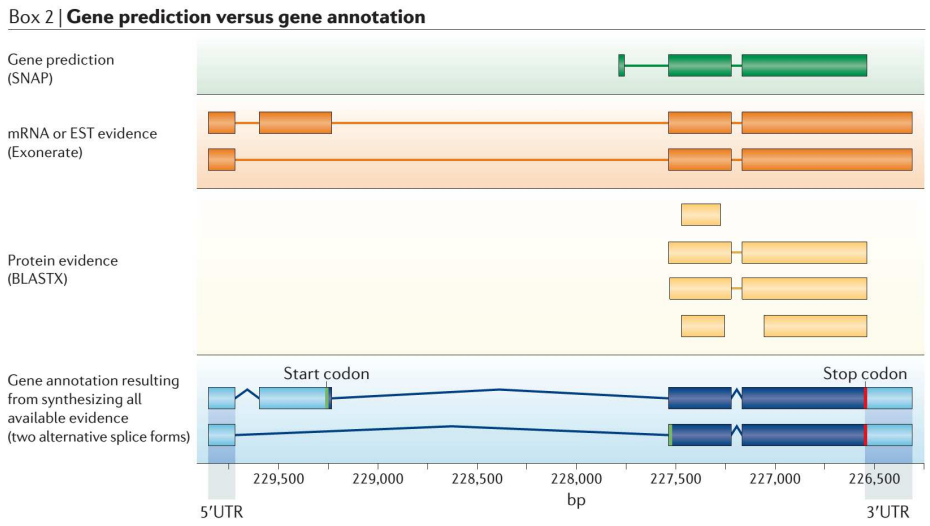
*Nature Reviews Genetics* 13:329-342 (2012)

GENSCAN, when it was first developed….

| Program | Sequences | Accuracy per base Sn | Sp | Accuracy per exon Sn | Sp |
|---|---|---|---|---|---|
| GENSCAN | 570 (8) | 0.93 | 0.93 | 0.78 | 0.81 |
| FGENEH | 569 (22) | 0.77 | 0.88 | 0.61 | 0.64 |
| GeneID | 570 (2) | 0.63 | 0.81 | 0.44 | 0.46 |
| Genie | 570 (0) | 0.76 | 0.77 | 0.55 | 0.48 |
| GenLang | 570 (30) | 0.72 | 0.79 | 0.51 | 0.52 |
| GeneParser2 | 562 (0) | 0.66 | 0.79 | 0.35 | 0.40 |
| GRAIL2 | 570 (23) | 0.72 | 0.87 | 0.36 | 0.43 |
| SORFIND | 561 (0) | 0.71 | 0.85 | 0.42 | 0.47 |
| Xpound | 570 (28) | 0.61 | 0.87 | 0.15 | 0.18 |
| GeneID+ | 478 (1) | 0.91 | 0.91 | 0.73 | 0.70 |
| GeneParser3 | 478 (1) | 0.86 | 0.91 | 0.56 | 0.58 |

*J. Mol. Biol.* (1997) **268**, 78–94

In general, we can do better with more data, such as mRNA and conservation




*Nature Reviews Genetics* 13:329-342 (2012)

How well do we know the genes now?

**In the year 2000**

## Genome Annotation Assessment in *Drosophila melanogaster*

= scientists from around the world held a contest ("GASP") to predict genes in part of the fly genome, then compare them to experimentally determined "truth"

**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 | | | | |

Genome Research 10:483–501 (2000)

How well do we know the genes now?

**In the year 2000**

"Over 95% of the coding nucleotides … were correctly identified by the majority of the gene finders."

"…the correct intron/exon structures were predicted for >40% of the genes."

Most promoters were missed; many were wrong.

"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."

Genome Research 10:483–501 (2000)

How well do we know the genes now?

**In the year *2006***

**EGASP: the … Assessment Project**

= scientists f… SP") to predict gene… are them to experimenta…

18 groups
36 programs

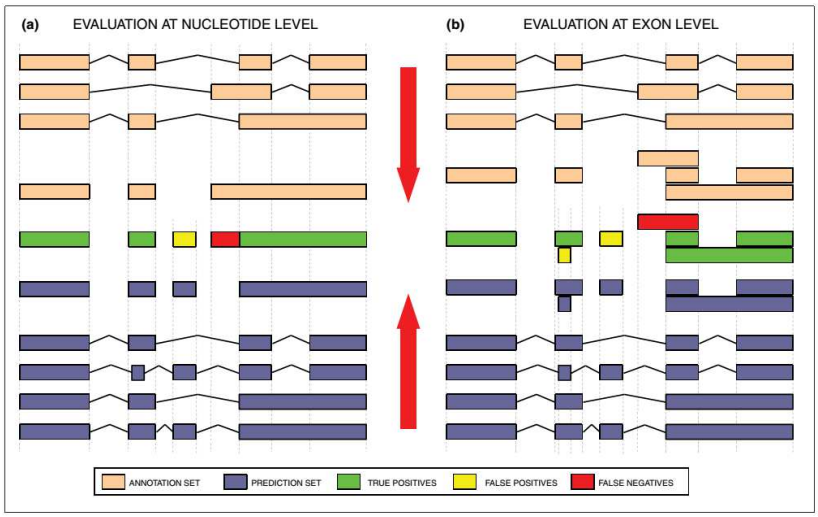
We discussed these earlier

**Table 3**

**Summary of programs used to determine predictions submitted for each EGASP category**

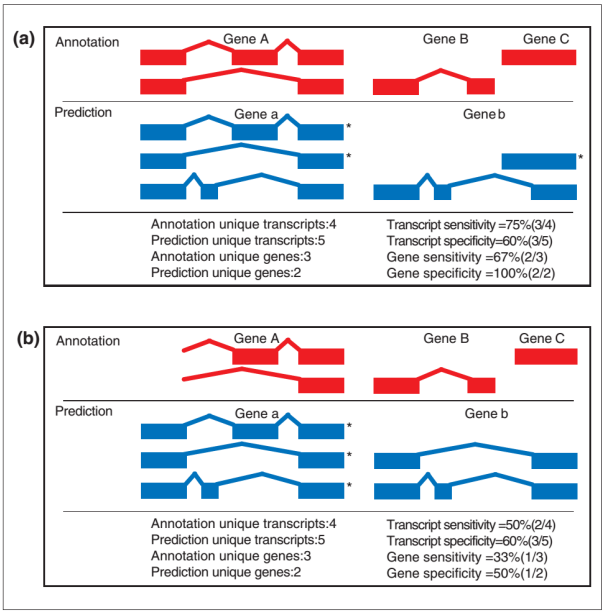
| Submission category | Program | Affiliation | Reference |
|---|---|---|---|
| 1 (AUGUSTUS-any) | AUGUSTUS | Georg-August-Universität, Göttingen | [58] |
| 2 (AUGUSTUS-abinit) | | | |
| 3 (AUGUSTUS-EST) | | | |
| 4 (AUGUSTUS-dual) | | | |
| 1 | FGENESH++ | Softberry Inc. | [56] |
| 1 | JIGSAW | The Institute for Genomic Research (TIGR) | [59] |
| 1 (PAIRAGON-any) | PAIRAGON and NSCAN_EST | Washington University, Saint Louis (WUSTL) | [57] |
| 3 (PAIRAGON+NSCAN_EST) | | | |
| 2 | GENEMARK.hmm | Georgia Institute of Technology | [60] |
| 2 | GENEZILLA | TIGR | [81] |
| 3 | ACEVIEW | National Center for Biotechnology Information (NCBI) | [52] |
| 3 | ENSEMBL | The Wellcome Trust Sanger Institute (WTSI) and European Bioinformatics Institute (EBI) | [64] |
| 3 | EXOGEAN | Ecole Normale Superieure, Paris | [62] |
| 3 | EXONHUNTER | University of Waterloo | [63] |
| 4 | ACESCAN* | Salk Institute | [82] |
| 4 | DOGFISH-C | WTSI | [67] |
| 4 | NSCAN | WUSTL | [57] |
| 4 | SAGA | University of California at Berkeley | [66] |
| 4 | MARS | WUSTL - EBI | [65] |
| 5 | GENEID-U12 | Institut Municipal d'Investigació Médica, Barcelona | – |
| 5 | SGP2-U12 | | |
| 6 | ASPIC† | Università degli Studi di Milano | [83] |
| 6 (AUGUSTUS-exon) | AUGUSTUS | Georg-August-Universität, Göttingen | [58] |
| 6 | CSTMINER‡ | Università degli Studi di Milano | [84] |
| 6 | DOGFISH-C-E§ | WTSI | [67] |
| 6 | SPIDA | EBI | [85] |
| 6 | UNCOVER‖ | Duke University | [86] |
| 1 | CCDSGene | UCSC tracks [7] | [55] |
| 1 | KNOWNGene | | [54] |
| 1 | REFSEQ (REFGene) | | [4] |
| 2 | GENEID | | [19] |
| 2 | GENSCAN | | [18] |
| 3 | ACEMBLT | | [52] |
| 3 | ECGene | | [53] |
| 3 | ENSEMBL (ENSGene) | | [6] |
| 3 | MGCGene | | [5] |
| 4 | SGP2 | | [9] |
| 4 | TWINSCAN | | [12,13] |
| - | CODING 20050607 | GENCODE annotation | [33] |
| - | GENES 20050607 | | |

*Genome Biology* 2006, **7(Suppl 1)**:S2




Genome Biology 2006, 7(Suppl 1):S2

## Transcripts vs. genes




Genome Biology 2006, 7(Suppl 1):S2

In the year *2006*

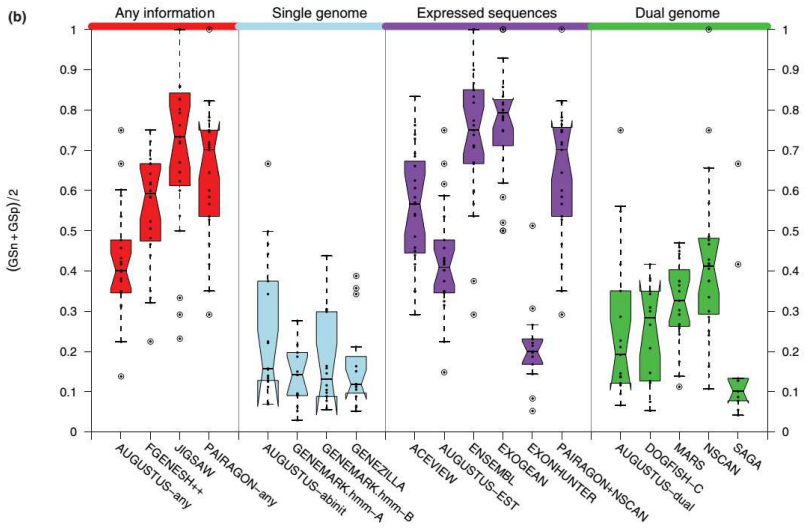
## So how did they do?

- "The best methods had at least one gene transcript correctly predicted for close to **70%** of the annotated genes."
- "…taking into account alternative splicing, … only approximately **40%** to **50%** accuracy.
- At the coding nucleotide level, the best programs reached an accuracy of **90%** in both sensitivity and specificity."

*Genome Biology* 2006, **7(Suppl 1)**:S2

In the year *2006*

## At the gene level, most genes have errors




*Genome Biology* 2006, **7(Suppl 1)**:S2

How well do we know the genes now?

**In the year *2008***

## nGASP – the nematode genome annotation assessment project

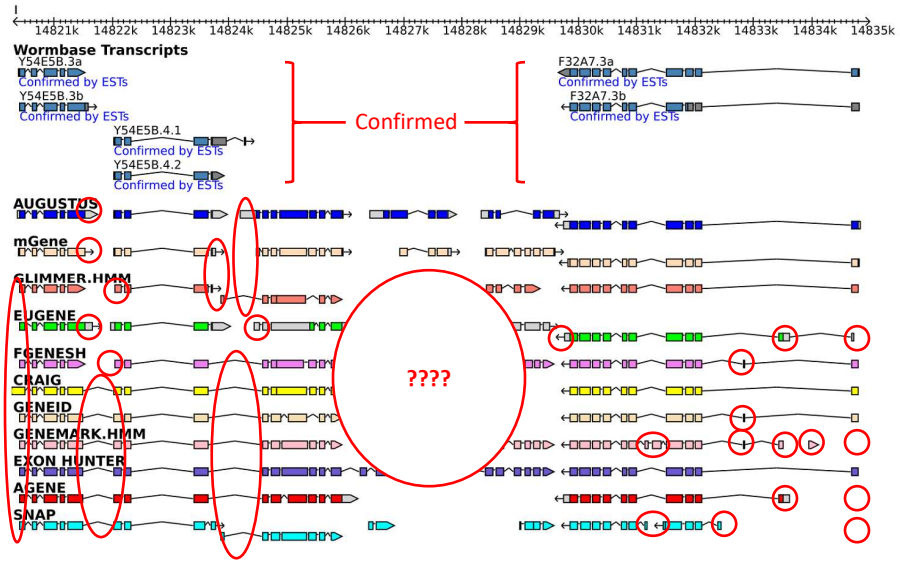
= scientists from around the world held a contest ("NGASP") to predict genes in part of the worm genome, then compare them to experimentally determined "truth"

- 17 groups from around the world competed
- "Median gene level sensitivity … was **78%**"
- "their specificity was **42%**", comparable to human

*BMC Bioinformatics* 2008, **9**:549

For example:

**In the year *2008***




*BMC Bioinformatics* 2008, **9**:549

How well do we know the genes now?

**In the year *2012***

# GENCODE: The reference human genome annotation for The ENCODE Project

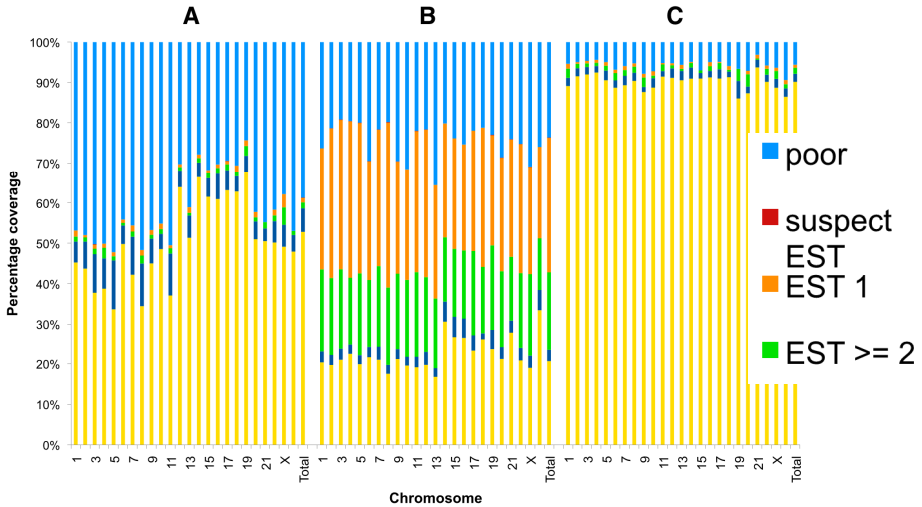
= a large consortium of scientists trying to annotate the human genome using a combination of experiment and prediction.
Best estimate of the current state of human genes.

*Genome Res.* 2012 22: 1760-1774

How well do we know the genes now?

**In the year *2012***

**Quality of evidence used to support automatic, manually, and merged annotated transcripts (probably reflective of transcript quality)**




23,855 transcripts 89,669 transcripts 22,535 transcripts

*Genome Res.* 2012 22: 1760-1774

How well do we know the genes now?

**In the year *2019***

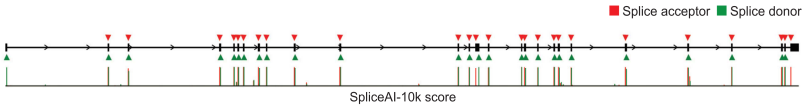
**The bottom line:**

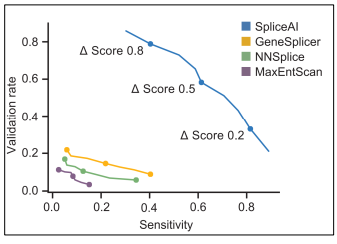
- **Gene prediction and annotation are hard**
- **Annotations for all organisms are still buggy**
- **Few genes are 100% correct; expect multiple errors per gene**
- **"even after 18 years of effort, the precise exon–intron structure of many human protein-coding genes is not settled. The annotation of most other eukaryotes—with the exception of small, intensively studied model organisms like yeast, fruit fly and Arabidopsis—is in worse shape than human annotation."**

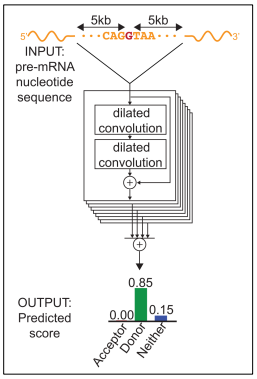
Next-generation genome annotation: we still struggle to get it right
SL Salzberg, *Genome Biology* (20) 92 (2019)

**But the algorithms are nonetheless getting better, e.g. new advances (at last!) in predicting splice sites using deep learning**

**In the year *2019***



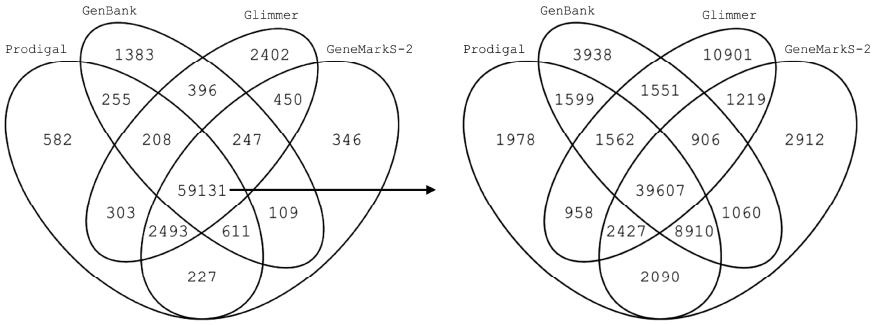




**Predicting Splicing from Primary Sequence with Deep Learning**

Kishore Jaganathan,[1,2] Sofia Kyriazopoulou Panagiotopoulou,[1,2] Jeremy F. McRae,[1,3] Siavash Fazel Darbandi,[2] David Knowles,[1] Yang I. Li,[1] Jack A. Kosmicki,[1,2] Juan Arbelaez,[1] Wenwu Cui,[1] Grace B. Schwartz,[1] Eric D. Chow,[4] Efstathios Kanterakis,[1] Hong Gao,[1] Amirali Kia,[1] Serafim Batzoglou,[1] Stephan J. Sanders,[5] and Kyle Kai-How Farh[1,2,*]
[1]Illumina Artificial Intelligence Laboratory, Illumina, Inc., San Diego, CA, USA
[2]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA
[3]Department of Genetics, Stanford University, Stanford, CA, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA
[†]These authors contributed equally
[*]Lead Contact
[*]Correspondence: kfarh@illumina.com
https://doi.org/10.1016/j.cell.2018.12.015

## What about the current state of prokaryote gene models?

**Here's the overlap in gene predictions from 4 algs on 20 test strains:**




Coding regions agree (shared stop)

Starts and stops agree

**AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions**

Deepank R. Korandla [1,2,3], Jacob M. Wozniak[4,5], Anaamika Campeau[4,5], David J. Gonzalez[4,5] and Erik S. Wright [3,*]

*Bioinformatics*, 36(4), 2020, 1022–1029

## What about the current state of prokaryote gene models?

- **"We applied AssessORF to compare gene predictions offered by GenBank, GeneMarkS-2, Glimmer and Prodigal on genomes spanning the prokaryotic tree of life.**
- **Gene predictions were 88–95% in agreement with the available evidence, with Glimmer performing the worst but no clear winner.**
- ***All programs were biased towards selecting start codons that were upstream of the actual start."***

*Bioinformatics*, 36(4), 2020, 1022–1029

**In practice, gene finding and genome annotation combines all lines of evidence, e.g. as for the frog genome:**

**Align frog RNA sequencing data (ESTs and cDNA) & BLAST genes from other animals vs. frog assembly** → **Define gene segments**

↙

**Integrate *ab initio* gene predictions & BLAST hits using Fgenesh and GenomeScan** (= GenScan successor, *Genome Research* 11:803 (2001))

↓

**Refine with RNA-seq and H3K4me3 data**

↓

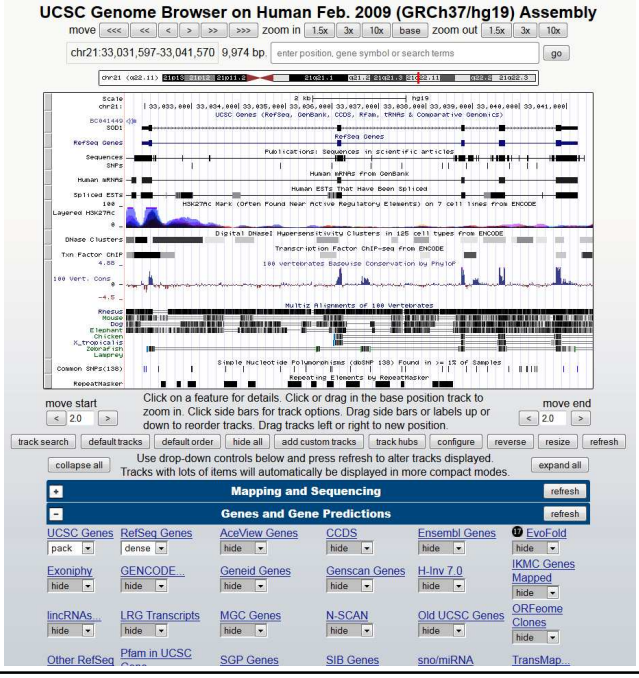
**Refine vs final genome assembly**

↓

**Manually curate 412 gene models**
**→ Estimate 96% accuracy overall**


Session *et al., Nature* 2016
Supplementary Info, pg. 22

The Univ of California Santa Cruz genome browser



The Univ of California Santa Cruz genome browser