Eukaryotic Gene Prediction

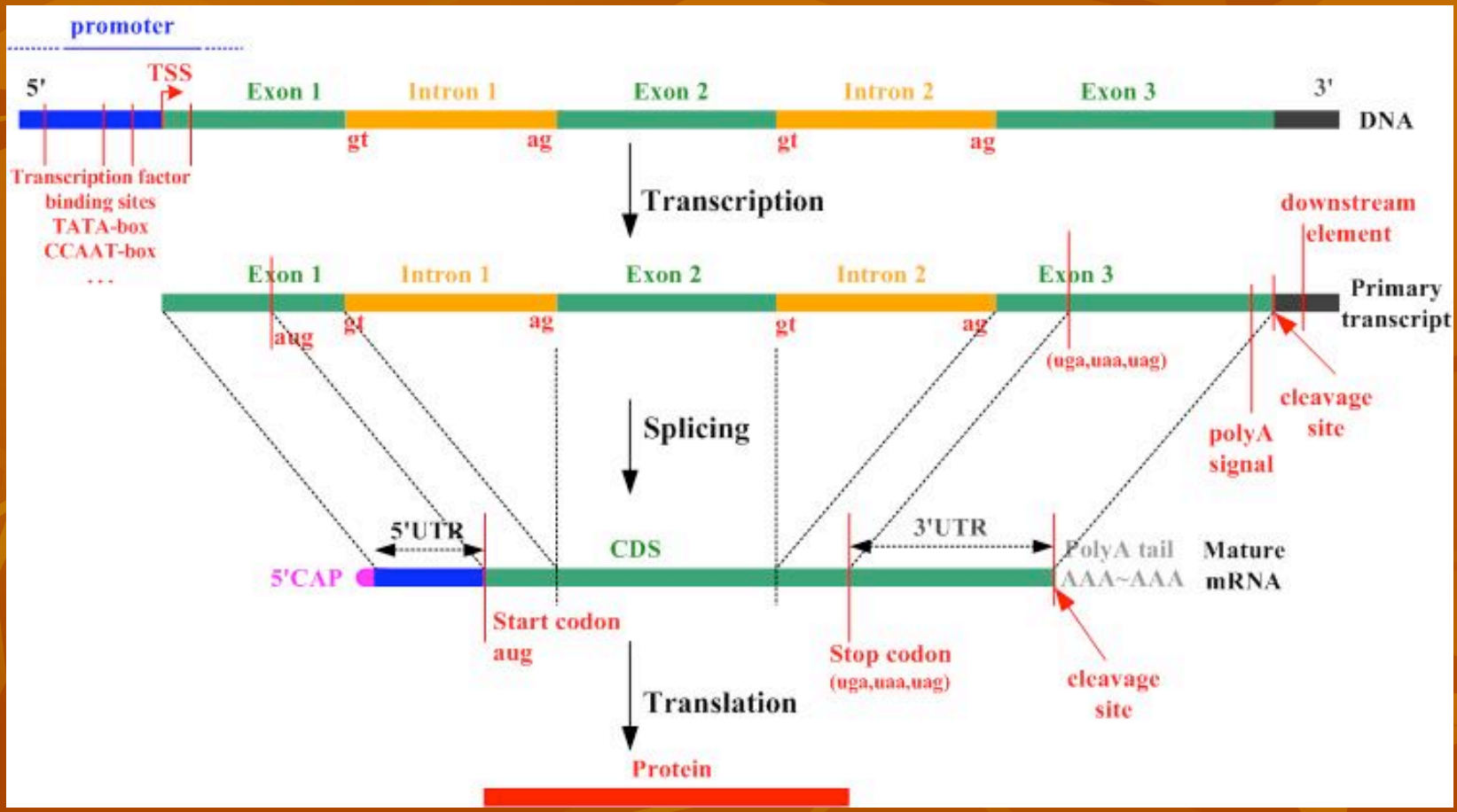
Wei Zhu

May 2007

“In nature, nothing is perfect ...”

- Alice Walker

Gene Structure



What is Gene Prediction?

Gene prediction is the problem of parsing a sequence into nonoverlapping *coding segments* (CDSs) consisting of exons separated by introns.

Signal Sensors

A *signal sensor* evaluates fixed-length features in DNA.

- Start codons
- Stop codons
- Donor sites
- Acceptor sites
- Promoters
- Poly-A signals

Content Sensors

A *content sensor* evaluates variable-length features which extend from one signal to another:

- Exons
- Introns
- Intergenic regions
- UTRs

Gene Prediction Approaches

- Intrinsic (*ab initio*)
  - GENSCAN, FGENESH, GeneMark.hmm GlimmerM, Genie;
- Extrinsic (similarity-based)
  - Spliced alignment: GenomeScan, EuGene, FGENESH+, FGENESH_C, GeneId+, etc;
  - Genomic comparison: TwinScan, TWAIN, SLAM, SGP, FGENESH_2, etc;
- Integrated
  - GeneScope, GeneMachine, JIGSAW, RiceGAAS, Ensembl, EVM etc.

ab initio Gene Prediction

- Adopt a rigorous probabilistic model of sequence structure and choose the most probable parse according to that probabilistic model.
- Pros
  - Fast and efficient
  - Remarkable accuracy at the nucleotide level
- Cons
  - Less than 50% accuracy at the gene level

Development of a Gene Finder

- Build the model
- Train the model to generate the related parameters
- Predict/Evaluate

Imperfect Model

GT……..AG

…A|G|T…

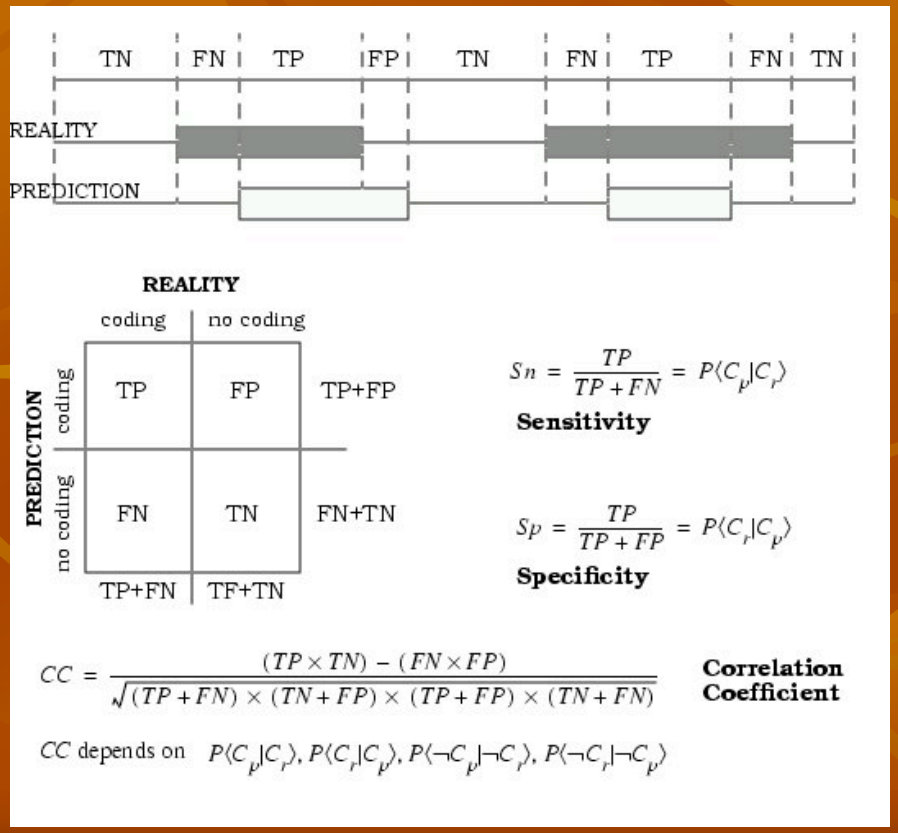
1-bp intron

Accuracy Evaluation

- Nucleotide level
- Exon level
- Gene level

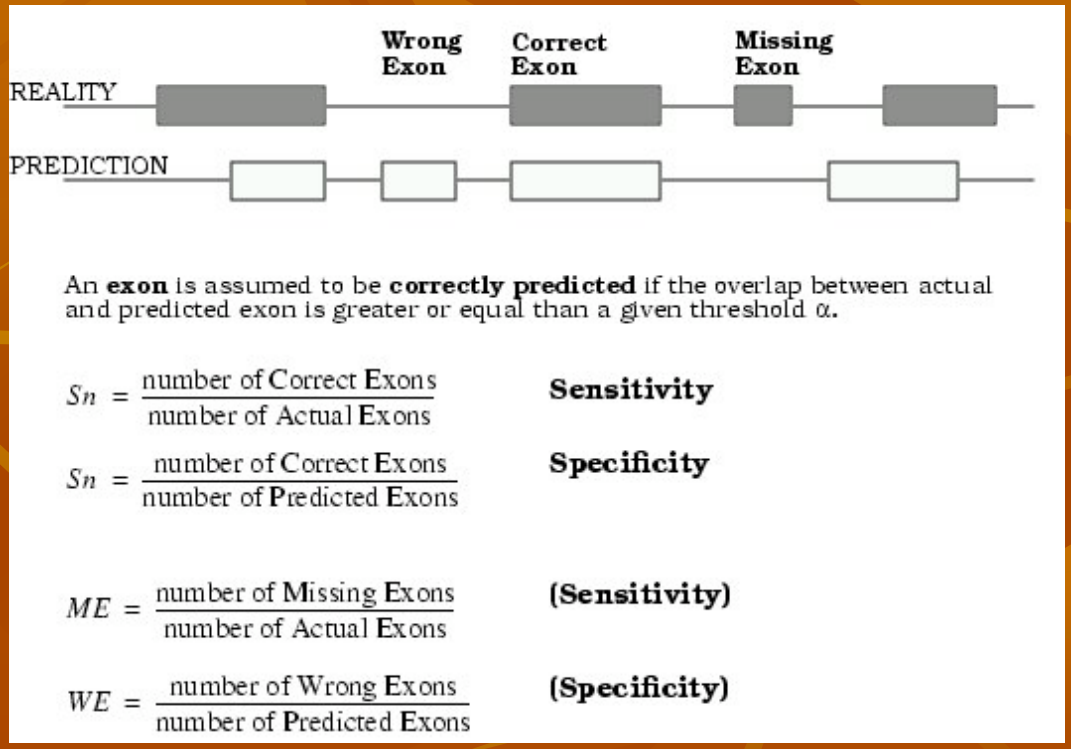
Nucleotide/Base Level

- **Prediction accuracy per base coding/non-coding**



Exon Level

- **Prediction accuracy with respect to exact prediction of exon start and end points**



An **exon** is assumed to be **correctly predicted** if the overlap between actual and predicted exon is greater or equal than a given threshold α.

$$Sn = \frac{\text{number of Correct Exons}}{\text{number of Actual Exons}}$$ **Sensitivity**

$$Sn = \frac{\text{number of Correct Exons}}{\text{number of Predicted Exons}}$$ **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)**

Gene/Protein Level

Prediction accuracy with respect to the protein product encoded by the predicted gene

A Simple Calculation

Given x accuracy at exon level, the accuracy of the prediction at the gene level is:

$$P = P \text{ (all exons correctly predicted)} = x^n ,$$

where n is the number of exons in the gene.

Typically, $x<90\%$ and $n=5$, then

$$P = 0.9 \times 0.9 \times 0.9 \times 0.9 \times 0.9 = 59\%$$

Performance

- Species-specific setting
  - GC content
  - Gene density
  - Gene/Exon/Intron length distribution
  - Codon usage
- Benchmark
  - training data set
  - test data set

Maize Gene Prediction

Plant Molecular Biology (2005) 57:445–460
DOI: 10.1007/s11103-005-0271-1

Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes

Hong Yao[1,4], Ling Guo[1,6,†], Yan Fu[1,4,†], Lisa A. Borsuk[1,6], Tsui-Jung Wen[2], David S. Skibbe[1,5], Xiangqin Cui[1,4,9], Brian E. Scheffler[8], Jun Cao[1,4], Scott J. Emrich[6], Daniel A. Ashlock[3,6] and Patrick S. Schnable[1,2,4–7,*]

[1]*Department of Genetics, Development, and Cell Biology (*author for correspondence; e-mail schnable@iastate.edu);* [2]*Department of Agronomy;* [3]*Department of Mathematics;* [4]*Inderdepartmental Graduate Programs in Genetics;* [5]*Department of Molecular, Cellular and Developmental Biology;* [6]*Department of Electrical and Computer Engineering and Department of Bioinformatics and Computational Biology;* [7]*Center for Plant Genomics, Iowa State University, Ames, Iowa 50011-3650;* [8]*USDA-ARS, Mid South Area Genomics Facility, Stoneville, MS 38776-0038, USA;* [9]*Present address: Department of Biostatistics, Birmingham, AL 35294, USA;* [†]*these authors contributed equally to this work*

Received 16 August 2004, accepted in revised form 6 January 2005

Gene Finders

Programs	Websites	Trained organisms	Type of prediction			Algorithm models
			Splice site	Exon	Gene model	
FGENESH	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	Monocots	Yes	Yes	Yes	GHMM[a]
GeneMark.hmm	http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi?org=H.sapiens	Maize	Yes	Yes	Yes	GHMM
GENSCAN	http://genes.mit.edu/GENSCAN.html	Maize	Yes	Yes	Yes	GHMM
GlimmerR	http://www.tigr.org/tdb/glimmerm/glmr_form.html	Rice	Yes	Yes	Yes	IMM[b]
Grail	http://compbio.ornl.gov/Grail-1.3/	*Arabidopsis*	Yes	Yes	No	neural networks

[a]GHMM, Generalized Hidden Markov Model.
[b]IMM, Interpolated Markov Model.

Accuracy

Programs	Nucleotide level			Exon level						
	SN	SP	CC	SN	SP	(SN + SP)/2	PE%	OE%	ME%	WE%
FGENESH	0.97	0.94	0.93	0.86	0.88	0.87	9.4	0	4.6	3.1
GeneMark.hmm	0.92	0.93	0.89	0.69	0.80	0.75	14	0	19	5.4
GENSCAN	0.81	0.95	0.82	0.54	0.81	0.68	12	0	39	7.0
GlimmerR	0.70	0.91	0.71	0.51	0.64	0.57	23	5.8	23	7.7
Grail	0.55	0.67	0.43	0.34	0.28	0.31	33	7.7	17	31

Challenges of Intrinsic Approaches

- Alternative splicing
- Nested/overlapped genes
- Extremely long/short genes
- Extremely long introns
- Extremely short exons
- Non-canonical introns
- Frame-shift errors
- Split start codons (that is, the start codon is split by an intron in the genomic sequence)
- UTR introns
- Non-ATG triplet as the start codon
- Polycistronic genes

Gene Prediction Approaches

- Intrinsic (*ab initio*)
  - GENSCAN, FGENESH, GeneMark.hmm GlimmerM, Genie;
- Extrinsic (similarity-based)
  - Spliced alignment: GenomeScan, EuGene, FGENESH+, FGENESH_C, GeneId+, etc;
  - Genomic comparison: TwinScan, TWAIN, SLAM, SGP, FGENESH_2, etc;
- Integrated
  - GeneScope, GeneMachine, JIGSAW, RiceGAAS, Ensembl, etc.

Similarity-based Gene Prediction

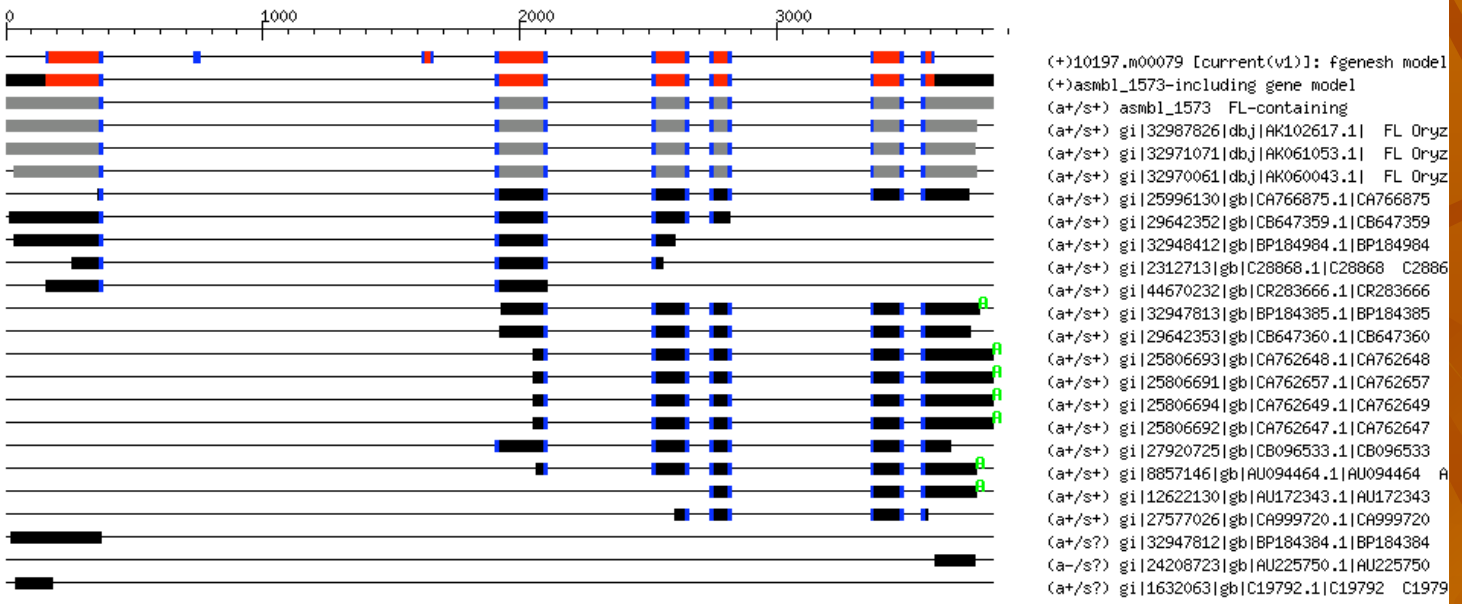
- EST/cDNA spliced alignment
- Protein spliced alignment
- Genomic comparison
  - Intra-genomic
  - Inter-genomic

EST/cDNA Spliced Alignment

Report for cDNA subcluster: 1258

of cluster: 20541 (annotdb_asmbl_id:10197 coords:116784-120621)

Subcluster view.



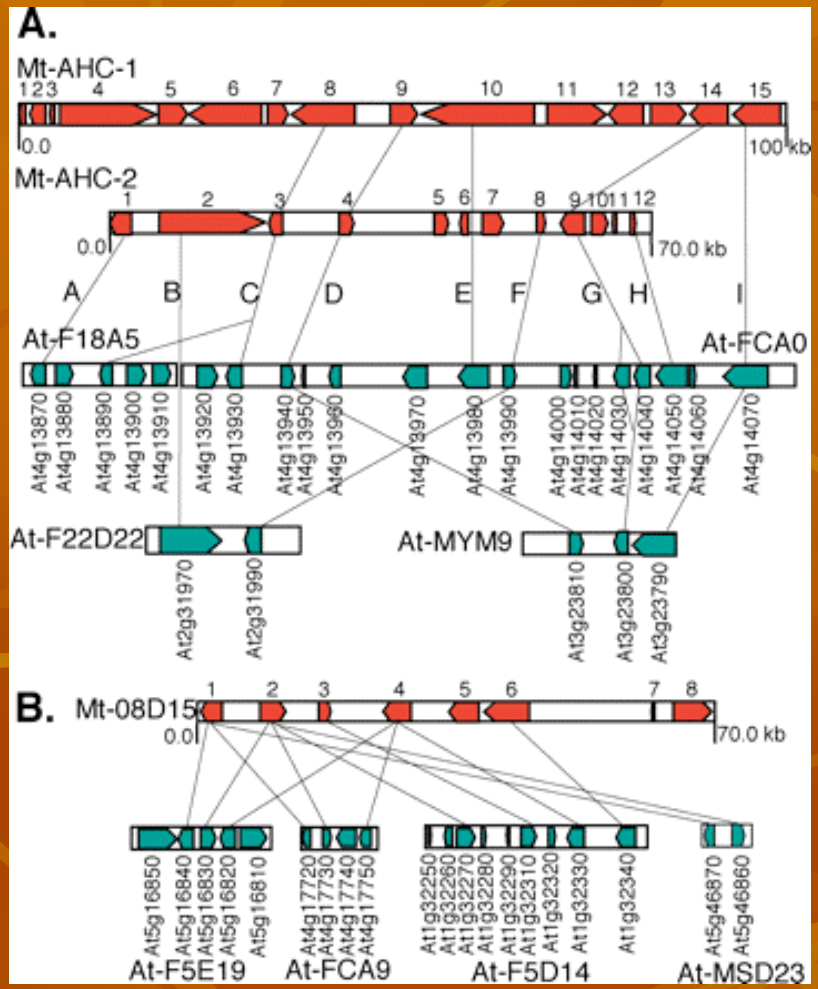
Assembly description

assembly	cdnas	annotations linked	status	
	gi|12622130|gb|AU172343.1|AU172343			

Pros and Cons

- Pros
  - High accuracy
- Cons
  - Unavailability or incompleteness of transcript sequence data
  - Extra computation to generate alignments
  - Diverse sequence quality
    - Incomplete full-length cDNA
    - Contamination
    - Incorrect sequence orientations

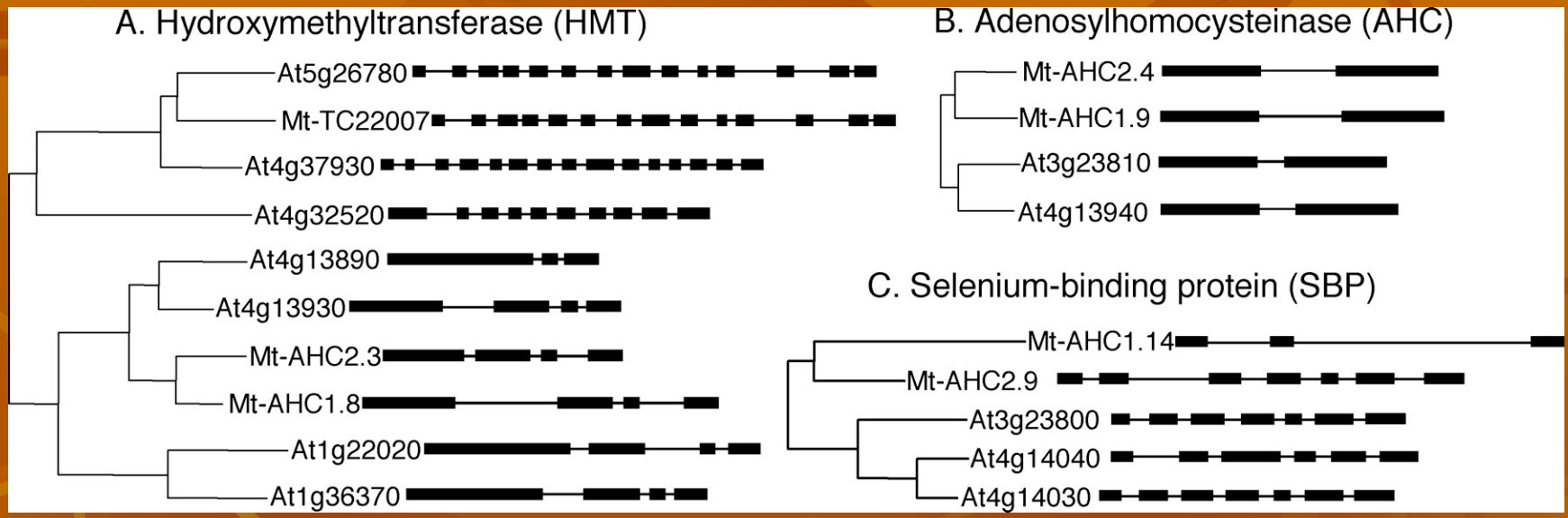
Genomic Comparison



Microsynteny between *M. truncatula* and Arabidopsis

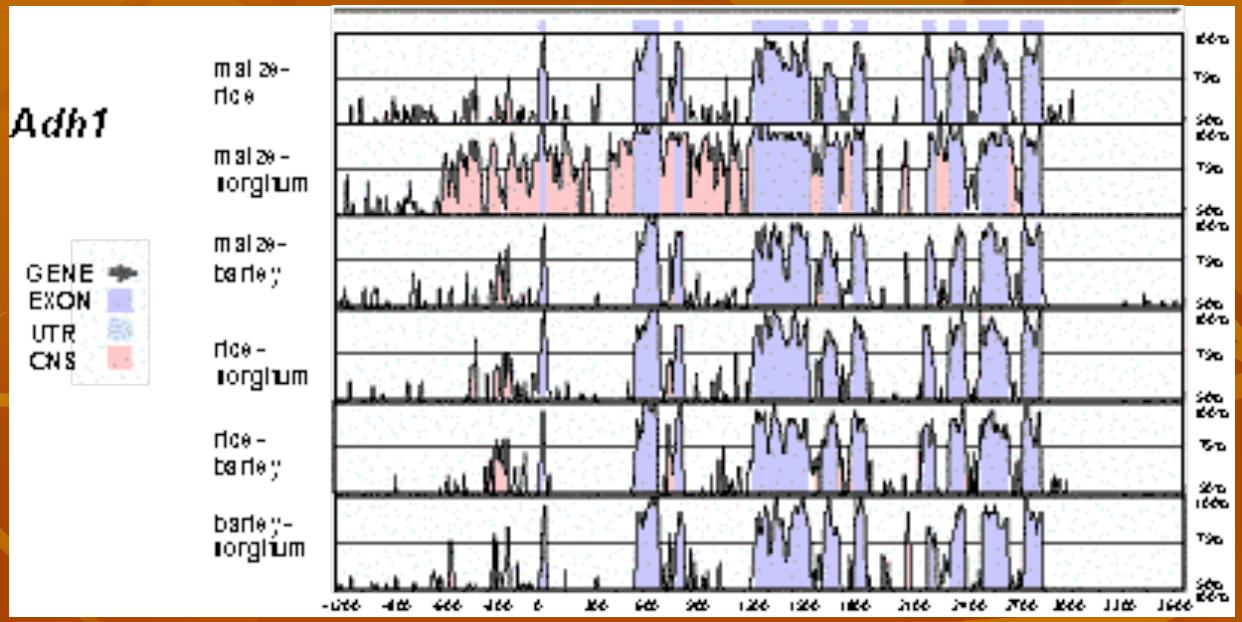
Hongyan et al, 2003

Gene Structure of Syntenic and non-Syntenic Homologous Genes

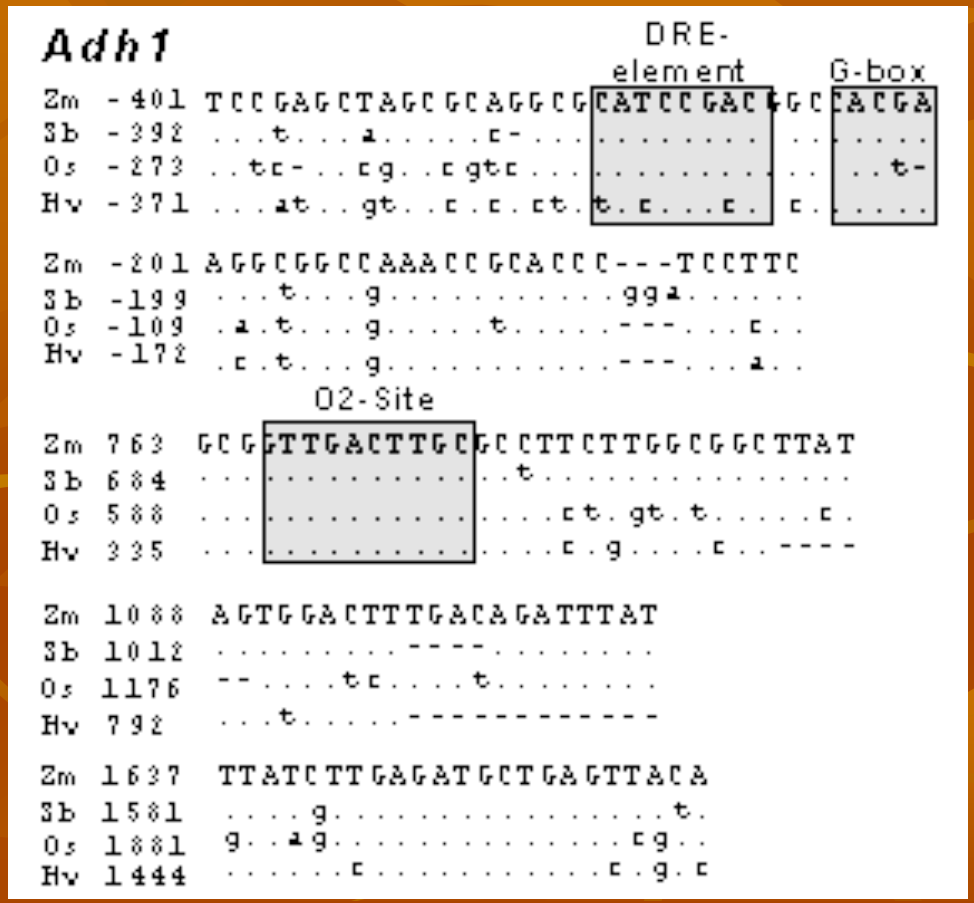


Hongyan et al, 2003

Comparative Analysis of Cereal Gene Structures



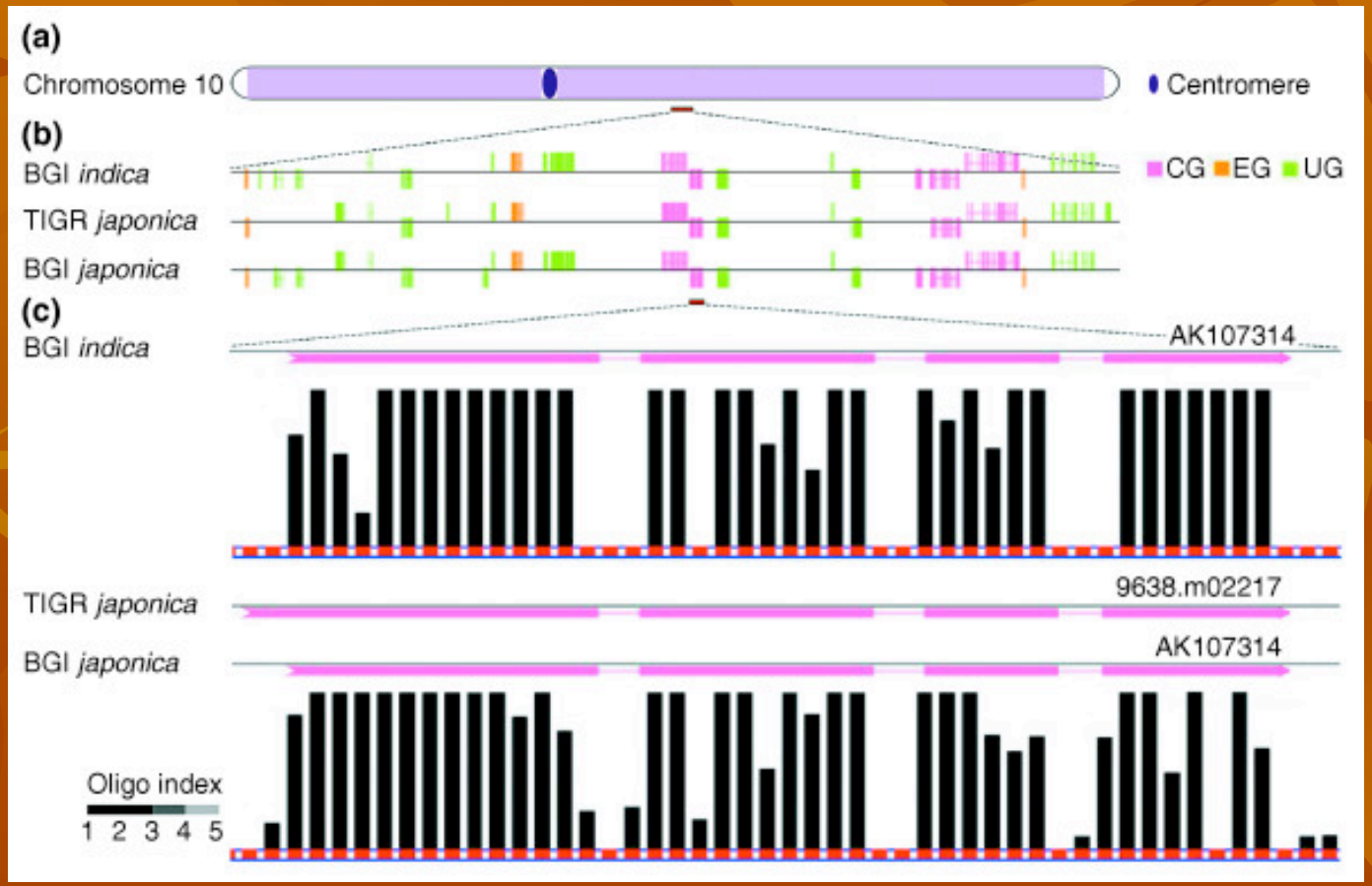
Comparative Analysis of Cereal Gene Promoters



Pros and Cons

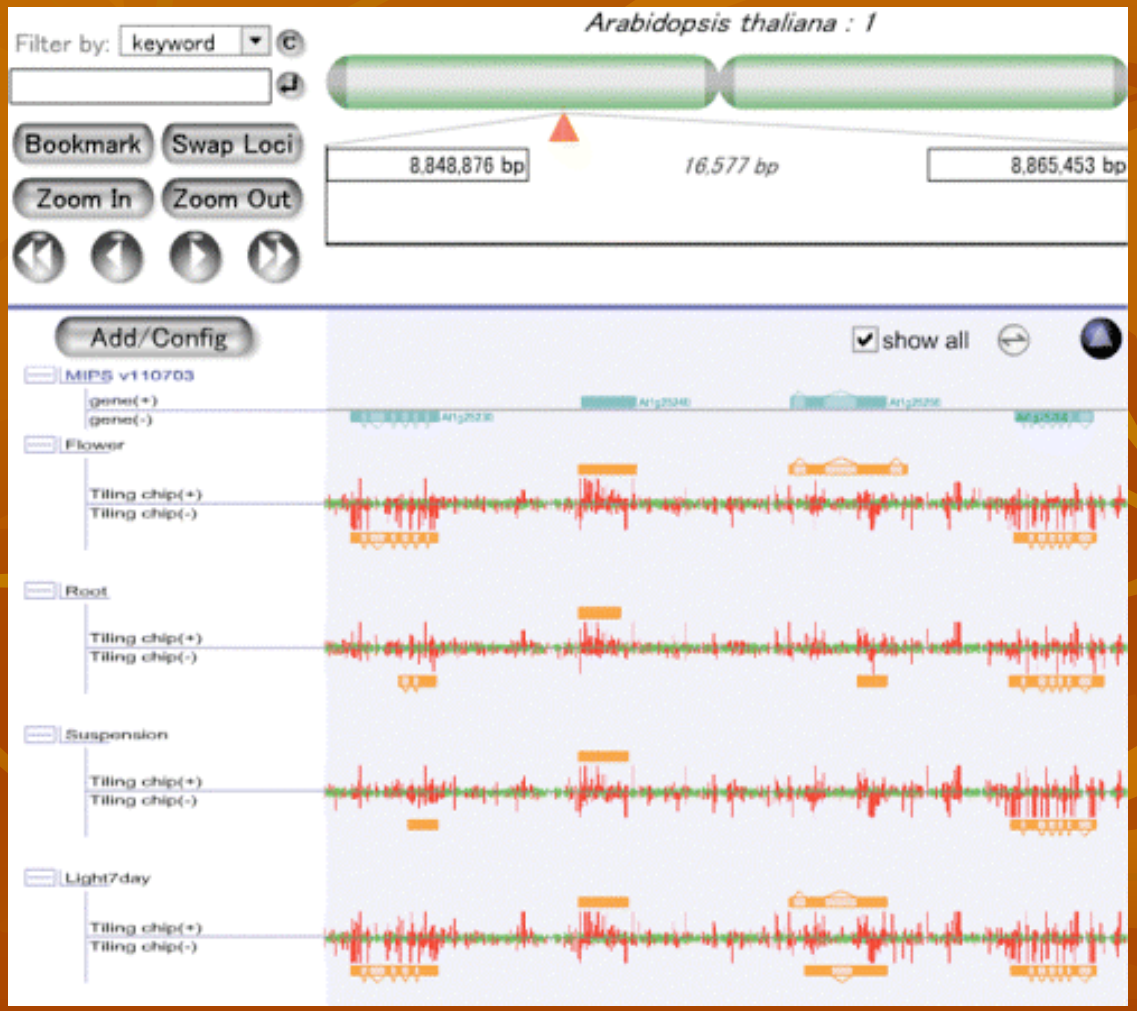
- Pros
  - Aid to identify low expressed genes
  - Identify genes in multiple species simultaneously
  - Aid to identify transcription factor binding sites
  - Uncover non-protein coding genes
- Cons
  - Performance will depend on the evolutionary distance between the compared sequences.
  - Exon/intron boundaries may not be conserved

Tiling Array



ARTADE

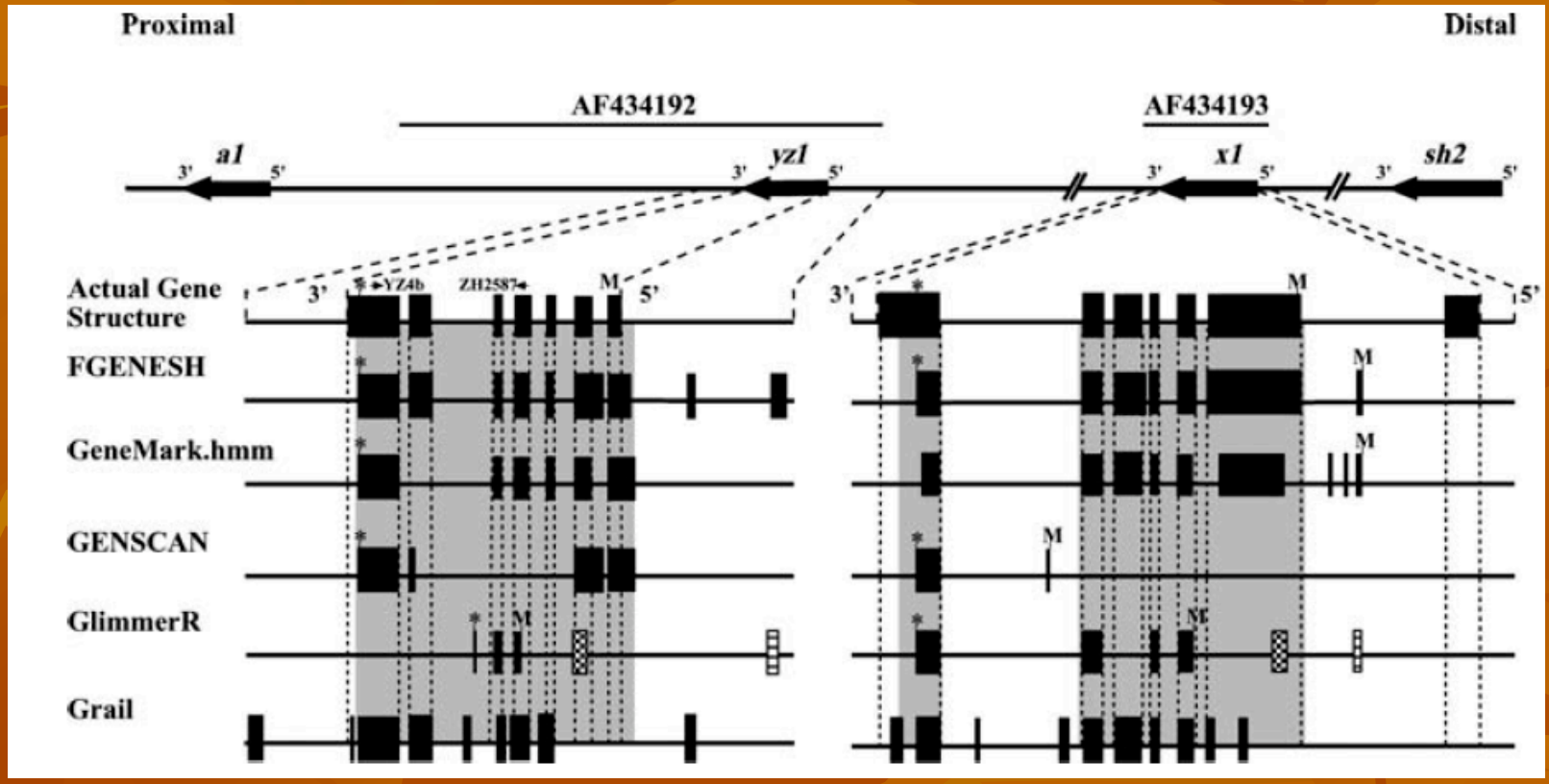
-ARabidopsis Tiling-Array-based Detection of Exons



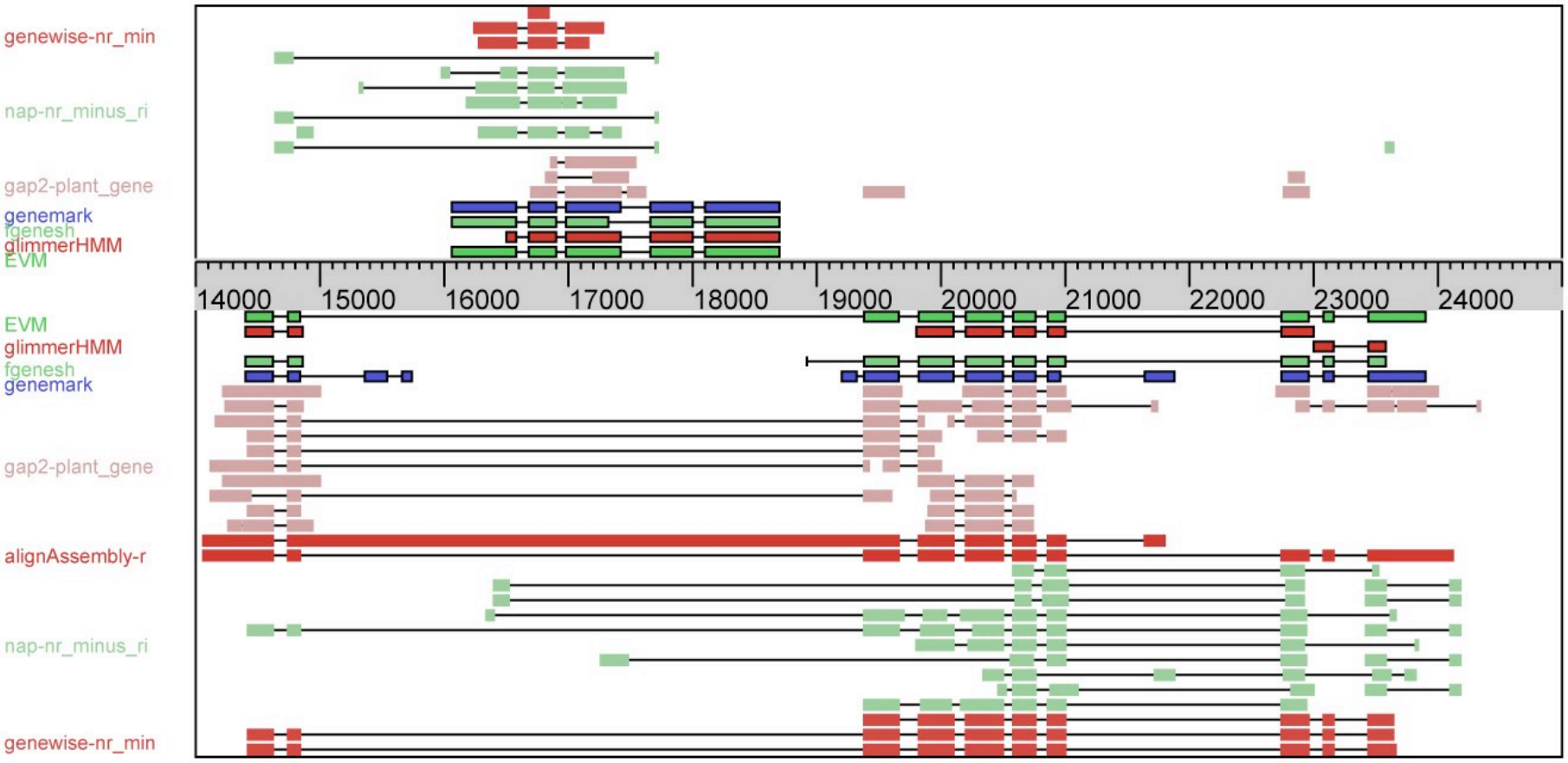
Gene Prediction Approaches

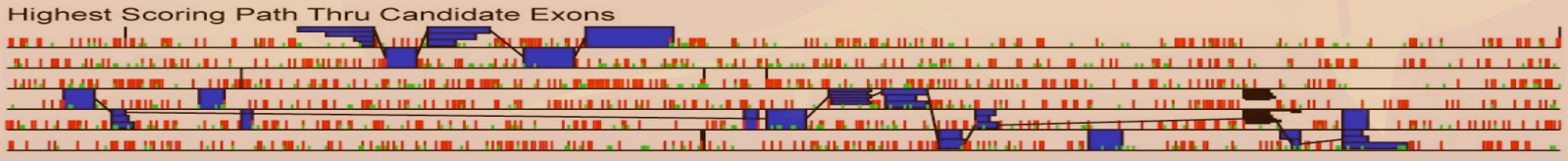
- Intrinsic (*ab initio*)
  - GENSCAN, FGENESH, GeneMark.hmm GlimmerM, Genie;
- Extrinsic (similarity-based)
  - Spliced alignment: GenomeScan, EuGene, FGENESH+, FGENESH_C, GeneId+, etc;
  - Genomic comparison: TwinScan, TWAIN, SLAM, SGP, FGENESH_2, etc;
- Integrated
  - GeneScope, GeneMachine, JIGSAW (combiner), RiceGAAS, Ensembl, etc.

Gene Discovery via Multiple Gene Finders

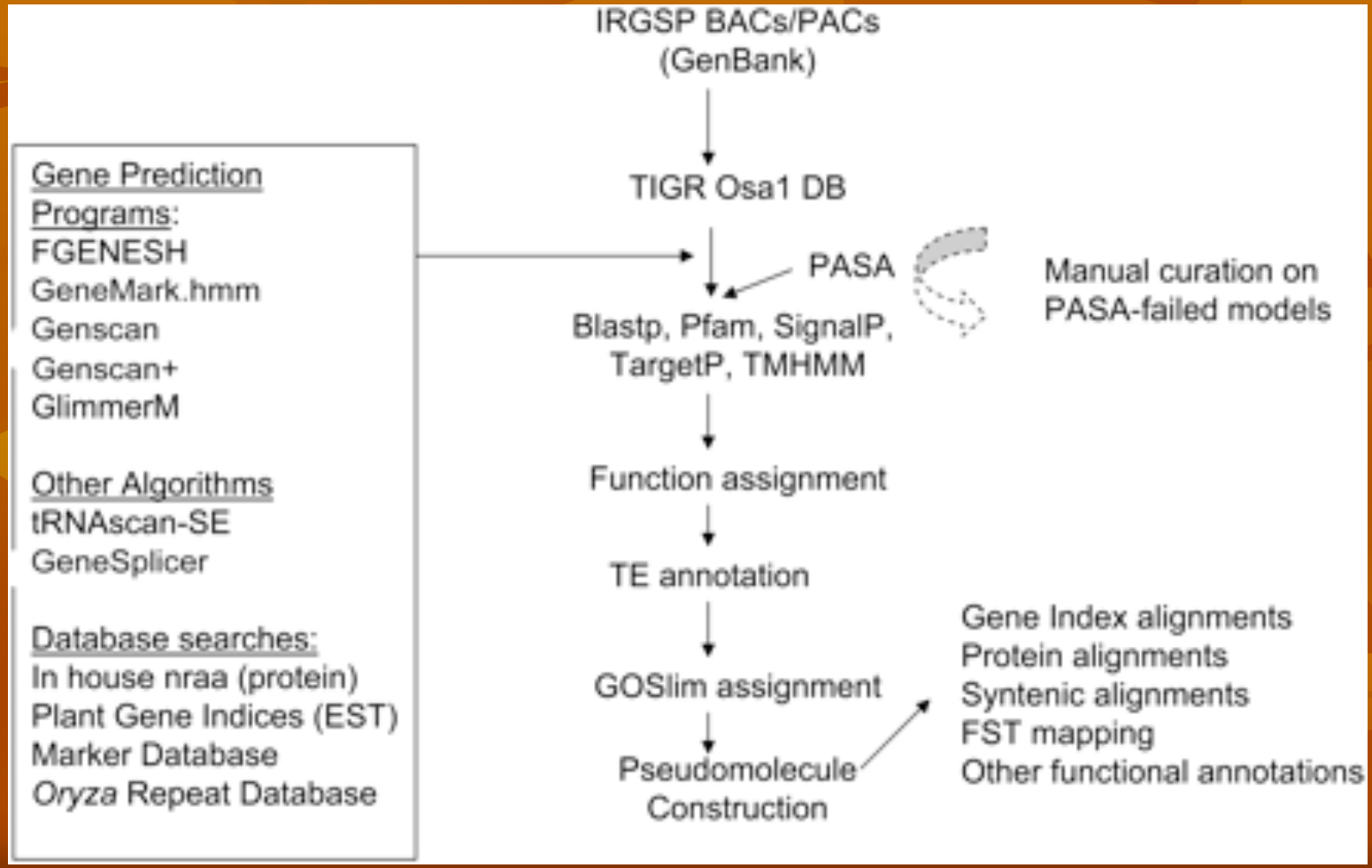


EVM

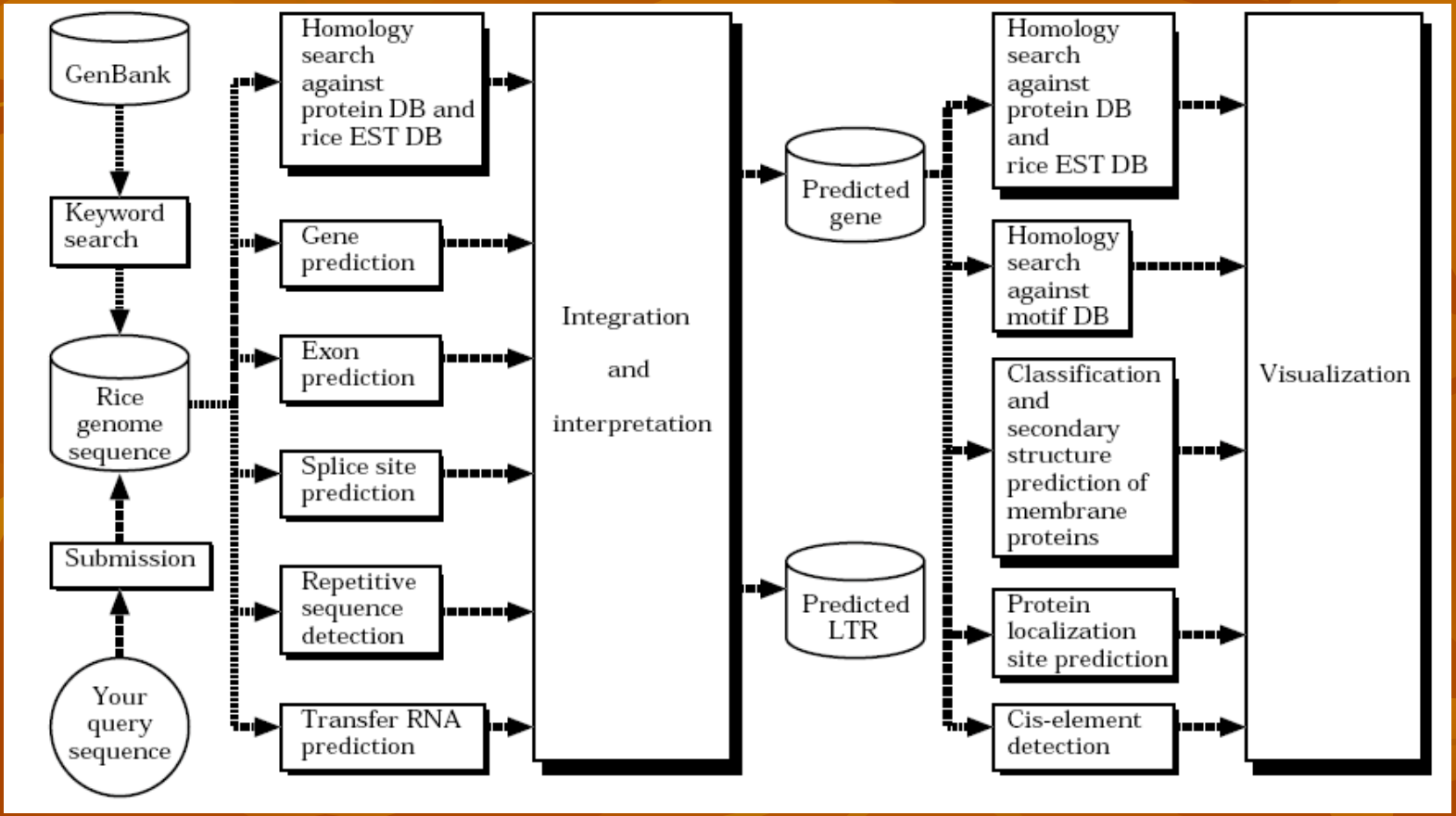

genewise-nr_min
nap-nr_minus_ri
gap2-plant_gene
genemark
fgenesh
glimmerHMM
EVM
14000
15000
16000
17000
18000
19000
20000
21000
22000
23000
24000
EVM
glimmerHMM
fgenesh
genemark
gap2-plant_gene
alignAssembly-r
nap-nr_minus_ri
genewise-nr_min


Highest Scoring Path Thru Candidate Exons

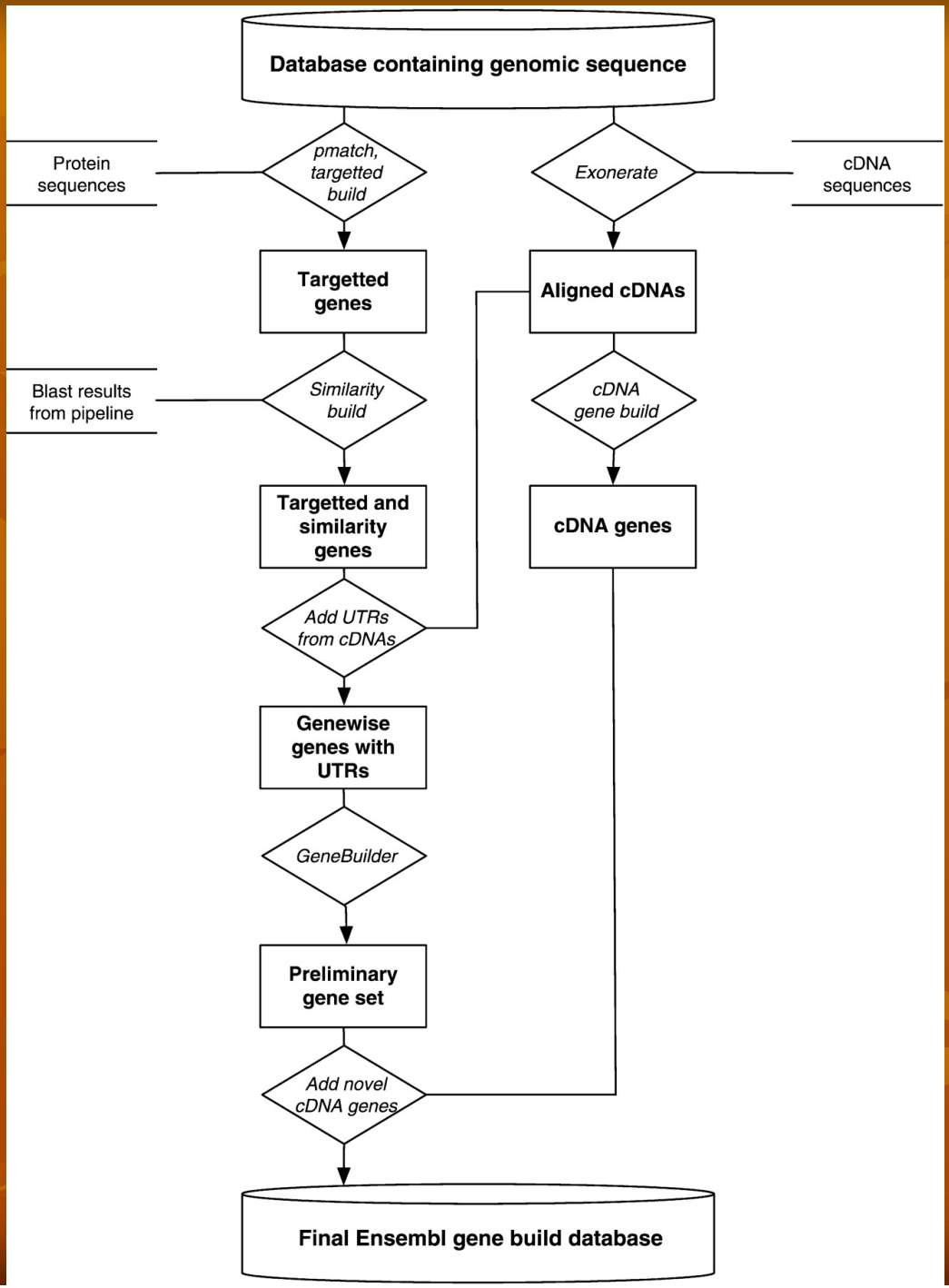
TIGR Rice Genome Annotation Pipeline



RiceGAAC


GenBank
Keyword search
Rice genome sequence
Submission
Your query sequence
Homology search against protein DB and rice EST DB
Gene prediction
Exon prediction
Splice site prediction
Repetitive sequence detection
Transfer RNA prediction
Integration and interpretation
Predicted gene
Predicted LTR
Homology search against protein DB and rice EST DB
Homology search against motif DB
Classification and secondary structure prediction of membrane proteins
Protein localization site prediction
Cis-element detection
Visualization

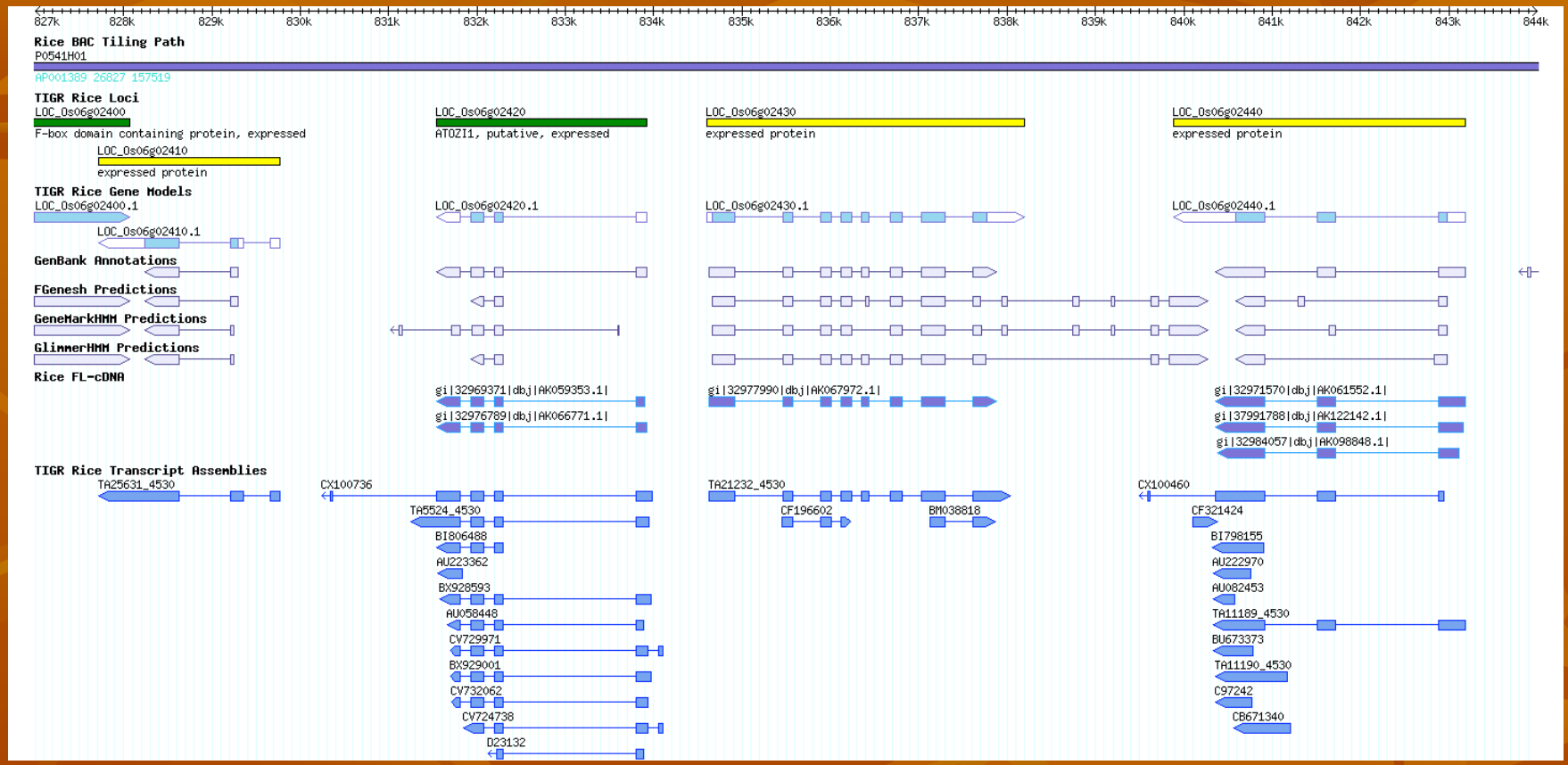
Ensembl Gene Prediction Procedure



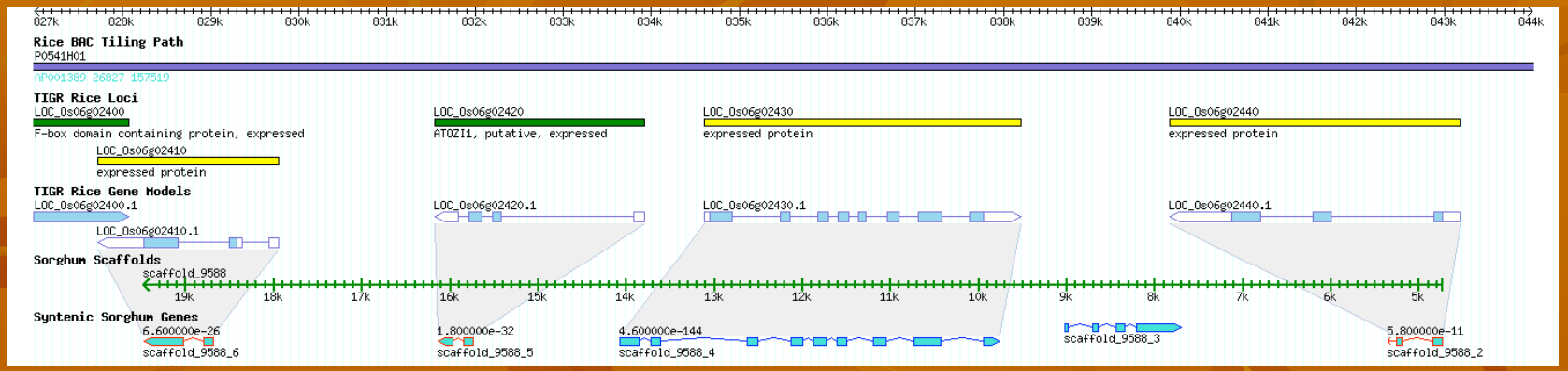
Summary

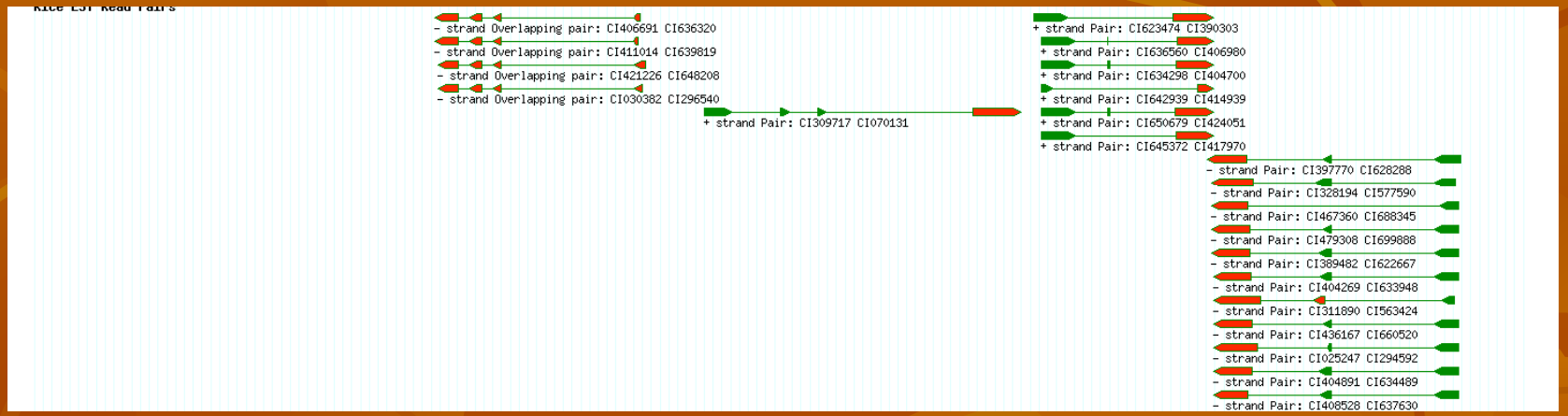
- Nothing is perfect
- Each gene identification approach has its own features and limitations;
- Genome annotation is an on-going process, and the accuracy is being improved along with the accumulation of the evidence data;tRNAsnoRNA

Case Study

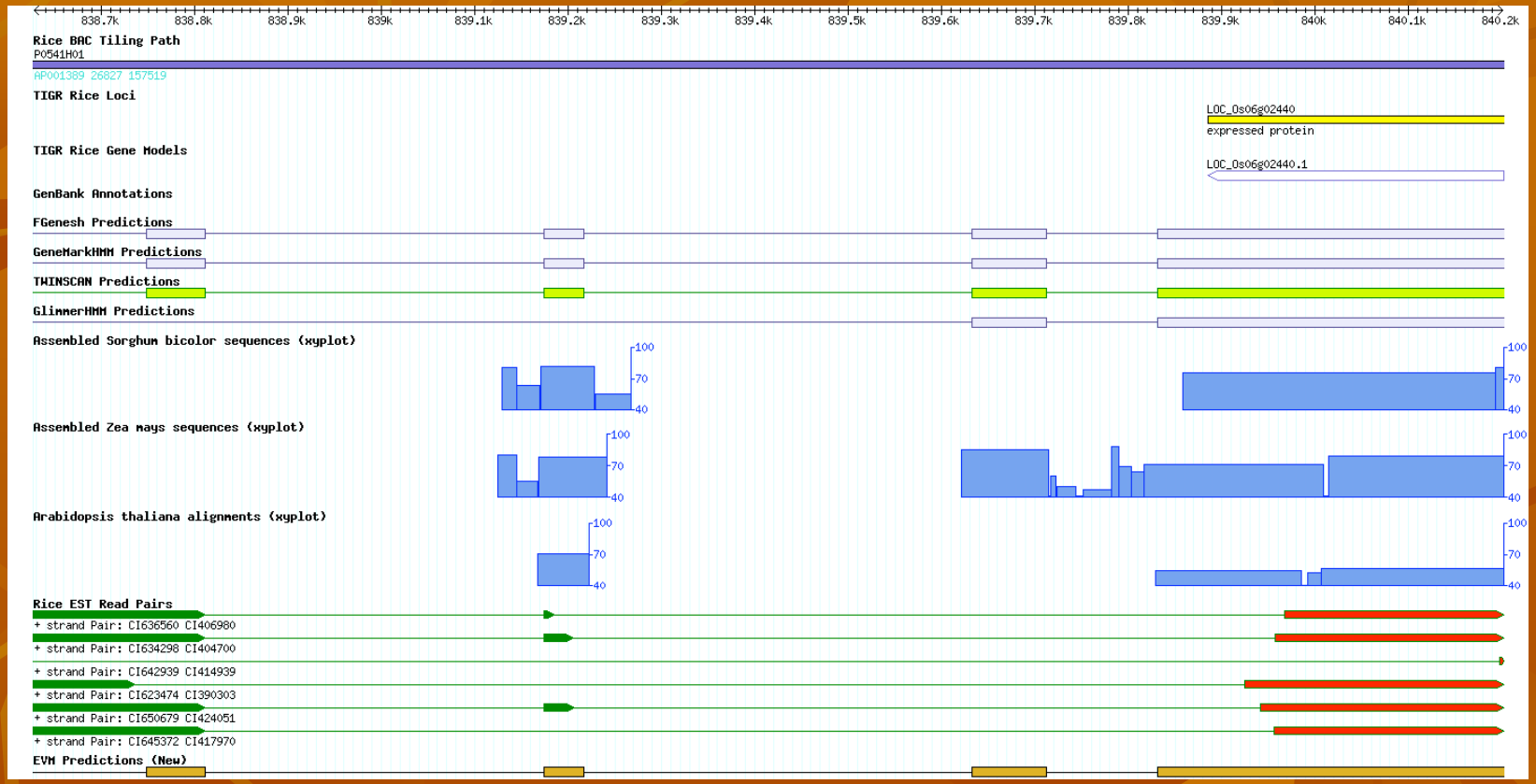

827k 828k 829k 830k 831k 832k 833k 834k 835k 836k 837k 838k 839k 840k 841k 842k 843k 844k
Rice BAC Tiling Path
P0541H01
AP001389 26827 157519
TIGR Rice Loci
LOC_Os06g02400
F-box domain containing protein, expressed
LOC_Os06g02410
expressed protein
LOC_Os06g02420
ATOZI1, putative, expressed
LOC_Os06g02430
expressed protein
LOC_Os06g02440
expressed protein
TIGR Rice Gene Models
LOC_Os06g02400.1
LOC_Os06g02410.1
LOC_Os06g02420.1
LOC_Os06g02430.1
LOC_Os06g02440.1
GenBank Annotations
FGenesh Predictions
GeneMarkHMM Predictions
GlimmerHMM Predictions
Rice FL-cDNA
gi|32969371|dbj|AK059353.1|
gi|32976789|dbj|AK066771.1|
gi|32977990|dbj|AK067972.1|
gi|32971570|dbj|AK061552.1|
gi|37991788|dbj|AK122142.1|
gi|32984057|dbj|AK098848.1|
TIGR Rice Transcript Assemblies
TA25631_4530
CX100736
TA5524_4530
BI806488
AU223362
BX928593
AU058448
CV729971
BX929001
CV732062
CV724738
D23132
TA21232_4530
CF196602
BM038818
CX100460
CF321424
BI798155
AU222970
AU082453
TA11189_4530
BU673373
TA11190_4530
C97242
CB671340

Sorghum-Rice Synteny and EST Read Pair





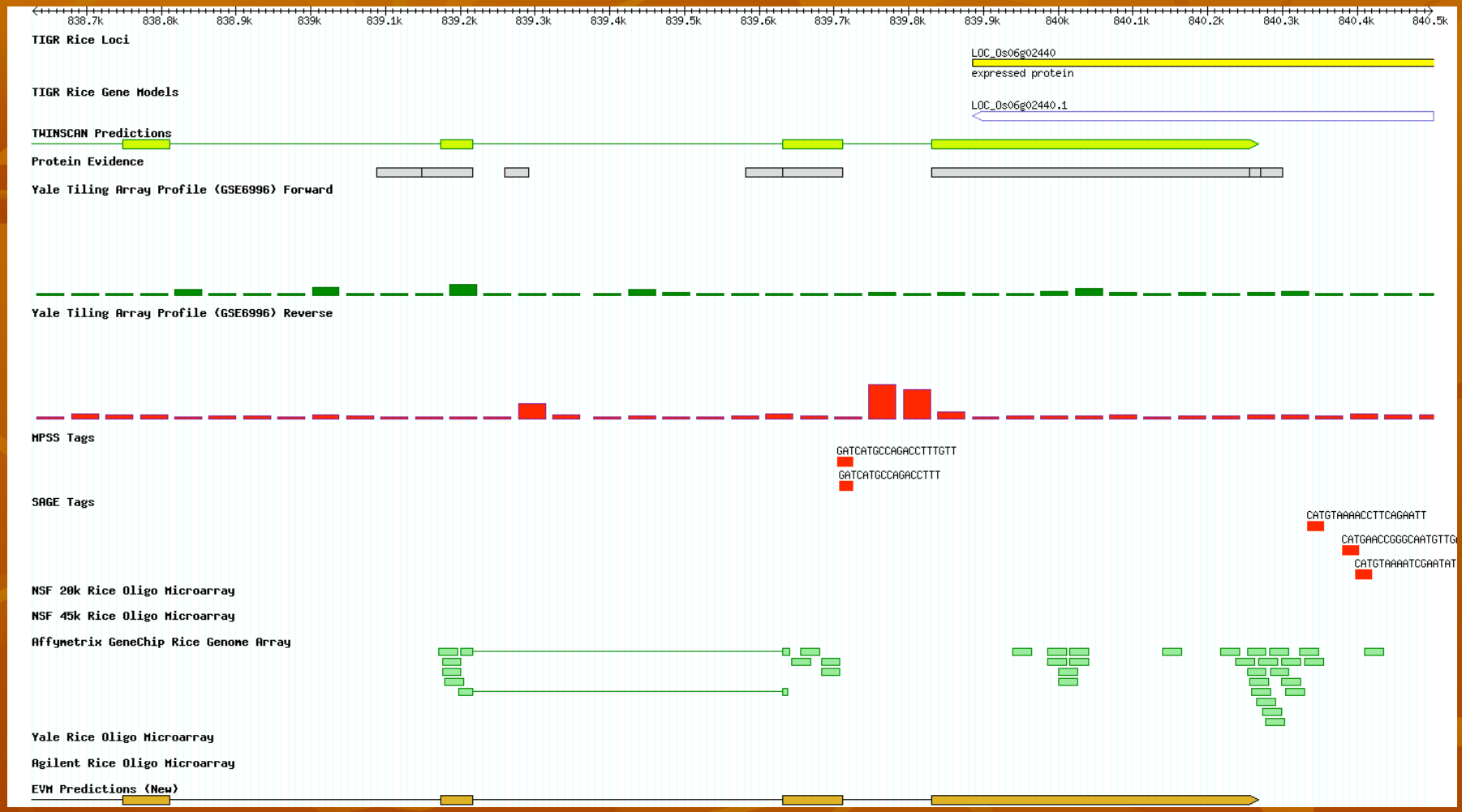
Create a Gene Model



Expression Data

Data Type	Data Source
EST/FL-cDNA	PASA/Manual curation
Peptide	Koller et al., PNAS, 2002 (6,296 peptides/2,528 fgenesh models)
MPSS	Blake Meyers (http://mpss.udel.edu/rice/)
SAGE	126,663 tags from MGOS (http://www.mgosdb.org/sage/)
Microarray	NSF Rice Oligonucleotide Array project (http://www.ricearrary.org)
Tiling array	Deng lab, Yale University

Expression Data in Gbrowse



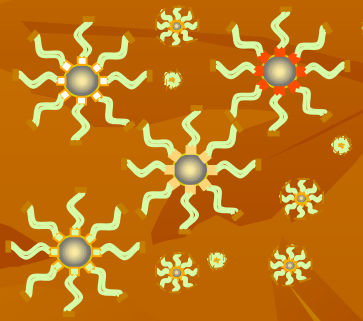
MPSS SEQUENCING TECHNOLOGY

I. Library construction

Brenner et al., *PNAS* 97:1665-70.

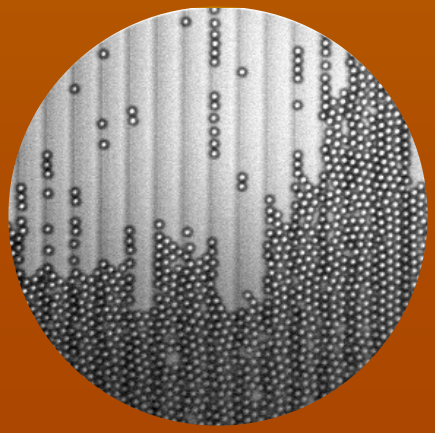
1) Cut w/ *Dpn*II
2) Ligate *Mme*I adapter

3) Cut to capture 21-22 bp "signature"
4) Add 'DNA barcode', amplify & capture on beads



Each bead contains the amplified product derived from the 3' end of a single transcript.

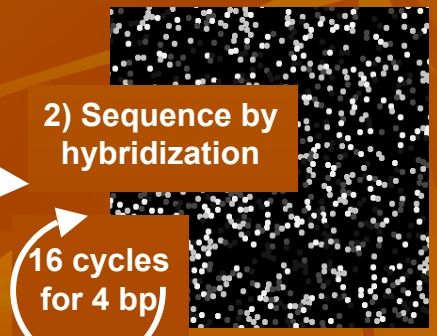
II. Loading the 'flow cell'



III. Sequencing of tags

Brenner et al., *Nat. Biotech.* 18:630-4.

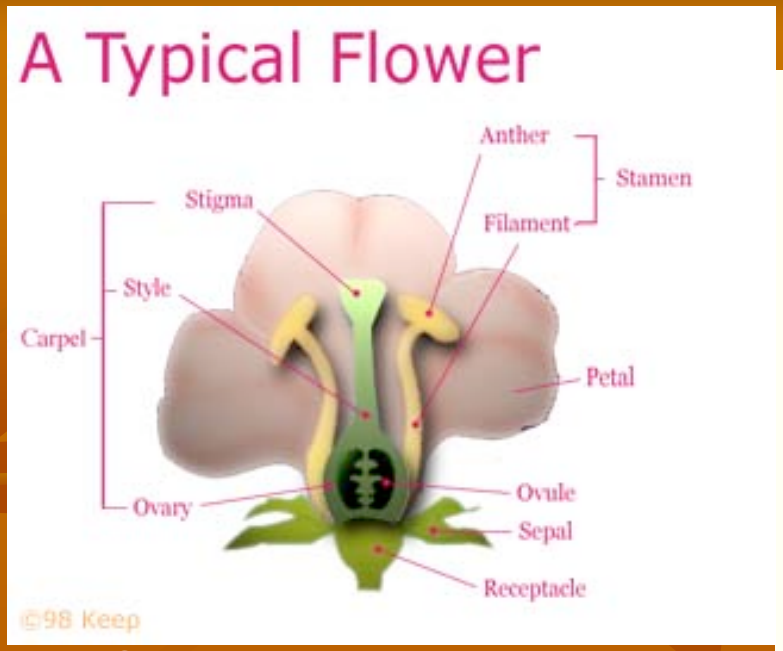
1) Add adaptors



2) Sequence by hybridization

16 cycles for 4 bp

3) Digest with Type IIS enzyme to uncover next 4 bases, repeat cycle



Ovary and mature stigma →

FME	(s)	0	0	744,319	0	1,482,167	1.000	0	2,226,486	0.0
FRO	(s)	0	0	702,046	0	1,200,706	1.000	0	1,902,752	0.0
FRR	(s)	0	0	690,560	0	1,017,481	1.000	0	1,708,041	0.0
MC00	(?)	0	0	473,119	0	639,651	1.000	0	1,112,770	0.0
MC24	(?)	0	0	616,970	0	550,712	1.000	0	1,167,682	0.0
MR03	(?)	0	0	549,558	0	617,519	1.000	0	1,167,077	0.0
MR06	(?)	0	0	383,788	0	450,596	1.000	0	834,384	0.0
MR12	(?)	0	0	411,516	0	583,952	1.000	0	995,468	0.0
MR24	(?)	0	0	399,688	0	663,844	1.000	0	1,063,532	0.0
MR48	(?)	0	0	544,207	0	592,538	1.000	0	1,136,745	0.0
MS03	(?)	0	0	583,551	0	706,204	1.000	0	1,289,755	0.0
MS06	(?)	0	0	336,582	0	489,874	1.000	0	826,456	0.0
MS12	(?)	0	0	360,120	0	452,446	1.000	0	812,566	0.0
MS24	(?)	0	0	366,078	0	446,143	1.000	0	812,221	0.0
MS48	(?)	0	0	500,054	0	676,083	1.000	0	1,176,137	0.0
MS96	(?)	0	0	333,441	0	476,251	1.000	0	809,692	0.0
NCA	(s)	0	0	665,983	0	857,387	1.000	0	1,523,370	0.0
NCL	(s)	0	0	726,299	0	993,805	1.000	0	1,720,104	0.0
NCR	(s)	0	0	819,719	0	1,005,156	1.000	0	1,824,875	0.0
NDL	(s)	0	0	868,702	0	1,057,364	1.000	0	1,926,066	0.0
NDR	(s)	0	0	636,111	0	944,735	1.000	0	1,580,846	0.0
NGD	(s)	0	0	809,426	0	1,154,506	1.000	0	1,963,932	0.0
NGS	(s)	0	0	552,552	0	718,782	1.000	0	1,271,334	0.0
NIP	(s)	0	0	888,823	0	1,153,636	1.000	0	2,042,459	0.0
NL4	(?)	0	0	0	0	0	1.000	0	0	0.0
NLA	(s)	0	0	332,531	0	442,190	1.000	0	774,721	0.0
NLB	(s)	0	0	444,641	0	556,924	1.000	0	1,001,565	0.0
NLC	(s)	0	0	532,992	0	425,854	1.000	0	958,846	0.0
NLD	(s)	0	0	443,578	0	500,829	1.000	0	944,407	0.0
NME	(s)	0	0	854,588	0	1,163,867	1.000	0	2,018,455	0.0
NOS	(s)	29	22	764,287	0	1,158,373	0.000	22	1,922,660	11.4
NPO	(s)	0	0	811,742	0	1,024,436	1.000	0	1,836,178	0.0
NR2	(?)	0	0	0	0	0	1.000	0	0	0.0
NRA	(s)	0	0	905,659	0	1,019,990	1.000	0	1,925,649	0.0
NRB	(s)	0	0	940,221	0	1,015,594	1.000	0	1,955,815	0.0
NSL	(s)	0	0	768,410	0	1,205,268	1.000	0	1,973,678	0.0
NSR	(s)	0	0	598,714	0	682,877	1.000	0	1,281,591	0.0
NST	(s)	0	0	692,117	0	746,500	1.000	0	1,438,617	0.0
NYL	(s)	0	0	626,418	0	1,048,895	1.000	0	1,675,313	0.0
NYR	(s)	0	0	737,011	0	593,122	1.000	0	1,330,133	0.0
PLA	(s)	0	0	424,016	0	503,578	1.000	0	927,594	0.0
PLC	(s)	0	0	429,898	0	550,824	1.000	0	980,722	0.0
PLW	(s)	0	0	363,097	0	436,803	1.000	0	799,900	0.0
PSC	(s)	0	0	454,566	0	591,489	1.000	0	1,046,055	0.0
PSI	(s)	0	0	435,252	0	556,580	1.000	0	991,832	0.0
PSL	(s)	0	0	368,498	0	516,690	1.000	0	885,188	0.0
PSN	(s)	0	0	366,140	0	603,782	1.000	0	969,922	0.0
PSY	(s)	0	0	411,871	0	558,414	1.000	0	970,285	0.0
XC00	(s)	0	0	448,936	0	453,985	1.000	0	902,921	0.0
XC06	(s)	0	0	511,865	0	539,475	1.000	0	1,051,340	0.0
XC24	(s)	0	0	437,295	0	476,855	1.000	0	914,150	0.0
XR03	(s)	0	0	389,774	0	484,771	1.000	0	874,545	0.0
XR06	(s)	0	0	467,130	0	495,044	1.000	0	962,174	0.0

Refine Gene Structure

“Have no fear of perfection - you'll never reach it.”

- Salvador Dalí