

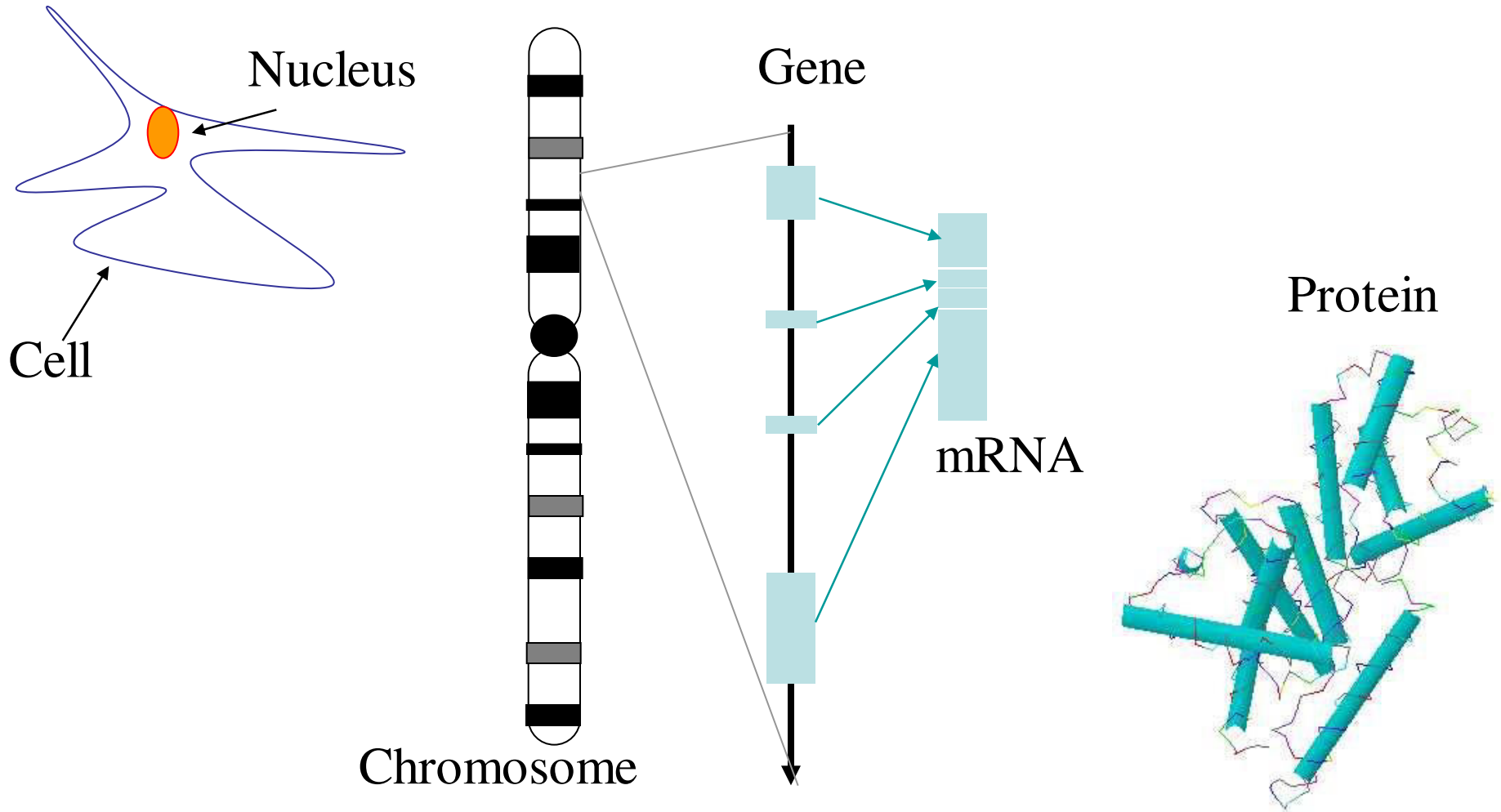


FACULTY OF ENGINEERING & TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY

Dr. Simranjit Singh
Assistant Professor
Department of Biotechnology
Rama University, Kanpur

Genome Annotation

Cells, Chromosomes, DNA, and Genes



Definitions

- Unless otherwise stated, annotation refers to prediction of protein-coding genes
- Methods exist to annotate
 - tRNA, rRNA
 - Several other small RNAs
 - Repetitive elements
 - [microRNAs]

Challenges

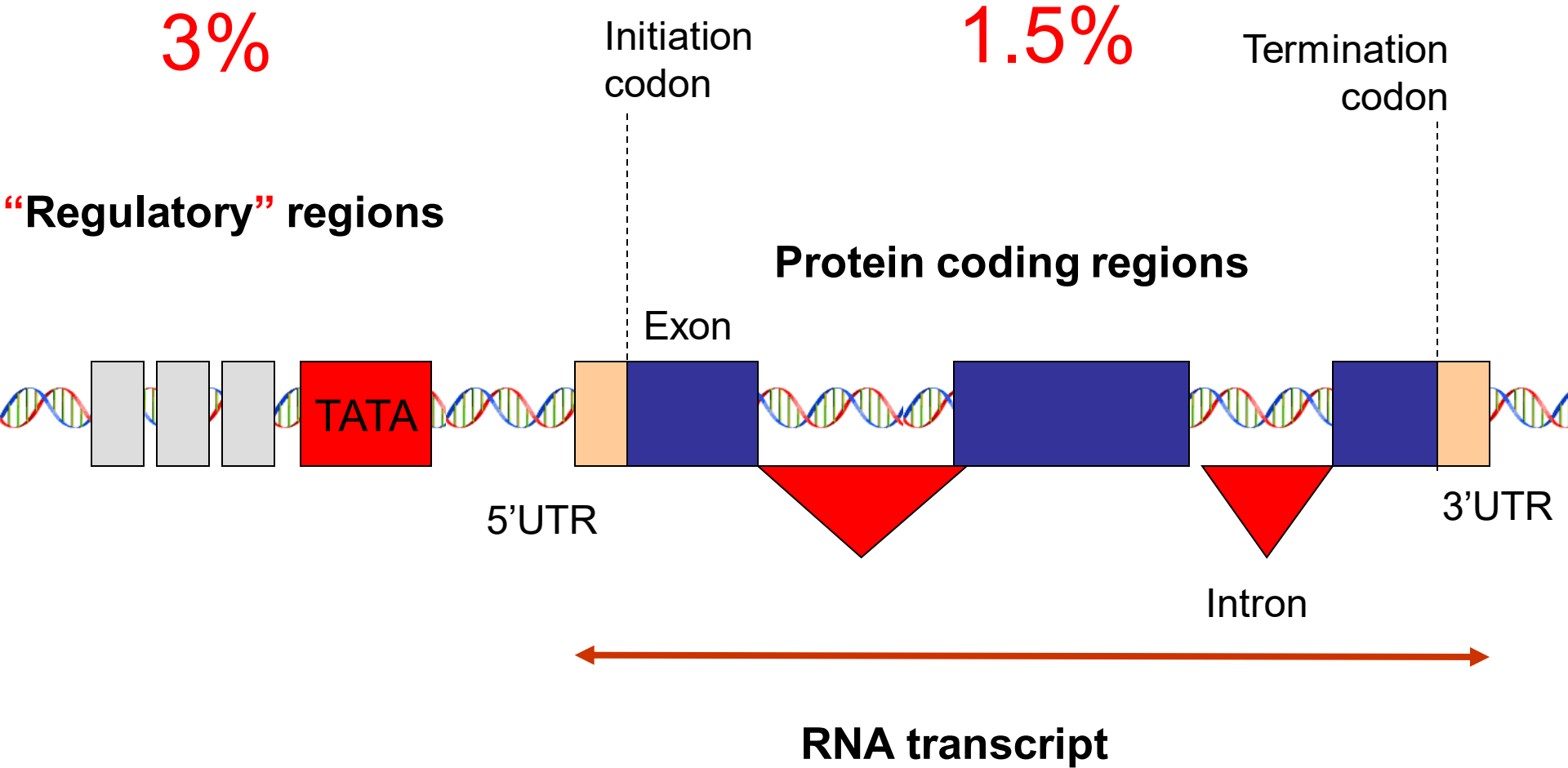
- Signal to noise
 - 1% protein coding sequence
 - Pseudogenes
- Splicing
 - Discontinuous nature of eukaryotic genes
 - Alternative splicing
- Non-uniform genome characteristics
 - Range of G+C content
 - Range of mutation rates
 - Gene/Genome segment duplication

1% codes for protein

```
ACACTCGCTTCTGGAACGTCTGAGGTTATCAATAAGCTCCTAGTCCAGACGCCATGGGTCATTTTCACAGAGGA  
GGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTGGAGGAGAAACCCTGGGAAGG  
CTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCA  
TGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACTTCCTTGGGAGATGCCACAAAGCACCTGGA  
TGATCTCAAGGCACCTTTGCCAGCTAGTGAACTGCACTGTGACAAGCTGCATGTGGATCCTGAGAACTT  
AAGCTCCTGGGAAATGTGCTGGTGACCGTTTTTGGCAATCCATTTCTGGCAAAGAATTCACCCCTGAGGTGCAGG  
CTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGTGCCCTGTCCTCCAGATACCACTGAGCTCACTGCCCATG  
ATTCAGAGCTTTCAAGGATAGGCTTTATTCTGCAAGCAATACAAATAATAAATCTATTCTGCTGAGAGATCAC  
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCCATGGTGCATCTGACTCCTGAGGA  
GAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTG  
CTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGG  
GCAACCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAA  
CCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG  
CTCCTGGGCAACGTCTCCTCTCTCTCTCTCCTCCCCATACCTTTCCAAACAATTCACCCACCAGTGCAGGCTG  
CCTATCAGAAAGTCTGCCTTTCTTGCTGTGCTTTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT  
CCAATTTCTATTAATGCATCTGGATTCTATGAAGGGCCTTGACTCTGAGGTTATCA  
ATAAGCTCCTAGTCCAGACGCCATGGGTCATTTTCACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGC  
AAGGTGAATGTGGAAGATGCTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGT  
TCTTTGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAA  
GAAGGTGCTGACTTCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCCAGCTGAGT  
GAACTGCACTGTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTT  
TGGCAATCCATTTCTGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCTGAGAGCTTTCAAGGATAGGCTTTATTCT  
CAGTGCCCTGTCCTCCAGATACCACTGAGCTCACTGCCCATGATTCAGAGCTTTCAAGGATAGGCTTTATTCT  
GCAAGCAATACAAATAATAAATCTATTCTGCTGAGAGATCACACATTTGCTTCTGACACAACTGTGTTCACTA  
GCACCTAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG  
GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCT  
TTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCTAAGGTGAAGGCTCATGGCAAGAA  
AGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAG  
CTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGG  
CCATCACTTTGGCAATAAATCTATTCTGCTGAGAGATCACACATTTGCTTCTGACACAACTGTGTTCACTAG  
CAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGG  
TGAACGTGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTT  
TGAGTCCTTTGGGGTCTGTCCACTCCTGATGCTGTTATGGGCAACCTAAGGTGAAGGCTCATGGCAAGAAA  
GTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGC
```

3% conserved non-coding

Structure of Genes



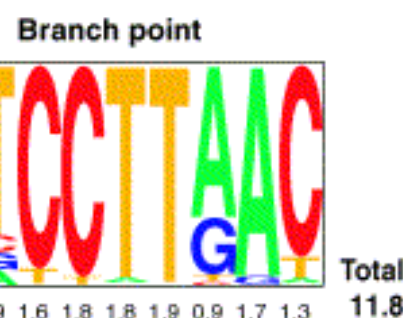
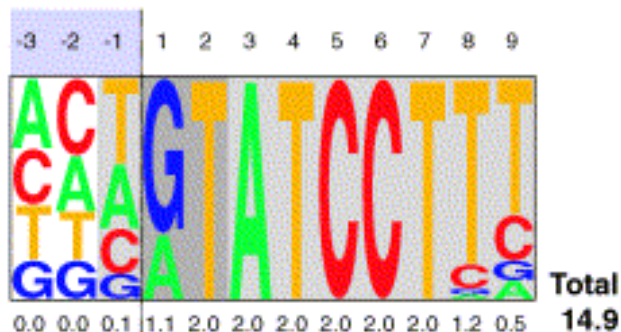
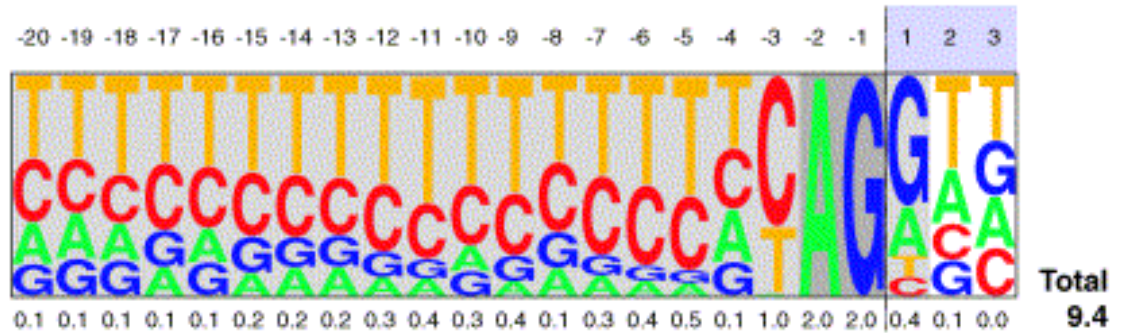
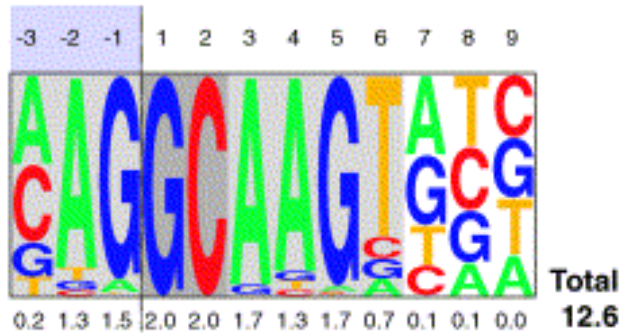
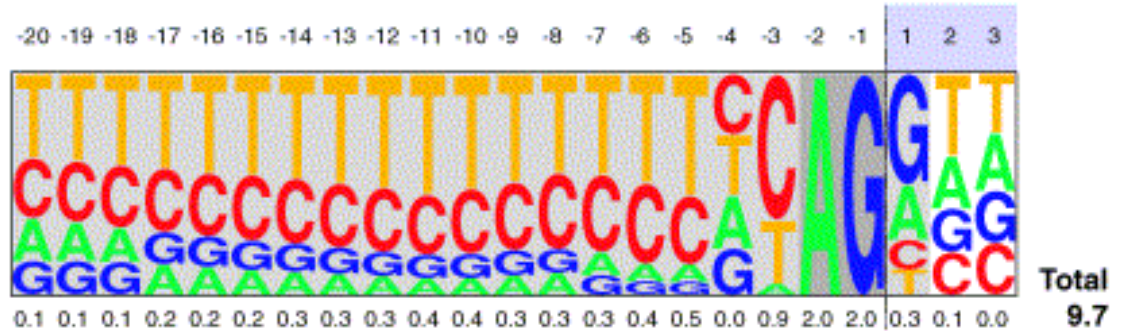
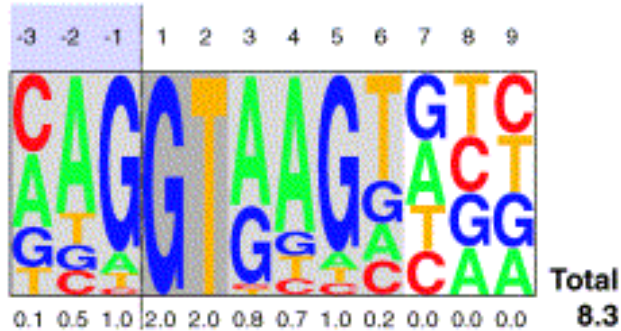
Genome Annotation Methods

- Known Genes
 - Blastn cDNA against genome
- Protein similarity
 - Blastx genome against SWISS-PROT
- Genome-Genome alignment
 - Blastz
- *De novo* prediction
 - GRAIL, Genscan, FGENESH
- Integrated methods
 - Expert review, ‘Combiner’ algorithms

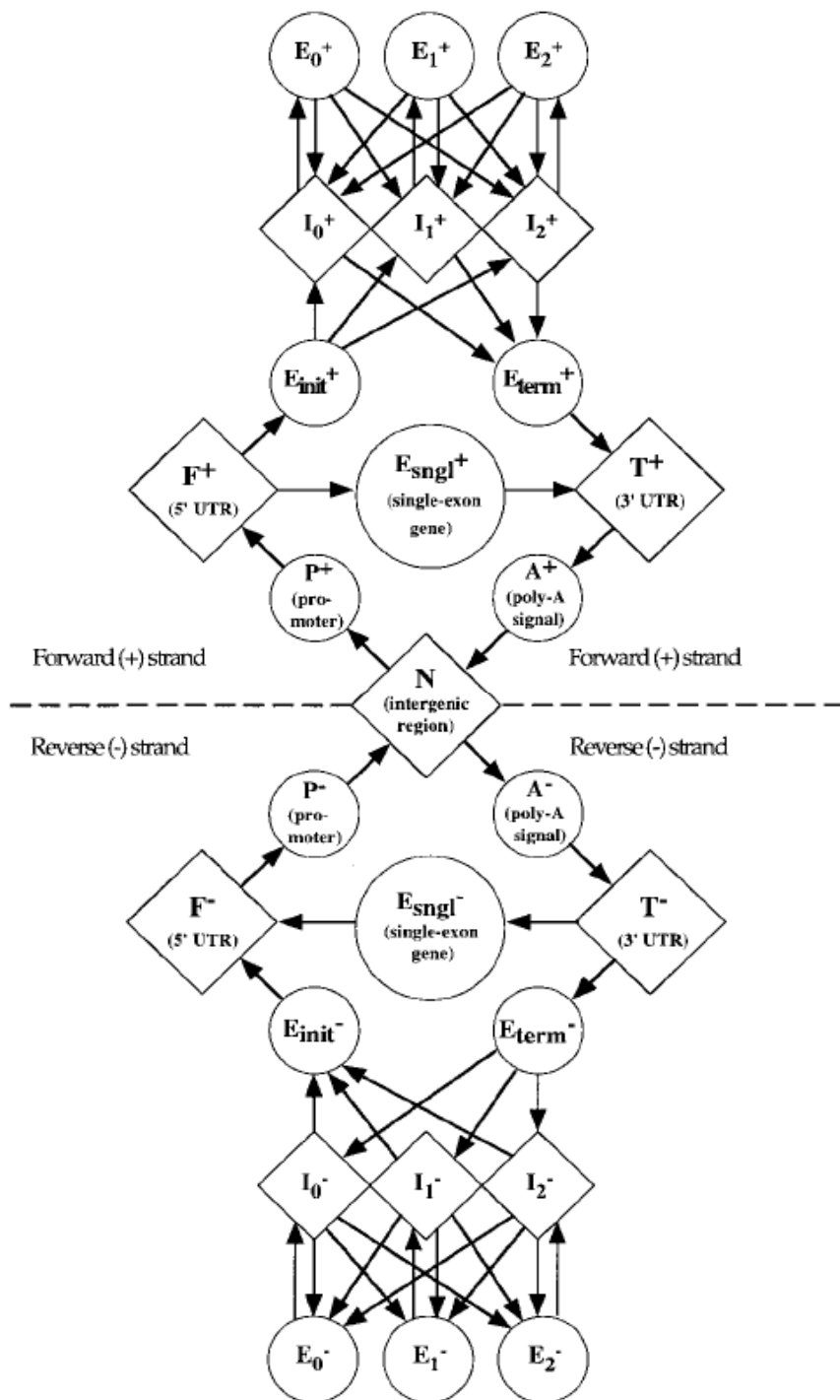
Signal Detection: Splice Sites

Donor

Acceptor



Genscan's View of a Gene



E = Exon
 I = Intron
 A = polyadenylation signal
 P = Promoter
 F, T = UTR
 N = Intergenic sequence

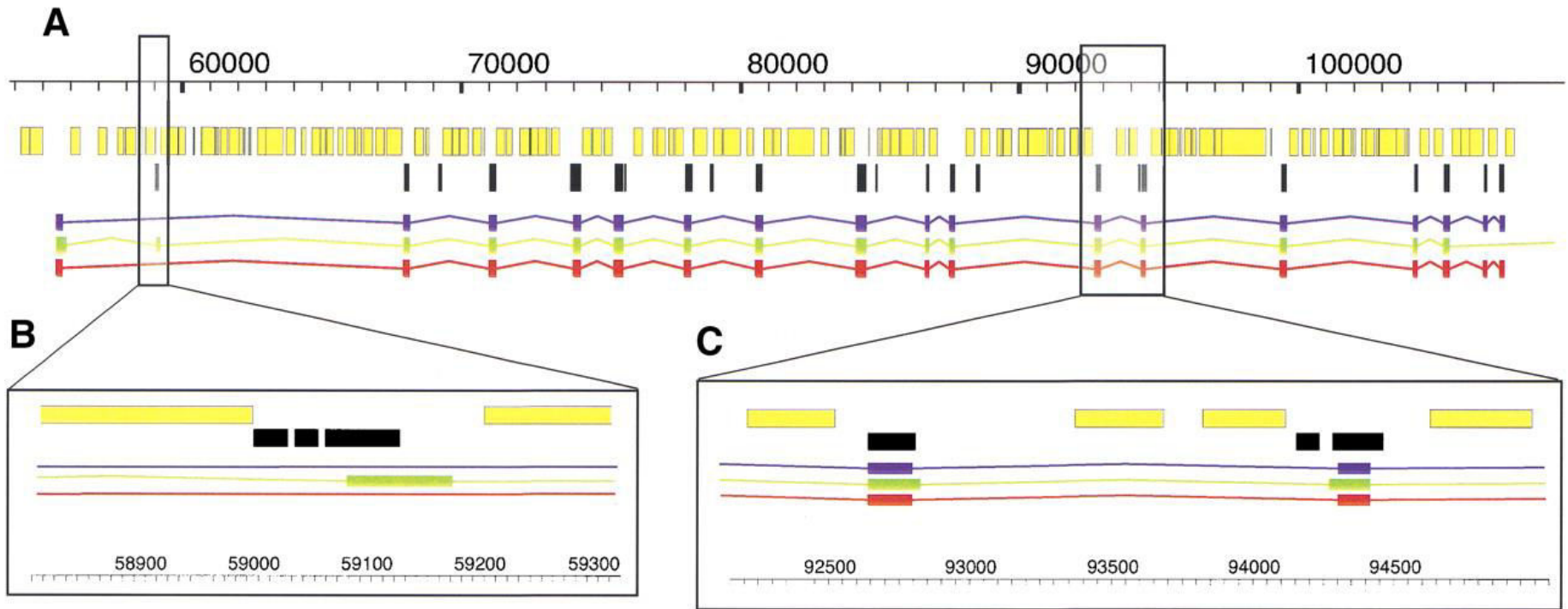
De novo methods: Single genome

- GRAIL
 - Uberbacher & Mural *PNAS* 88:11261, 1991
 - Neural network
 - Trained using known gene structures
- GENSCAN
 - Burge & Karlin *J. Mol. Biol.* 268:78, 1997
 - Generalized hidden Markov model approach
 - Probabilistic model of gene structure
 - Uses descriptions of transcriptional, translational, and splicing signals
 - Distinct parameter sets for varying gene density and structure across G+C ranges
 - Allows for partial genes, multiple genes, and genes on both strands

De novo methods: Dual genome

- Pair HMM approach (SLAM)
 - Joint probability model for sequence alignment and gene structure definition
 - Dynamic programming algorithm combines classic alignment algorithms and HMM decoding
- ‘Informant genome’ approach (SGP-2, TWINSCAN)
 - Alignments performed first (BLASTN, TBLASTX)
 - Alignments ‘inform’ prediction algorithms based on single-genome predictors (e.g. GENSCAN)

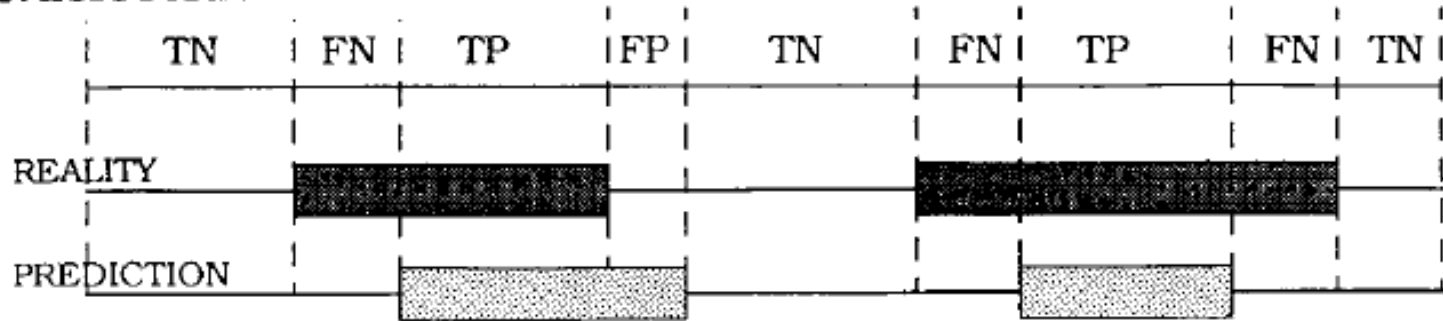
Genscan vs Twinscan



A detailed view of a TWINSKAN prediction (red), a GENSCAN prediction (green), and an aligned RefSeq transcript (blue). Masked repetitive and low-complexity regions (yellow) and mouse alignments (black) are indicated. (A) Complete gene prediction at the *KIAA1630* gene (NM_018706) from *Homo sapiens* 10p14. Note that the presence of conservation is neither a necessary (e.g., the first exon), nor a sufficient (e.g., the first alignment block condition) for TWINSKAN to predict an exon. (B) A magnified region around the second exon predicted by GENSCAN. TWINSKAN correctly omits this exon because the conserved region ends within it. (C) A magnified region around the 11th and 12th RefSeq exons. TWINSKAN correctly predicts both splice sites because they are within the aligned regions.

Evaluation of Predictions

Nucleotide Level



		REALITY		
		coding	no coding	
PREDICTION	coding	TP	FP	TP+FP
	no coding	FN	TN	FN+TN
		TP+FN	FP+TN	

$$S_n = \frac{TP}{TP + FN}$$

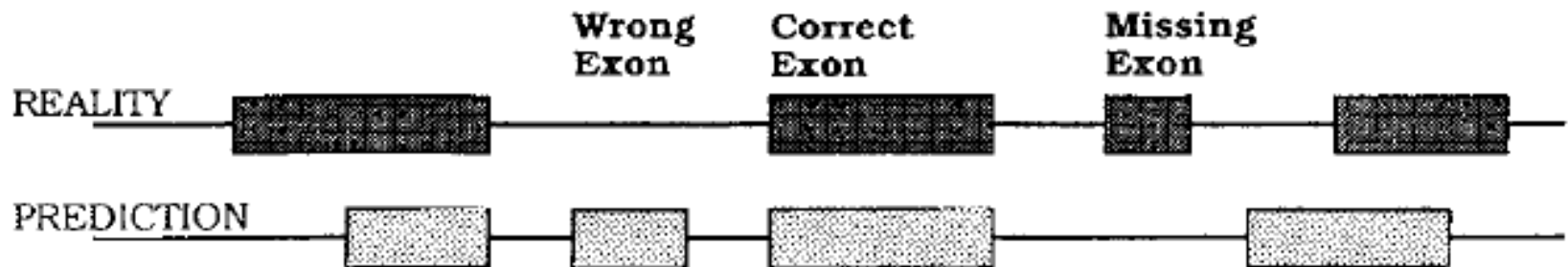
Sensitivity

$$S_p = \frac{TN}{TN + FP}$$

Specificity

Evaluation of Predictions

Exon Level



$$S_n = \frac{\text{number of Correct Exons}}{\text{number of Actual Exons}}$$

Sensitivity

$$S_n = \frac{\text{number of Correct Exons}}{\text{number of Predicted Exons}}$$

Specificity

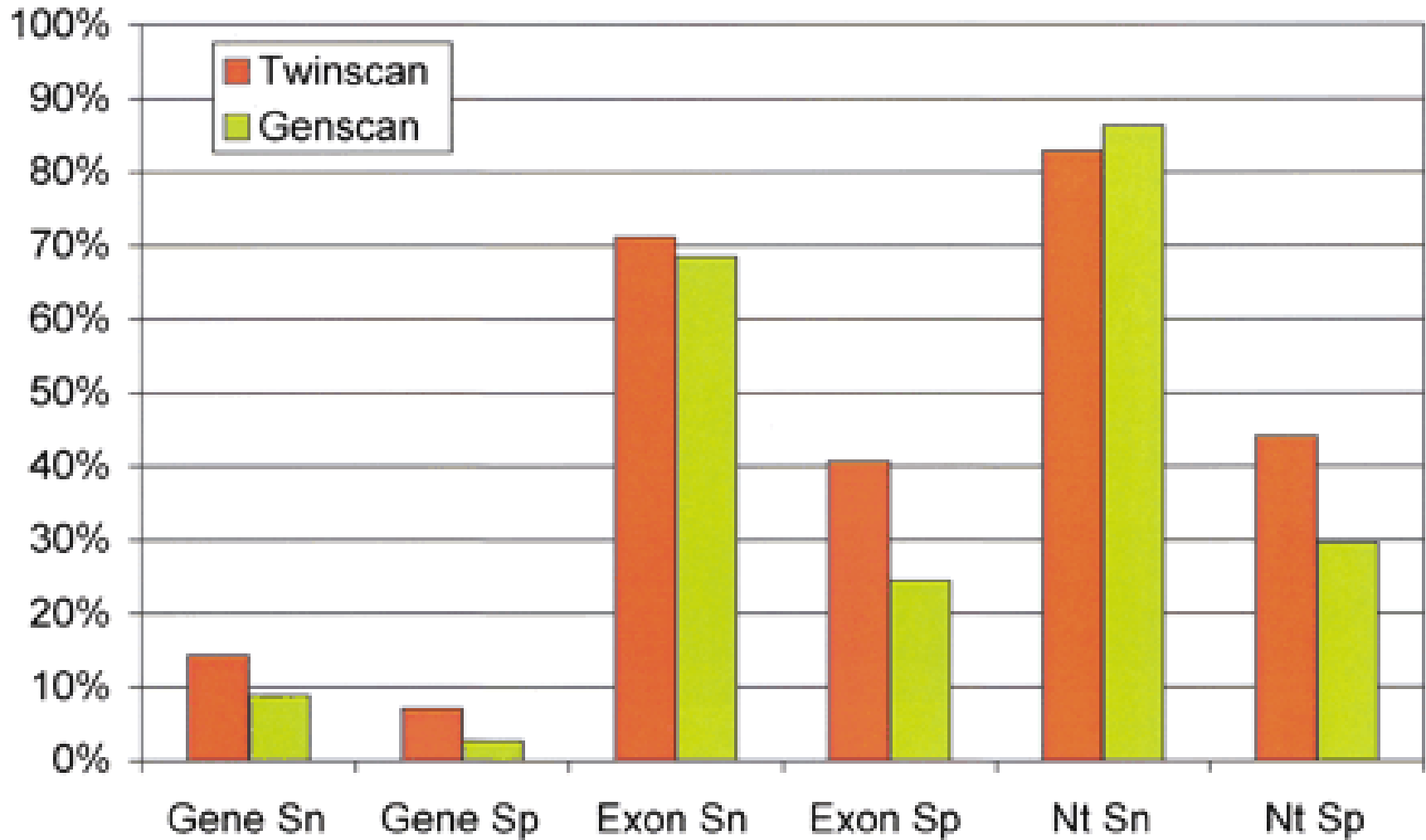
$$ME = \frac{\text{number of Missing Exons}}{\text{number of Actual Exons}}$$

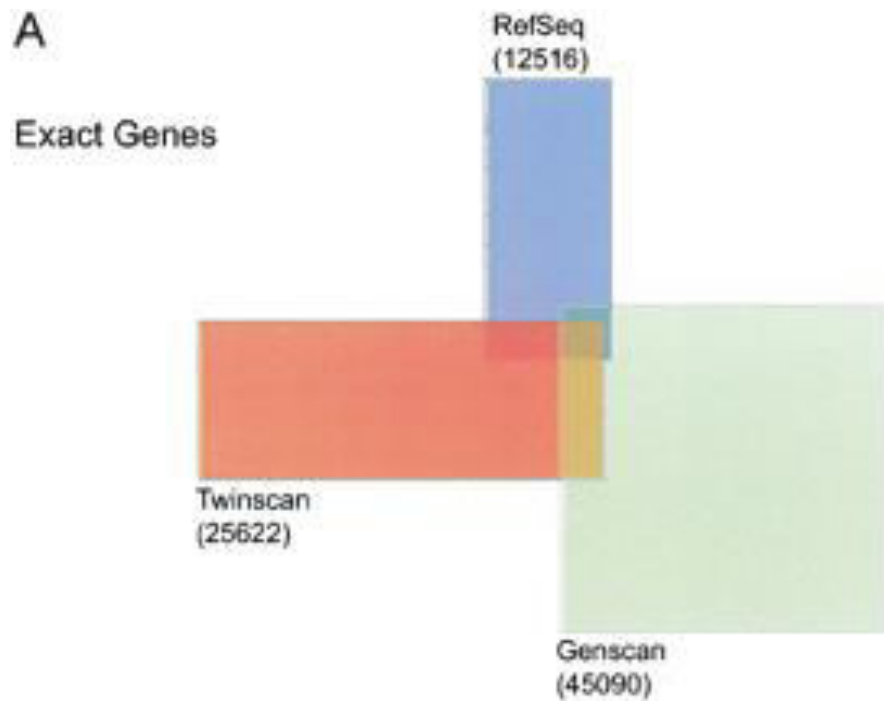
(Sensitivity)

$$WE = \frac{\text{number of Wrong Exons}}{\text{number of Predicted Exons}}$$

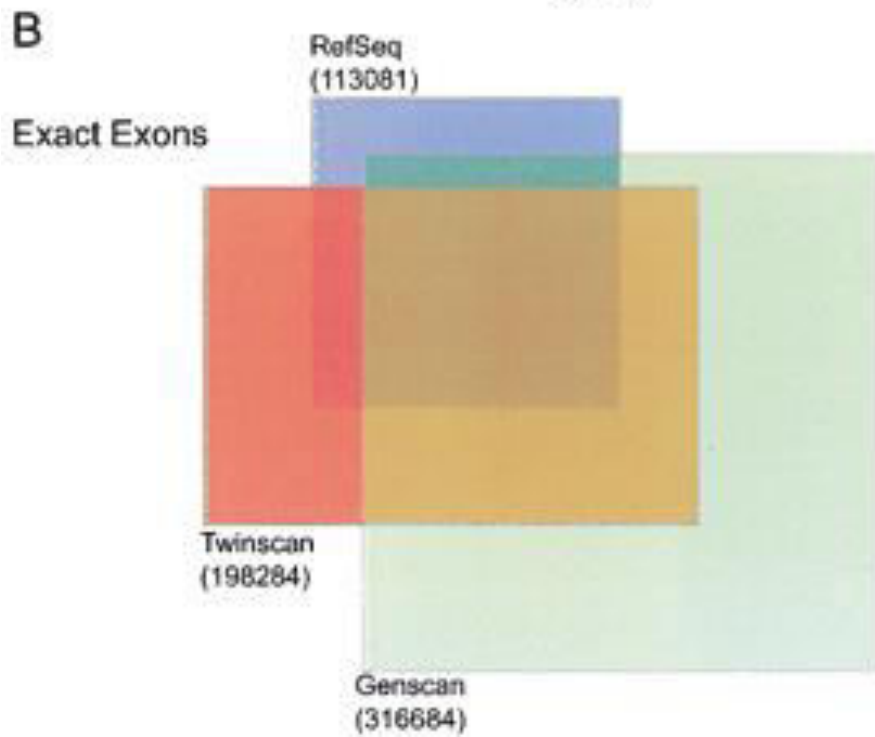
(Specificity)

Annotation of the Mouse Genome





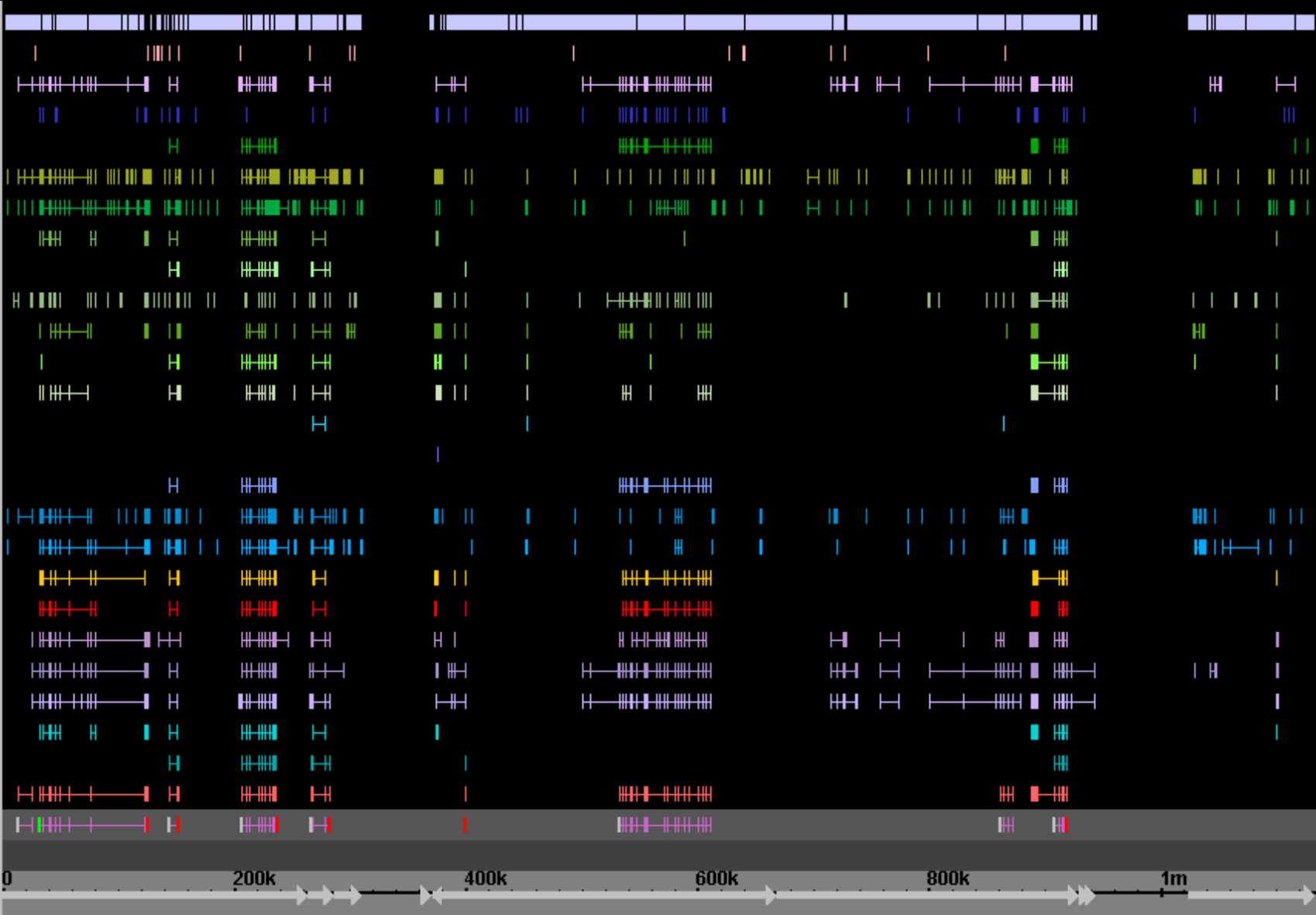
Assessment of Genscan and Twinscan



Addition of Evidence

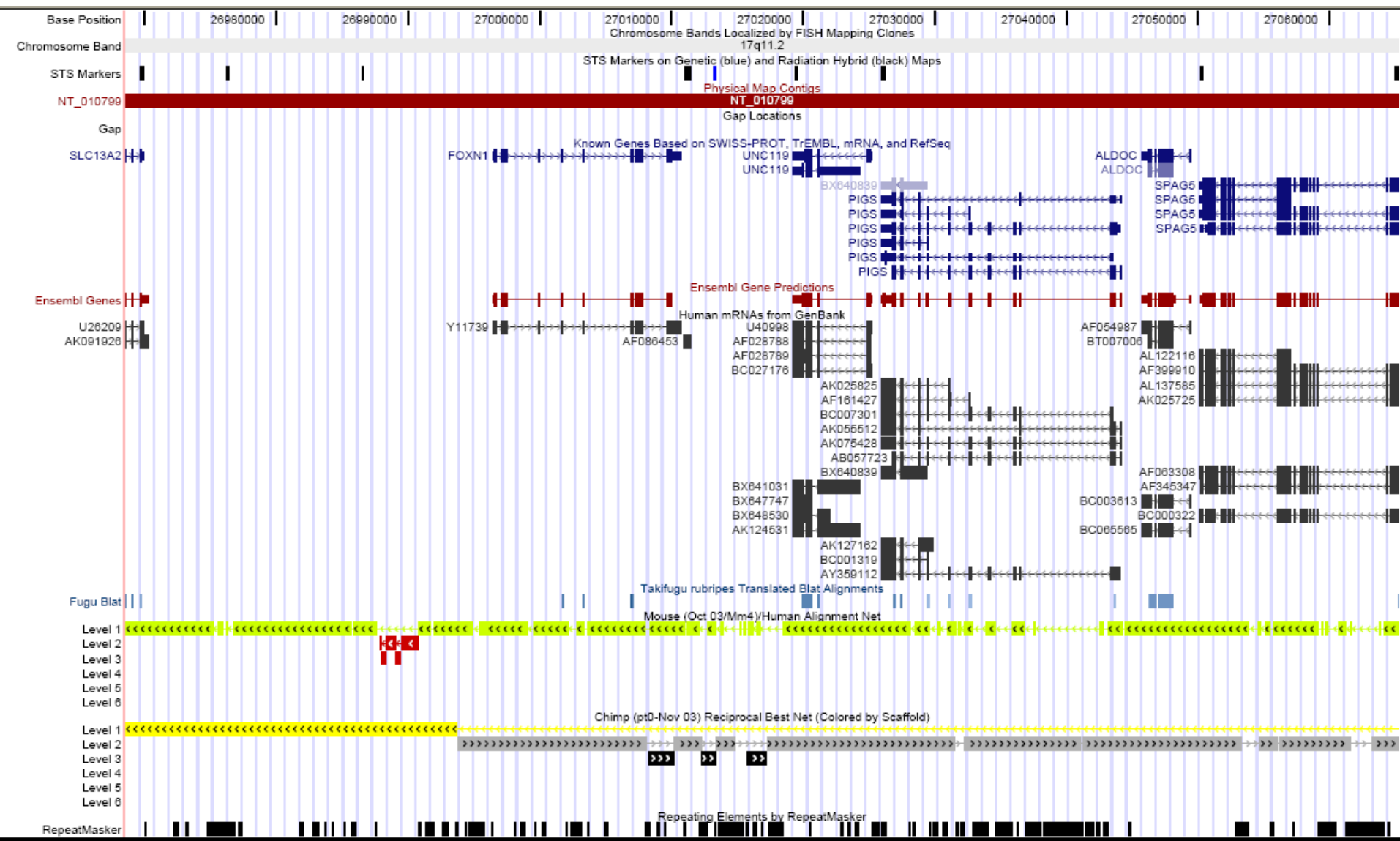
- Known cDNAs
- ESTs (partial cDNA sequence)
- Known genes
- [Predicted genes from other species]
- Genome comparison

- Repeat-masking



repeat
CpG
GenScan
BlastN
Bn:GBCDS
Bn:h_EST
Bn:CHGI
Bn:ensembl
Bn:refSeq
Bn:Mouse
Bn:r_EST
Bn:CRGI
Bn:CMGI
S4:r_EST
S4:CMGI
S4:GBCDS
S4:h_EST
S4:CHGI
BlastX:nraa
LAP:nraa
Grail
FgenesH
TigrCombiner
S4:ensembl
S4:refSeq
Otto
Promoted
Workspace
Axis

Genome Browser



Additional Reading

- Brent & Guigo, Recent advances in gene structure prediction.
Curr. Op. Struct. Biol. 14(3) 264-272, 2004
- Fickett, JW. The gene identification problem: an overview for developers.
Comput. Chem. 20:103-118, 1996