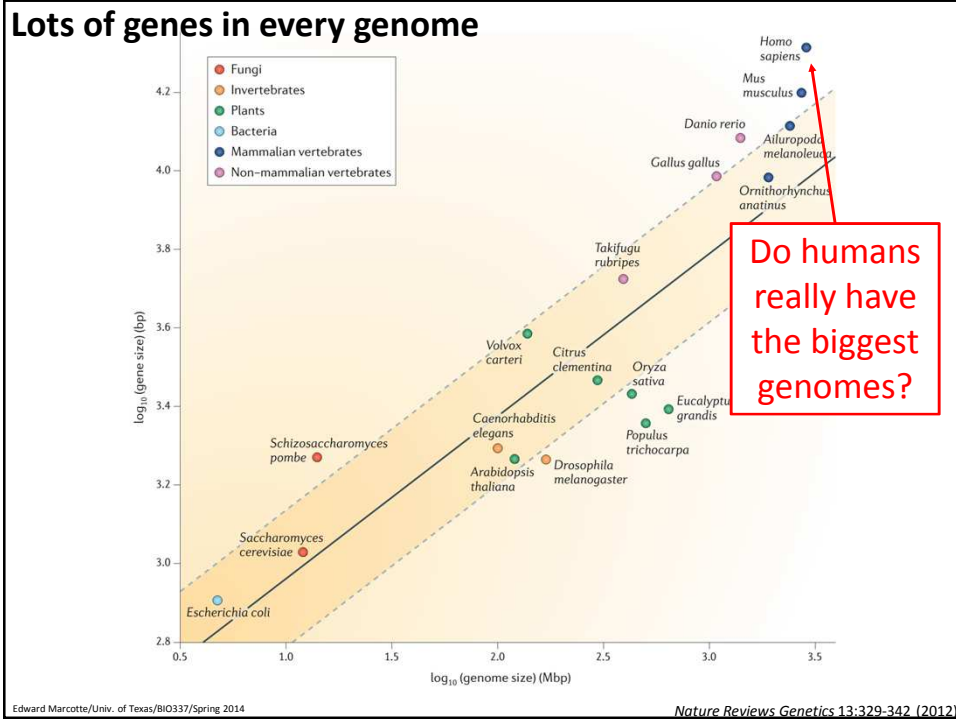


Gene Finding

BIO337 Systems Biology / Bioinformatics – Spring 2014

Edward Marcotte, Univ of Texas at Austin

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014



Lots of genes in every genome

animal Largest Genome Ever Sequenced Belongs To Locust Species

January 17, 2014
f t+ d+ v+ p+ 12 Print



6.5 Gb
(2X human)
17K genes

Image Credit: Thinkstock.com

April Flowers for redOrbit.com – Your Universe Online

The whole genome sequence of Locust (*Locusta migratoria*), the most widespread locust species, has been successfully decoded by researchers from the Institute of Zoology, Chinese Academy of Sciences, BGI and other institutes. The researchers were surprised by the remarkably large (6.5 gigabases) yielded genome, which is the largest animal genome sequenced so far.

gigabases (not gigabytes)

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

The Telegraph

Home News World Sport Finance Comment Culture Travel Life Women Fashion
Politics Investigations Obits Education Earth Science Defence Health Scotland Royal
Science News Space Night Sky Roger Highfield Dinosaurs Evolution Steve Jones Science

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.



Print this article
Share 304
Facebook 248
Twitter 56
Email
LinkedIn 0
G+ 0

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

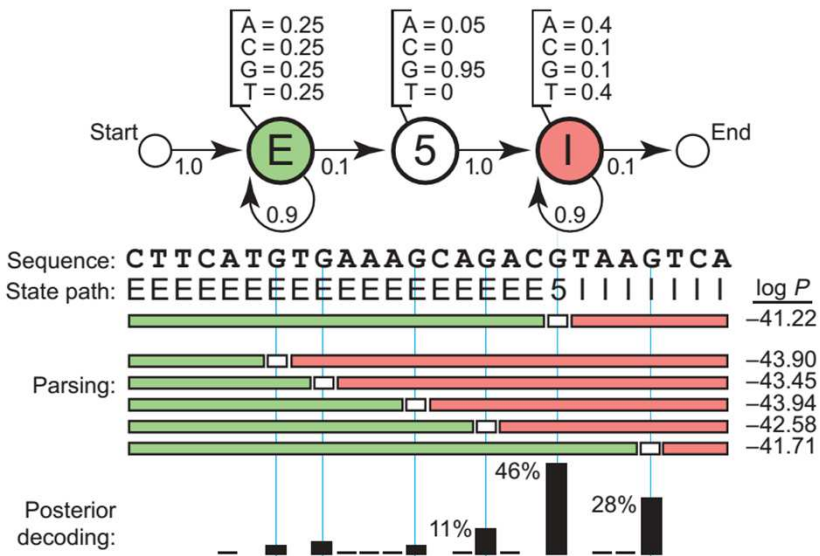
Science News
News » UK News »
Science »
Earth News »

Where are the genes? How can we find them?

```
GATCACTTGATAAATGGGCTGAAGTAACTCGCCAGATGAGGAGTGTGCTGCCAGCAAT
CCAAACAGGCCCACTAGGCCCGAGACACCTTGTCTCAGATGAAACTTTGGACTCGGAATT
TTGAGTTAATGCCGGAATGAGTTCAGACTTTGGGGGACTGTTGGGAAGGCATGATTGGTT
TCAAATGTGAGAAGGACATGAGATTTGGGAGGGGCTGGGGCAGAATGATATAGTTTG
GCTCTGCGTCCCCACCAATCTCATGTCAAATTGTAATCCTCATGTGCAGGGGAGAGGCCT
GGTGGGATGTGATTGGATCATGGGAGTGGATTTCCCTCTTGCACTTCTCGTGATAGTGAGT
GAGTTCTCACGAGATCTGGTTGTTTGAAAGTGTGCAGCTCCTCCCCCTTCGCGCTCTCTCT
TCCCCTGTCCACCATGGTGAGACGTGCTTGCCTCCCTTTGCCTTCTGCCATGATTGTAAG
CTTCCTCAGGCGTCTAGCCACGCTTCTGTACAGCCTGAGGAACTGGGAGTCAATGAAA
CCTCTTCTTTCATAAATTACCCAGTTTCAGGTAGTTCTTTCTAGCAGTGTGATAATGGACGA
TACAAGTAGAGACTGAGATCAATAGCATTTCACACTGGGCCTGGAACACACTGTTAAGAAC
GTAAGAGCTATTGCTGTCATTAGTAATATTCTGTATTATTGGCAACATCATCACAATACACTGC
TGTGGGAGGGTCTGAGATACTTCTTTGCACTCCAATTTGTCAAACATAAAATCAGG
AGCCTCATGAATAGTGTAAATTTTACATAATAATACATTGCACCATTTGGTATATGAGTCT
TTTTGAAATGGTATATGCAGGACGGTTTCTAATATACAGAATCAGGTACACCTCCTCTTCCA
TCAGTGCCTGAGTGTGAGGGATTGAATCCTCTGGTTAGGAGTTAGCTGGCTGGGGGTTTC
TACTGCTGTTGTACCCACAGTGCACCTCAGACTCACGTTTCTCCAGCAATGAGCTCCTGTT
CCCTGCACTTAGAGAAGTCAGCCCGGGGACCAGACGGTTCTCTCTTTCCTGCTGCCAG
CCTTGGCCTTCAGCAGTCTGGATGCCTATGACACAGAGGGCATCCTCCCAAGCCCTGGTC
CTTCTGTGAGTGGTGGTGGTCTGTTAATCCAAAAGGACAGGTGAAAACATGAAAGCC...
```

Edward Marcotte/Univ. of Texas/BI0337/Spring 2014

A toy HMM for 5' splice site recognition (from Remember this! linked on the course web page)



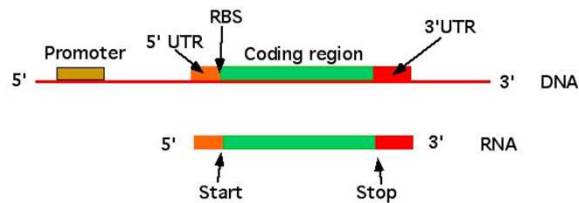
Edward Marcotte/Univ. of Texas/BI0337/Spring 2014cc

Let's start with prokaryotic genes

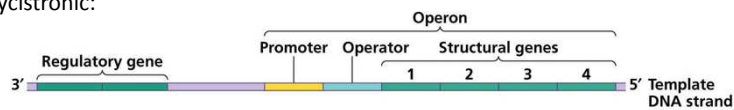
What elements should we build into an HMM to find bacterial genes?

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Let's start with prokaryotic genes



Can be polycistronic:



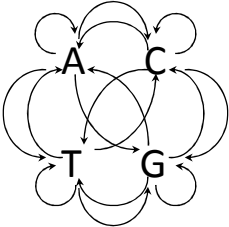
Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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

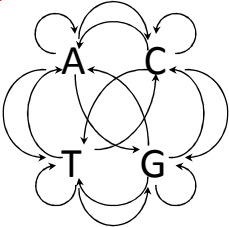
A CpG island model might look like: **Remember this?**

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



p(C→G) is higher

CpG island model



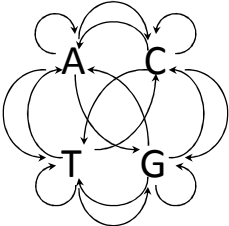
p(C→G) is lower

Not CpG island model

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

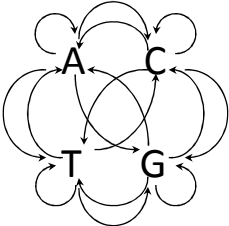
Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

One way to build a minimal gene finding Markov model



Transition probabilities reflect codons

Coding DNA model



Transition probabilities reflect intergenic DNA

Intergenic DNA model

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

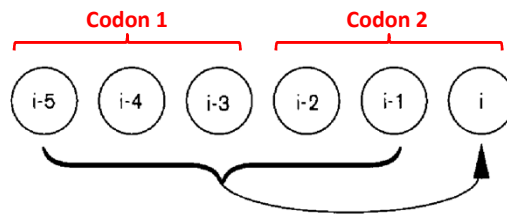
Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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.

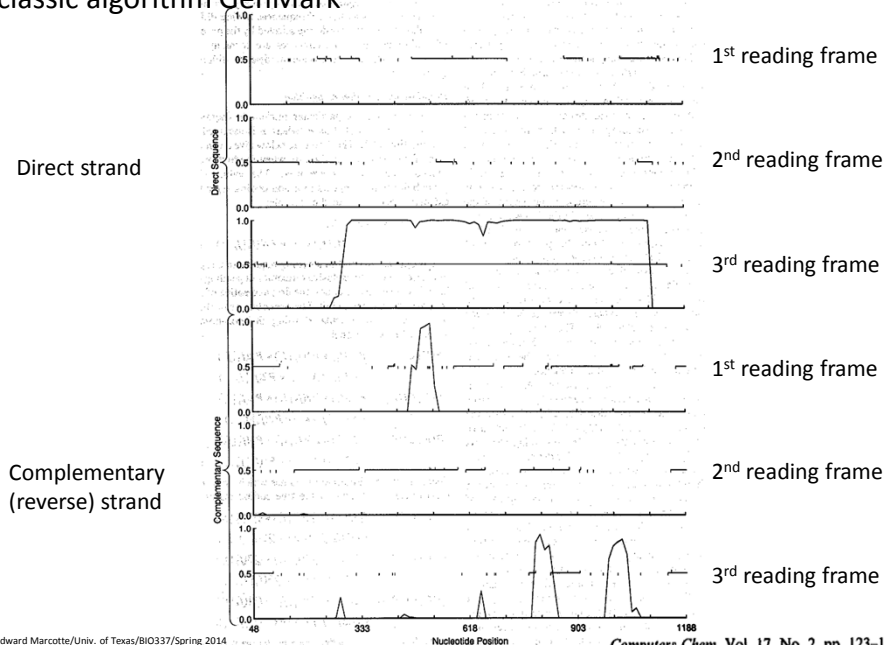
An n^{th} -order Markov process takes into account the past n nucleotides, e.g. as for a 5th order:



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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

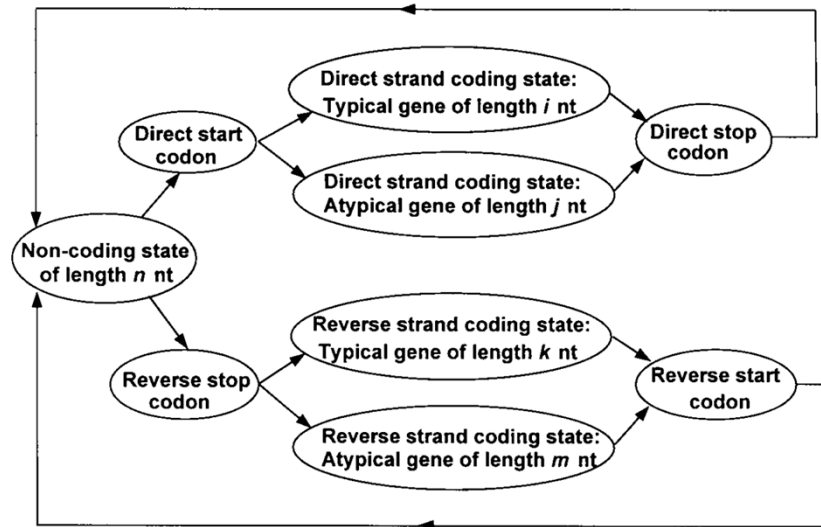
5th order Markov chain, using models of coding vs. non-coding using the classic algorithm GenMark



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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

Edward Marcotte/Univ. of Texas/BI0337/Spring 2014

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

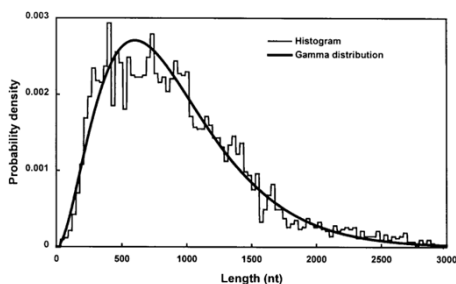
Alexander V. Lukashin and Mark Borodovsky^{1*}

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

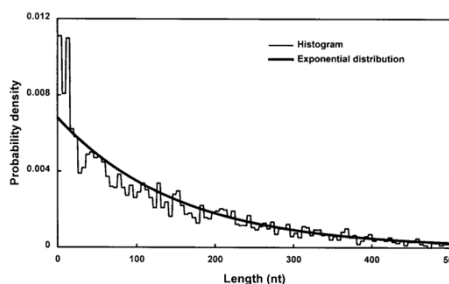
Edward Marcotte/Univ. of Texas/BI0337/Spring 2014

... and variation in gene lengths

Length distributions (in # of nucleotides)



Coding (ORFs)



Non-coding (intergenic)

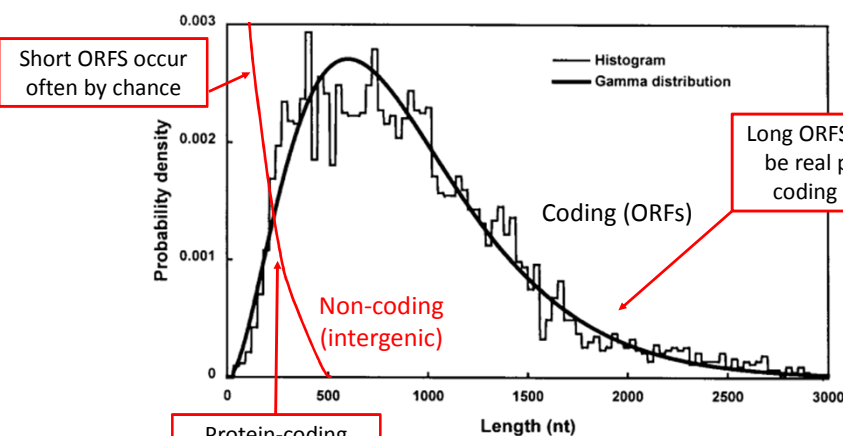
GeneMark.hmm: new solutions for gene finding

Alexander V. Lukashin and Mark Borodovsky^{1*}

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

Edward Marcotte/Univ. of Texas/BI0337/Spring 2014

(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

Edward Marcotte/Univ. of Texas/BI0337/Spring 2014

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

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

How well does it do on well-characterized genomes?

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

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

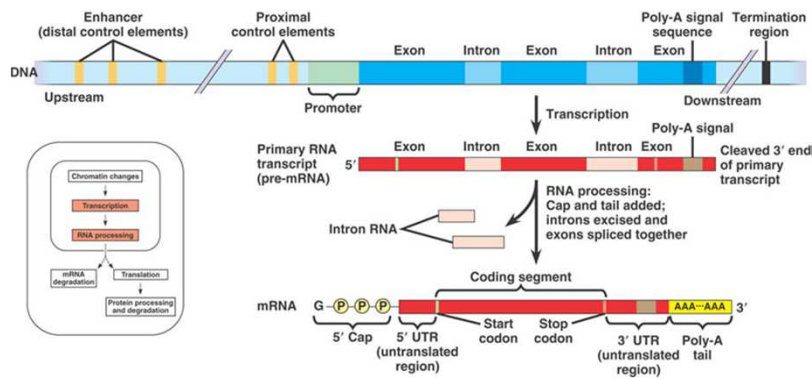
But this was a long time ago!

Eukaryotic genes

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

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

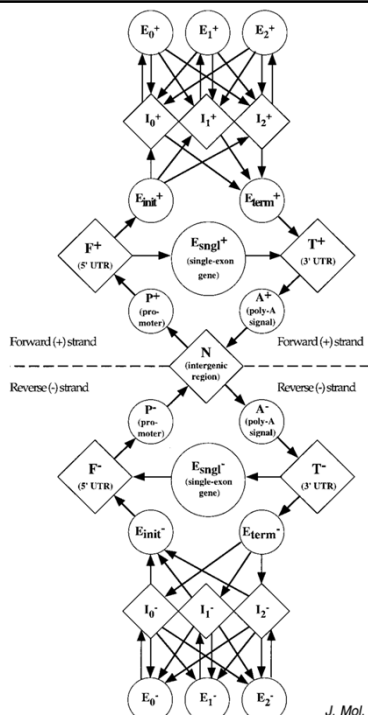
Eukaryotic genes



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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

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

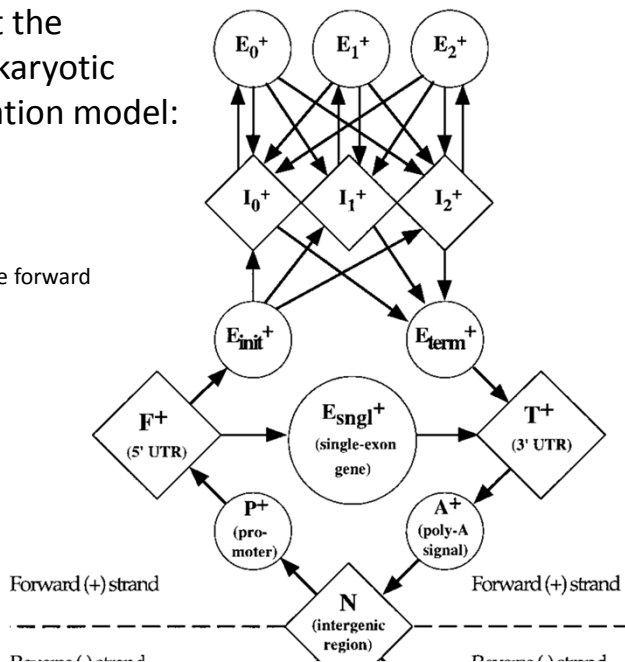


Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

J. Mol. Biol. (1997) 268, 78-94

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

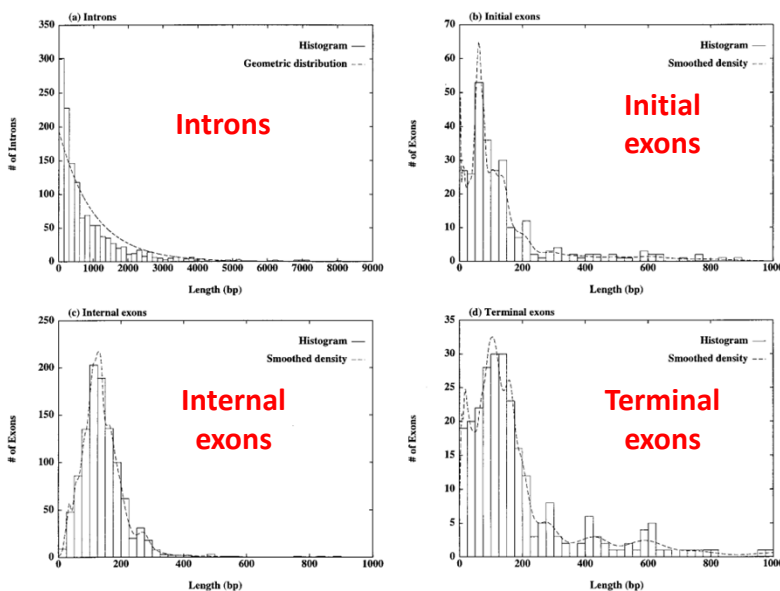
Zoomed in on the forward strand model...



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

J. Mol. Biol. (1997) 268, 78-94

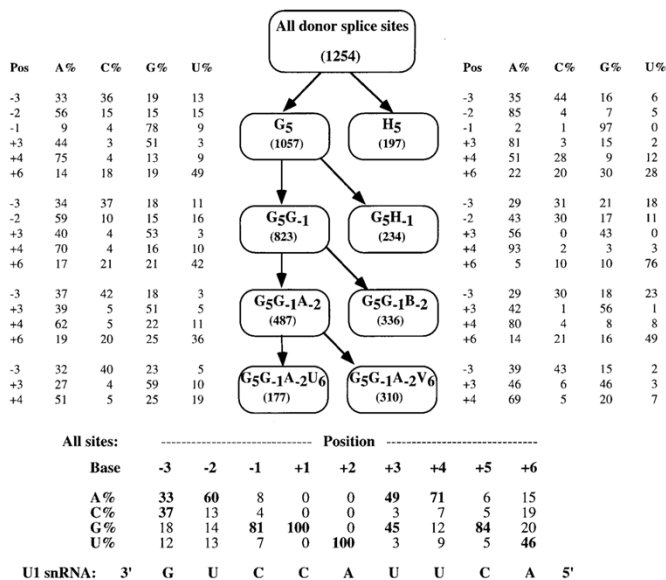
Introns and different flavors of exons all have different typical lengths



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

J. Mol. Biol. (1997) 268, 78–94

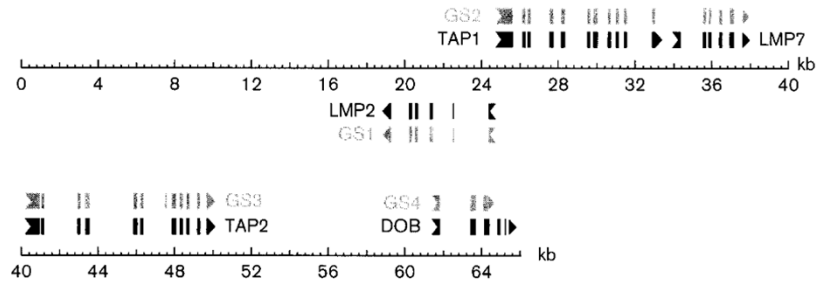
Taking into account donor splice sites



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

J. Mol. Biol. (1997) 268, 78–94

An example of an annotated gene...



⌞	□
⌞	□
⌞	⌞
■	■

⌞ Initial exon □ Internal exon
 ⌞ Terminal exon ⌞ Single-exon gene
 ■ GENSCAN predicted exon
 ■ GenBank annotated exon

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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:

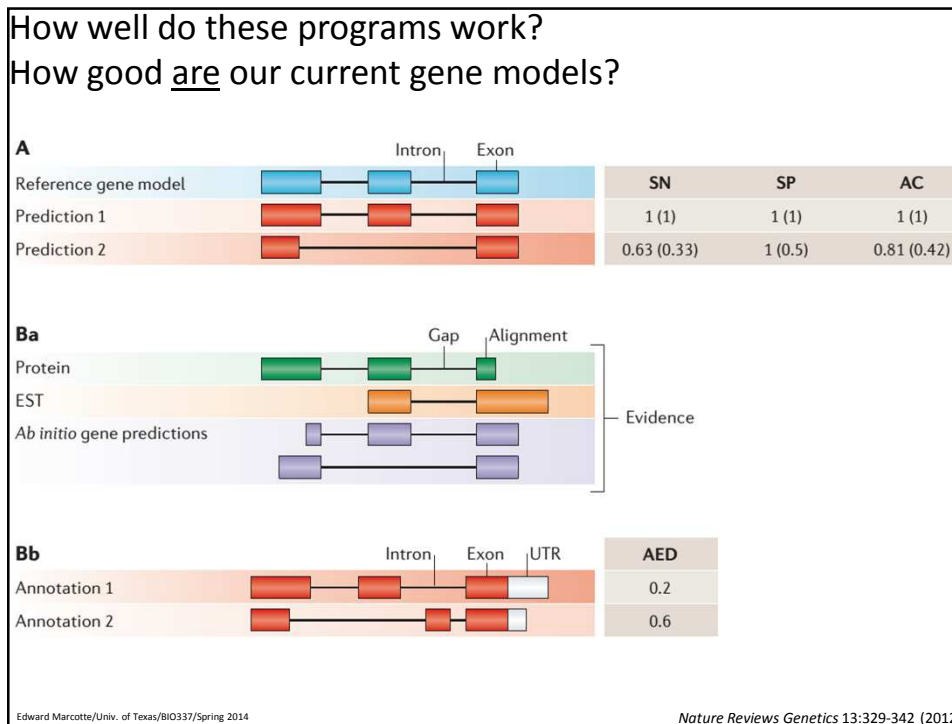
		Positive	Negative
Algorithm predicts:	Positive	True positive	False positive
	Negative	False negative	True negative

Specificity = $TP / (TP + FP)$

Sensitivity = $TP / (TP + FN)$

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Nature Reviews Genetics 13:329-342 (2012)

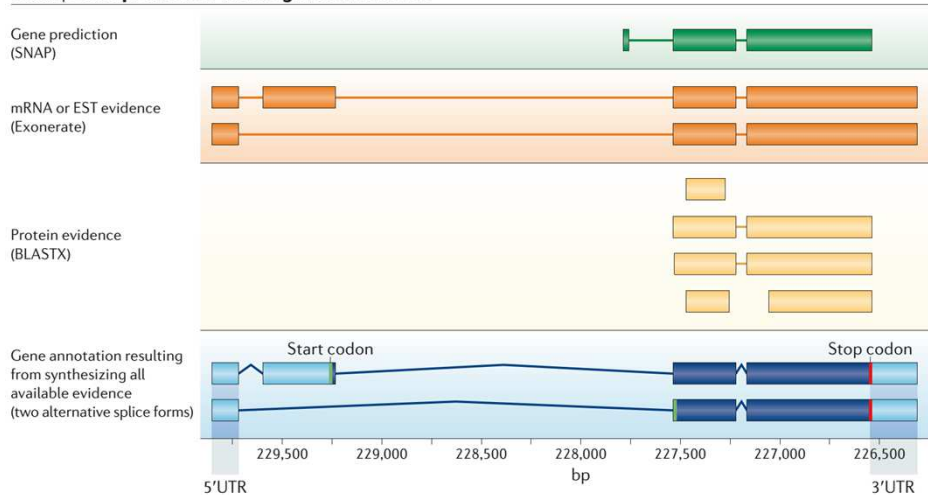


GENSCAN, when it was first developed....

Program	Sequences	Accuracy per base		Accuracy per exon	
		Sn	Sp	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

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

Box 2 | Gene prediction versus gene annotation



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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	HMWGene	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
Resse et al. Berkeley/Santa Cruz, US	Gene	X	X				

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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.”

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Genome Research 10:483-501 (2000)

How well do we know the genes now?

In the year 2006

EGASP: the Project

= scientists predict gene experimental

18 groups
36 programs

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-absent)			
3 (AUGUSTUS-EST)			
4 (AUGUSTUS-dual)			
1	FGENESH++	Solberry 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.intron	Georgia Institute of Technology	[60]
2	GENEMARK	TIGR	[61]
2	ACEVIEW	National Center for Biotechnology Information (NCBI)	[52]
2	ENSEMBL	The Wellcome Trust Sanger Institute (WTSI) and European Bioinformatics Institute (EBI)	[64]
3	EXONSCAN	Ecole Normale Supérieure, 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	-
5	SGP2-LI12	Médica, Barcelona	-
6	ASPICT	Università degli Studi di Milano	[83]
6 (AUGUSTUS-exon)	AUGUSTUS	Georg-August-Universität, Göttingen	[58]
6	CSTMINDER	Università degli Studi di Milano	[84]
6	DOGfish-C-EP	WTSI	[67]
6	SPIDA	EBI	[85]
6	UNCOVER1	Duke University	[86]
1	CCDSGene	UCSC tracks [7]	[55]
1	KNOWNGene		[54]
1	REFSEQ (REFGene)		[4]
2	GENEID		[19]
2	GENSCAN		[18]
3	TRIPBLT		[52]
3	ECGene		[53]
3	ENSEMBL (ENSGene)		[6]
3	HGCCGene		[5]
4	SGP2		[9]
4	TWINSCAN		[12,13]
-	CODING 20050607	GENCODE annotation	[33]
-	GENES 20050607		

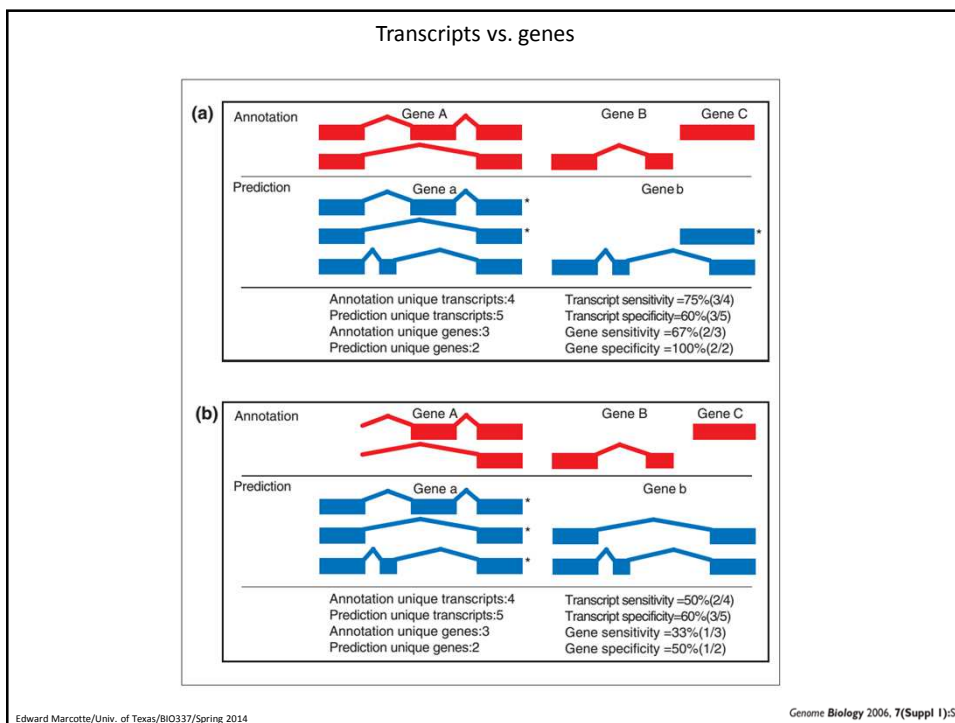
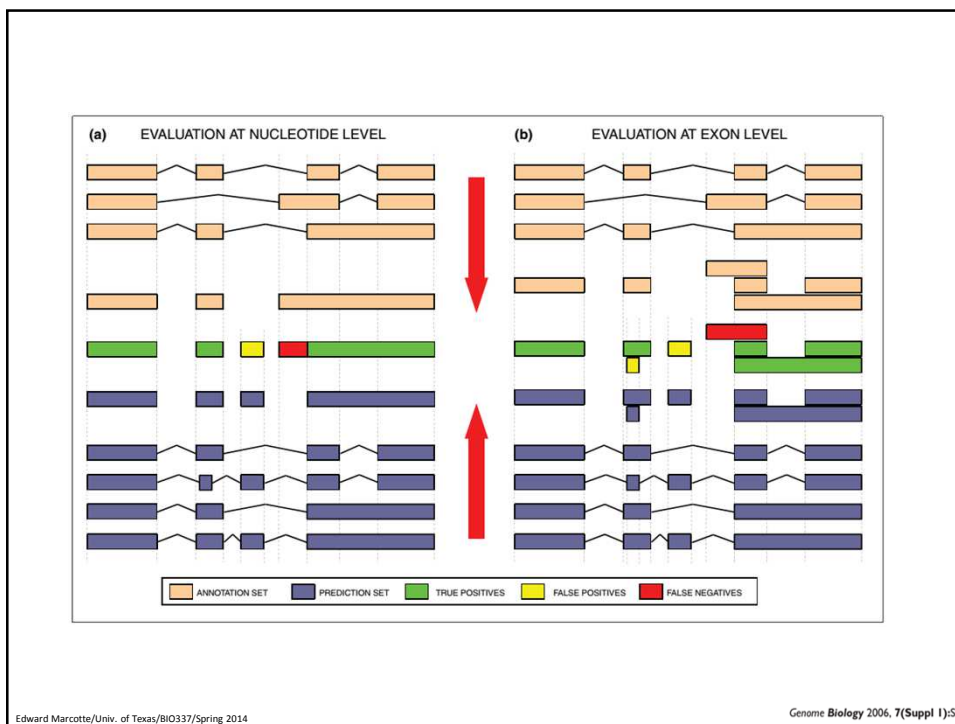
We discussed these earlier

Assessment

SP”) to are them to

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Genome Biology 2006, 7(Suppl 1):S2



In the year 2006

So how did they do?

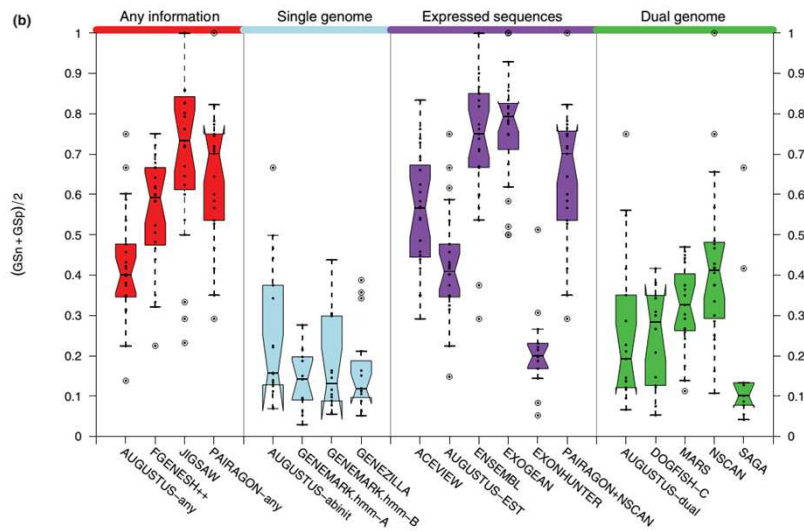
- “The best methods had at least one gene transcript correctly predicted for close to **70%** of the annotated genes.”
- “...taking into account alternative splicing, ... only approximately **40% to 50%** accuracy.
- At the coding nucleotide level, the best programs reached an accuracy of **90%** in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

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Genome Biology 2006, 7(Suppl 1):S2

In the year 2006

At the gene level, most genes have errors



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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”

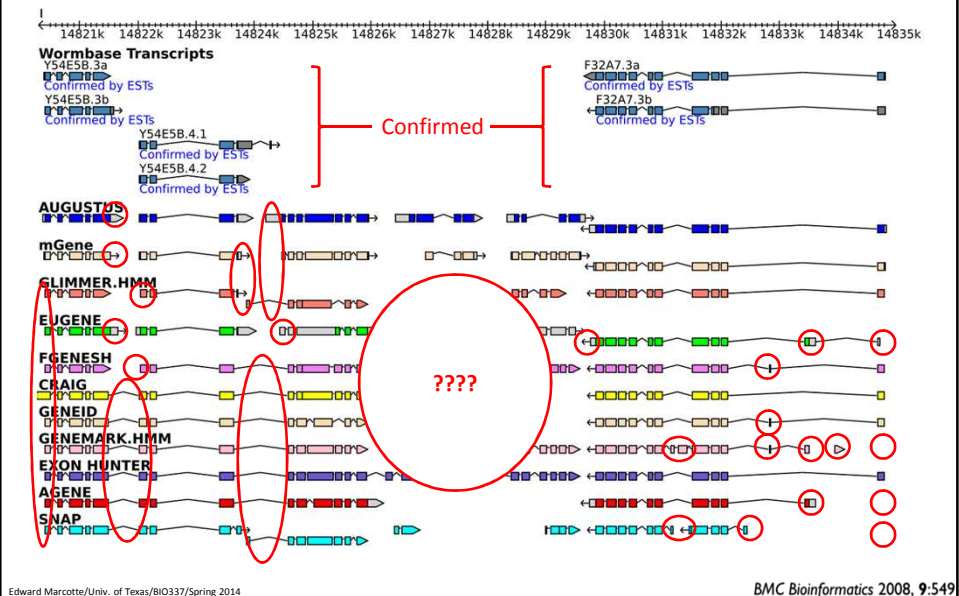
- 17 groups from around the world competed
- “Median gene level sensitivity ... was **78%**”
- “their specificity was **42%**”, comparable to human

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

BMC Bioinformatics 2008, 9:549

For example:

In the year **2008**



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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.

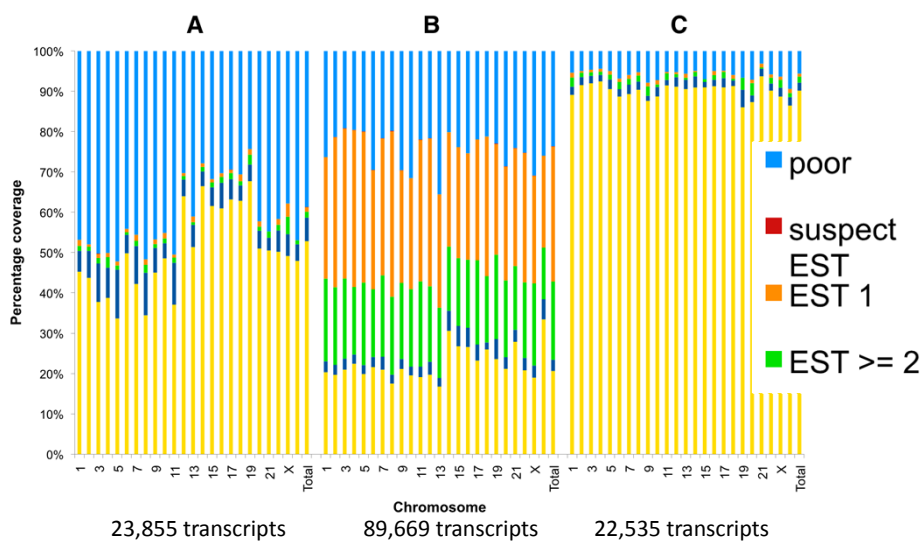
Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

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)



Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

Genome Res. 2012 22: 1760-1774

How well do we know the genes now?

In the year **2014**

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
- Most organisms' gene annotations are probably much worse than for humans

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

The Univ of California Santa Cruz genome browser

The screenshot displays the UCSC Genome Browser interface for a region on chromosome 21 (chr21:33,031,597-33,041,570, 9,974 bp). The browser shows multiple tracks for gene annotations and genomic features. Key tracks include RefSeq Genes, Ensembl Genes, UCSC Genes, Human ESTs, HDZTNC Mark, DNase ChIP, TF ChIP, 100 vertebrate conservation, and RepeatMasker. The interface includes navigation controls such as zoom in/out, move, and search, as well as options to collapse or expand tracks.

Edward Marcotte/Univ. of Texas/BIO337/Spring 2014

The Univ of California Santa Cruz genome browser

