Finding Features in DNA Sequences using Computational Methods

Overview

- What is automated annotation?
- Why do automated annotation?
- TIGR's Eukaryotic automated annotation pipeline
  - Mask repeats
  - Develop training set
  - Train and run gene finders
  - Run database comparisons & searches
  - Consensus Prediction
  - EST based refinement
  - Load and view automated gene Structures
- How good are automated gene predictions?

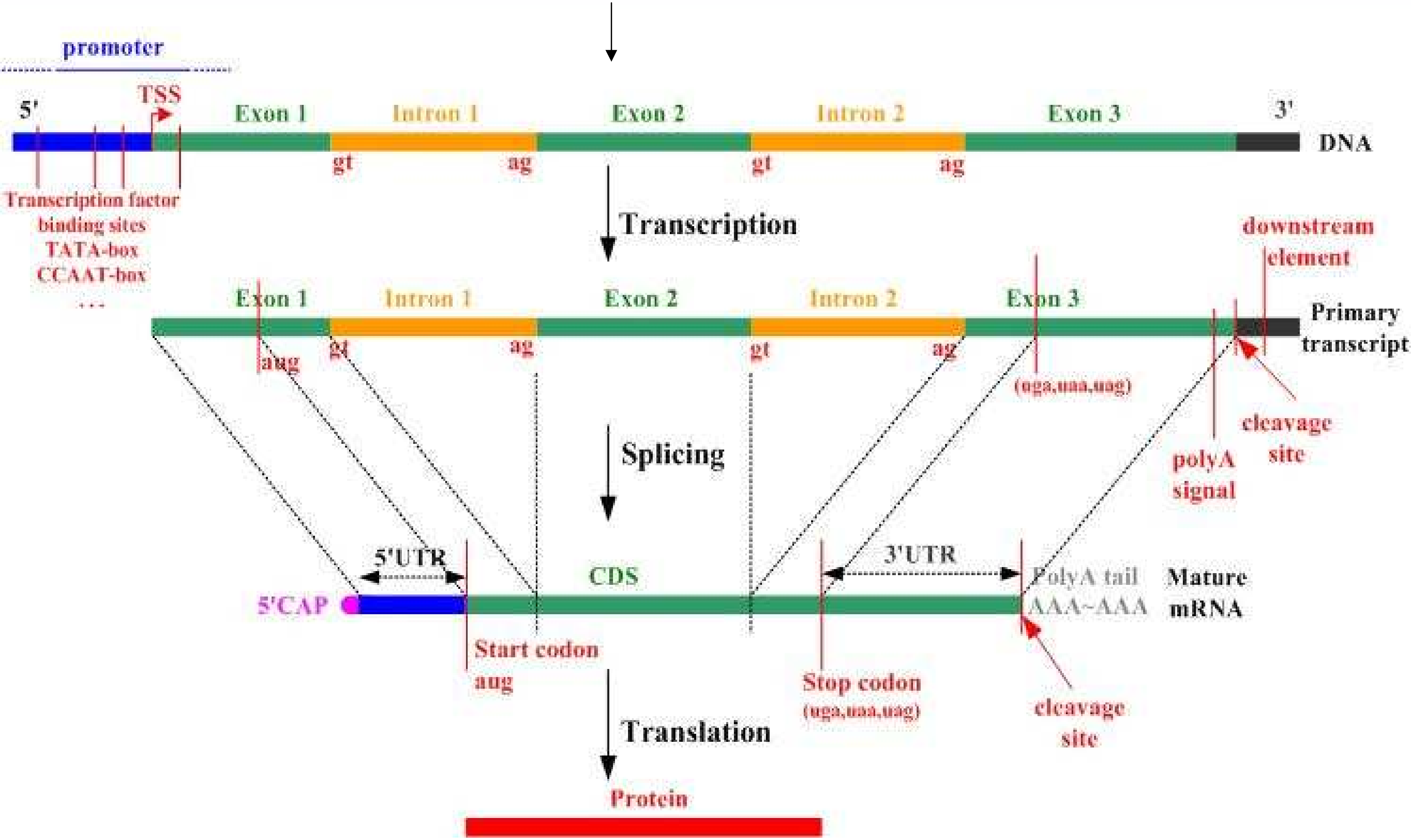
What is automated annotation?

The raw sequences after closure are run through a series of programs and scripts which we call the “pipeline” in an automated way to generate a basic working gene set that serve as a beginning point for further work.

In real world nothing is completely automated!!!

Gene Structure

AGCGTGGTAGCGCGAGTTTGCGAGCTAGCTAGGCTCCGGATGCGA
CCAGCTTTGATAGATGAATATAGTGTGCGCGACTAGCTGTGTGTT
GAATATATAGTGTGTCTCTCGATATGTAGTCTGGATCTAGTGTTG
GTGTAGATGGAGATCGCGTAGCGTGGTAGCGCGAGTTTGCGAGCT

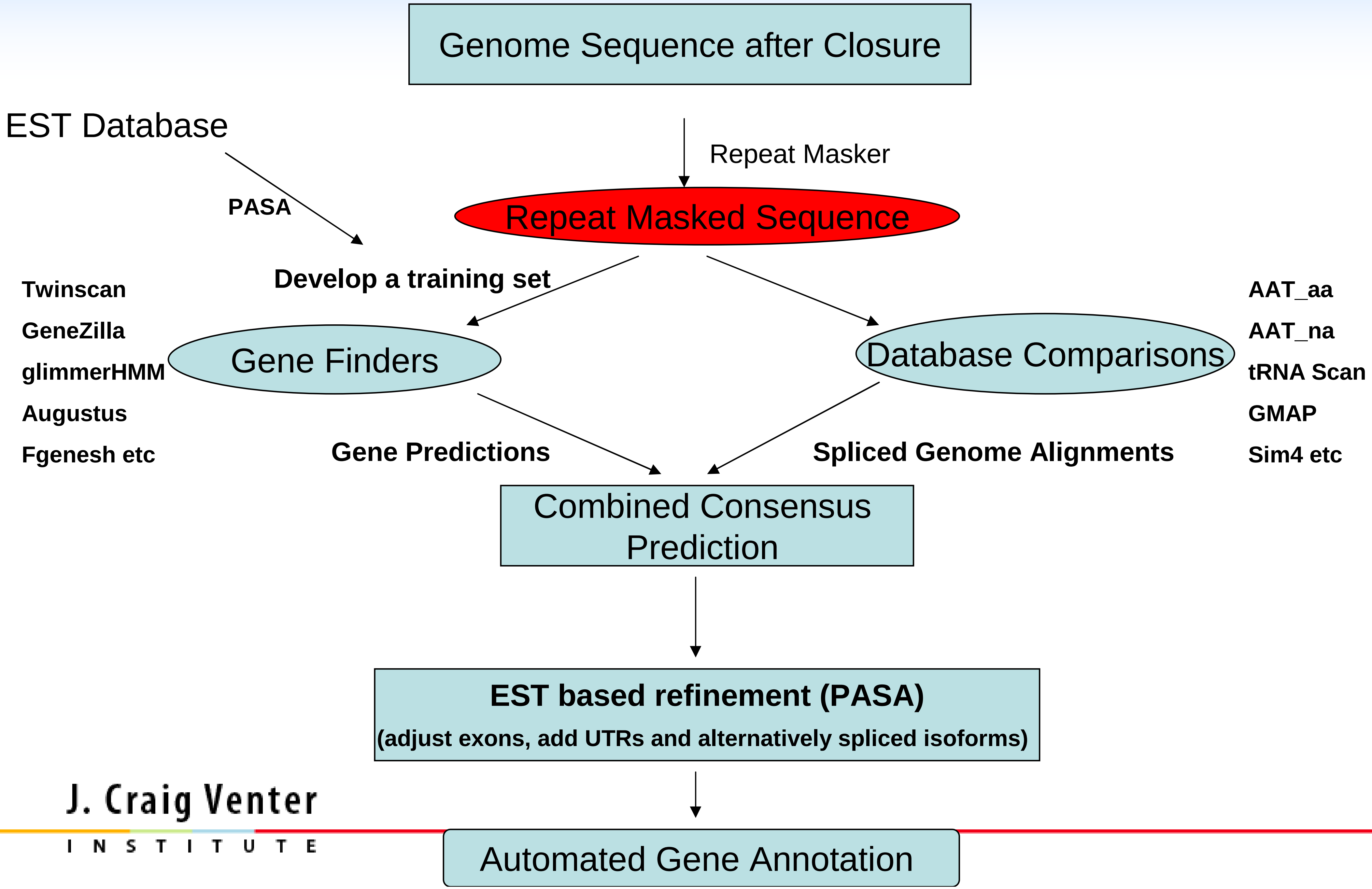


Why do automated annotation?

- It is fast and cost effective
- Allows immediate insights to genome biology.
- Provides a starting point for more accurate manual curation efforts.

Eukaryotic Structural Annotation pipeline

(Genome Level Computes)



Genome Sequence after Closure

EST Database

Repeat Masker

PASA

Repeat Masked Sequence

Develop a training set

Twinscan

GeneZilla

glimmerHMM

Augustus

Fgenesh etc

Gene Finders

Database Comparisons

AAT_aa

AAT_na

tRNA Scan

GMAP

Sim4 etc

Gene Predictions

Spliced Genome Alignments

Combined Consensus Prediction

EST based refinement (PASA)

(adjust exons, add UTRs and alternatively spliced isoforms)

Automated Gene Annotation

J. Craig Venter INSTITUTE

Repeats

- A repeat is a substring that occurs multiple times within a sequence or a collection of sequences.
- Repetitive sequences account for roughly half of the human genome sequence.
- Repeats can interfere with genome analyses in several ways.

Repeats : Example

- GCAGCCTCAGTGGTGGTAGAAGCCGCCTCAGTGGTGGTAGAAGCACCTCAGTGGTGGTAGAGGCAGCCTCAGTGGTGGTAGAAGCCGCCTAGTGGTGGTAGAAGCAGCCTCAGTGGTGGTAGAGGCAGCCTCAGTGTGGTAGAAGCCGCCTCAGTGGTGGTAGAAGCAGCCTCAGTGGTGTAGAGGCAGCCTCAGTGGTGGTAGAAGC

- CCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGGCGGGGGCCTCAGTAGTAGTGG

Repeats

Tandem Repeats (Satellite DNA)

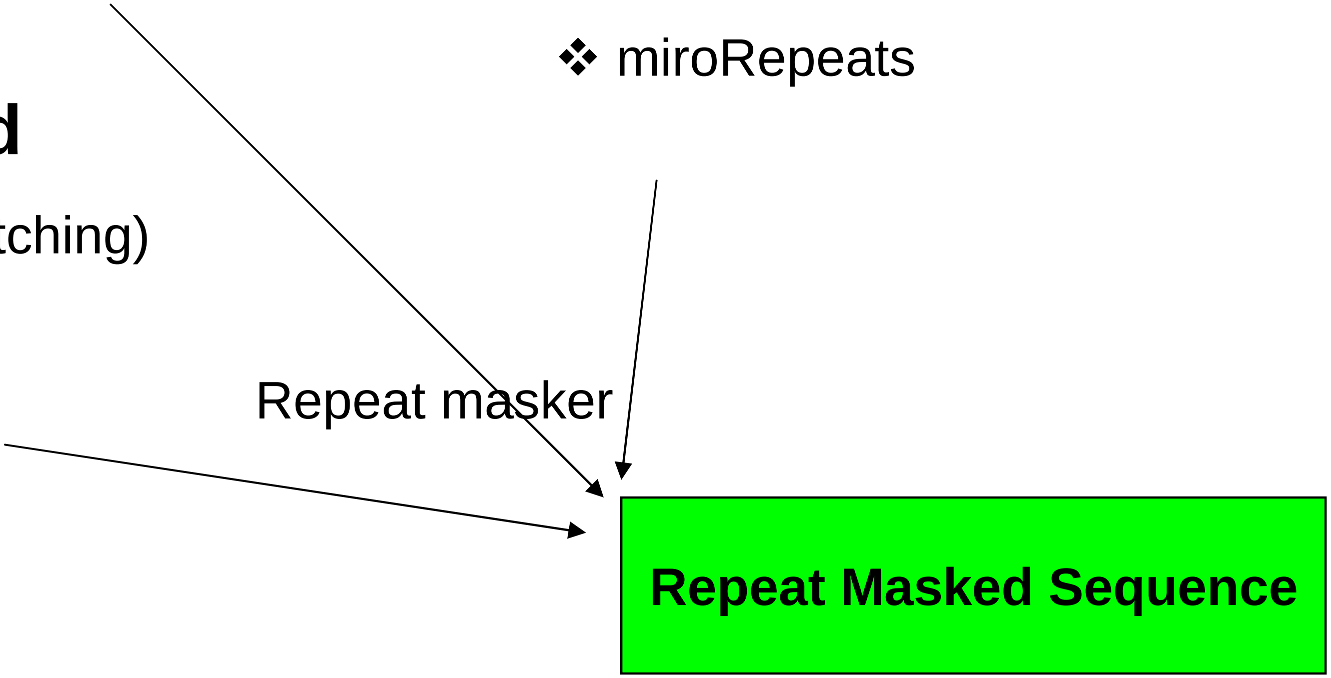
- mreps
- tandem repeat finder (trf)

Dispersed Repeats (Transposable Elements)

- Repeat Scout
- transposonPSI
- miroRepeats

Repeats of any kind

- VMATCH (fast string matching)



Repeat masker

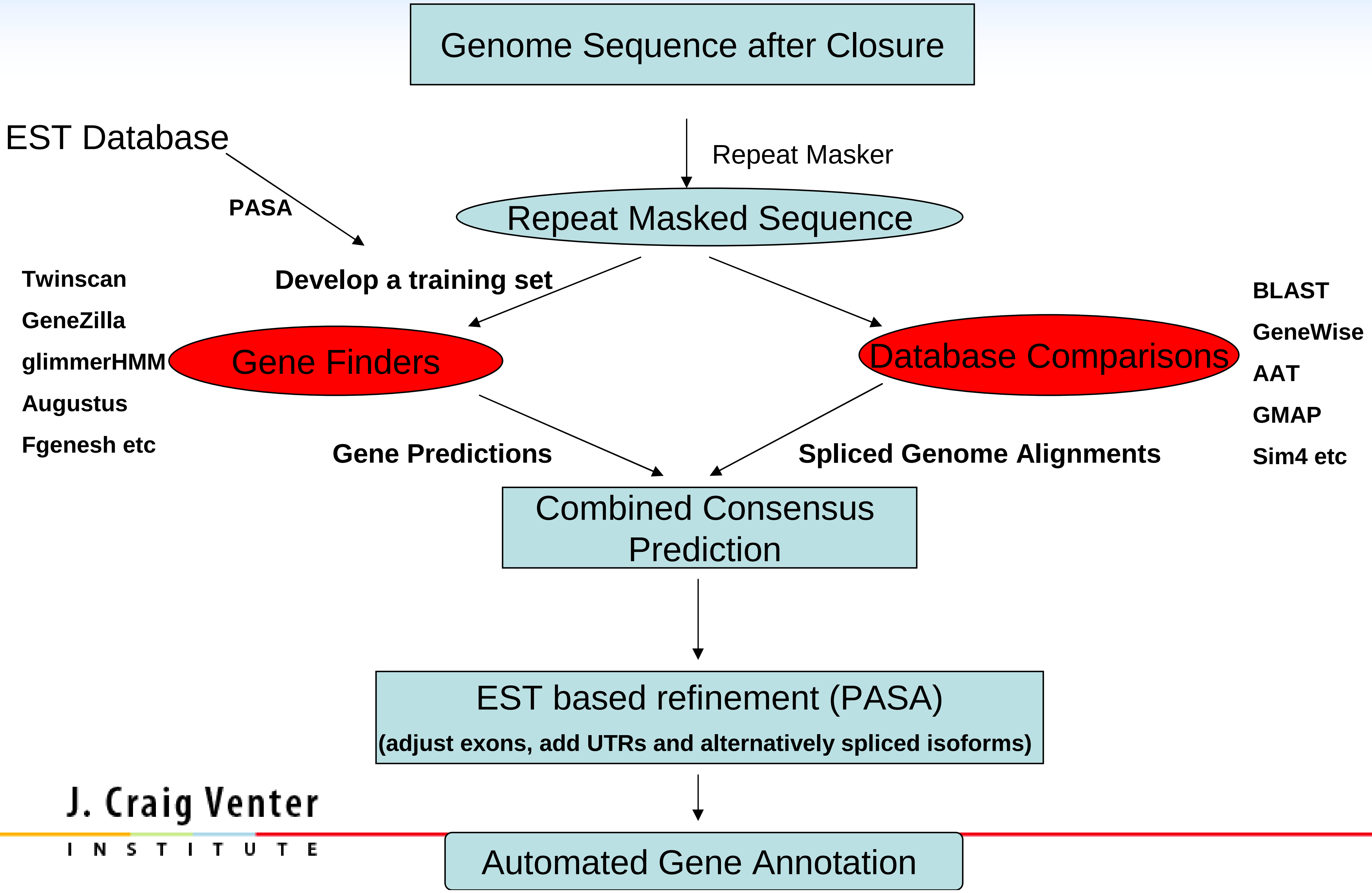
Repeat Masked Sequence

Why mask repeats?

- Repeats tend to confuse *ab initio* gene finders
  - calling exons or complete genes in repeat regions.
  - fragmenting gene predictions.
- Repeats confound sequence alignment, such as in searches for synteny or segmental duplications.

Eukaryotic Structural Annotation pipeline

(Genome Level Computes)



Genome Sequence after Closure

EST Database

PASA

Repeat Masker

Repeat Masked Sequence

Develop a training set

Twinscan

GeneZilla

glimmerHMM

Augustus

Fgenesh etc

Gene Finders

Database Comparisons

BLAST

GeneWise

AAT

GMAP

Sim4 etc

Gene Predictions

Spliced Genome Alignments

Combined Consensus Prediction

EST based refinement (PASA)

(adjust exons, add UTRs and alternatively spliced isoforms)

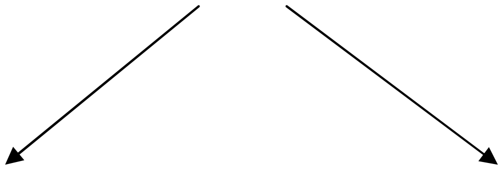
Automated Gene Annotation

J. Craig Venter INSTITUTE

Gene Predictions

A gene finding software parses a sequence into non overlapping **coding segments** (CDSs) consisting of exons separated by introns.

It predicts different features within a gene.



Signal Sensor

(Fixed length features)

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

Content Sensor

(Variable length features)

Exons
Introns
Intergenic regions
UTRs (UnTranslated Regions)

Protein coding gene identification

Intrinsic (Signal and Content)

- **Ab-Initio Gene Prediction**

  Genscan
  GenemarkHMM
  Fgenesh
  GlimmerHMM
  GeneZilla
  SNAP
  PHAT
  AUGUSTUS
  Genie

Hybrid (Intrinsic and Extrinsic Evidence)

Twinscan
N-scan
SLAM
TWAIN
ExonHunter
shortHMM

Extrinsic (Similarity based)

- **Protein homology**

  AAT
  GeneWise → **Hands on this afternoon**

- **EST, full-length cDNA alignments**

  est_genome
  AAT
  sim4
  geneseqer
  BLAT
  GMAP

- **Genome Sequence Comparisons**

  blastZ
  MUMmer
  Avid
  [SM]Lagan
  Vista

Ab initio gene prediction

Adopt a rigorous probabilistic model of gene structure and choose the most likely labeling of sequence states (exon, intron, start, stop) according to that model.

- Pros
  - Fast and efficient
  - Remarkable accuracy at the nucleotide level
- Cons
  - Less than 50% accuracy at the gene level

Species-specific setting like GC content, gene density, Gene/Exon/Intron length distribution and codon usage are taken into consideration

How do you train a gene finder?

- ❖ **Run alignment assembly program against the EST databases (PASA) – EST's can also be downloaded from the genbank**

- ❖ **Build a training set with genes based on PASA (EST/cDNA) evidence**

  Confirmed splice sites, exons from EST alignments
  Complete gene structures from cDNA alignments
  Used to build statistical models of exons, introns, intergenic regions, etc.

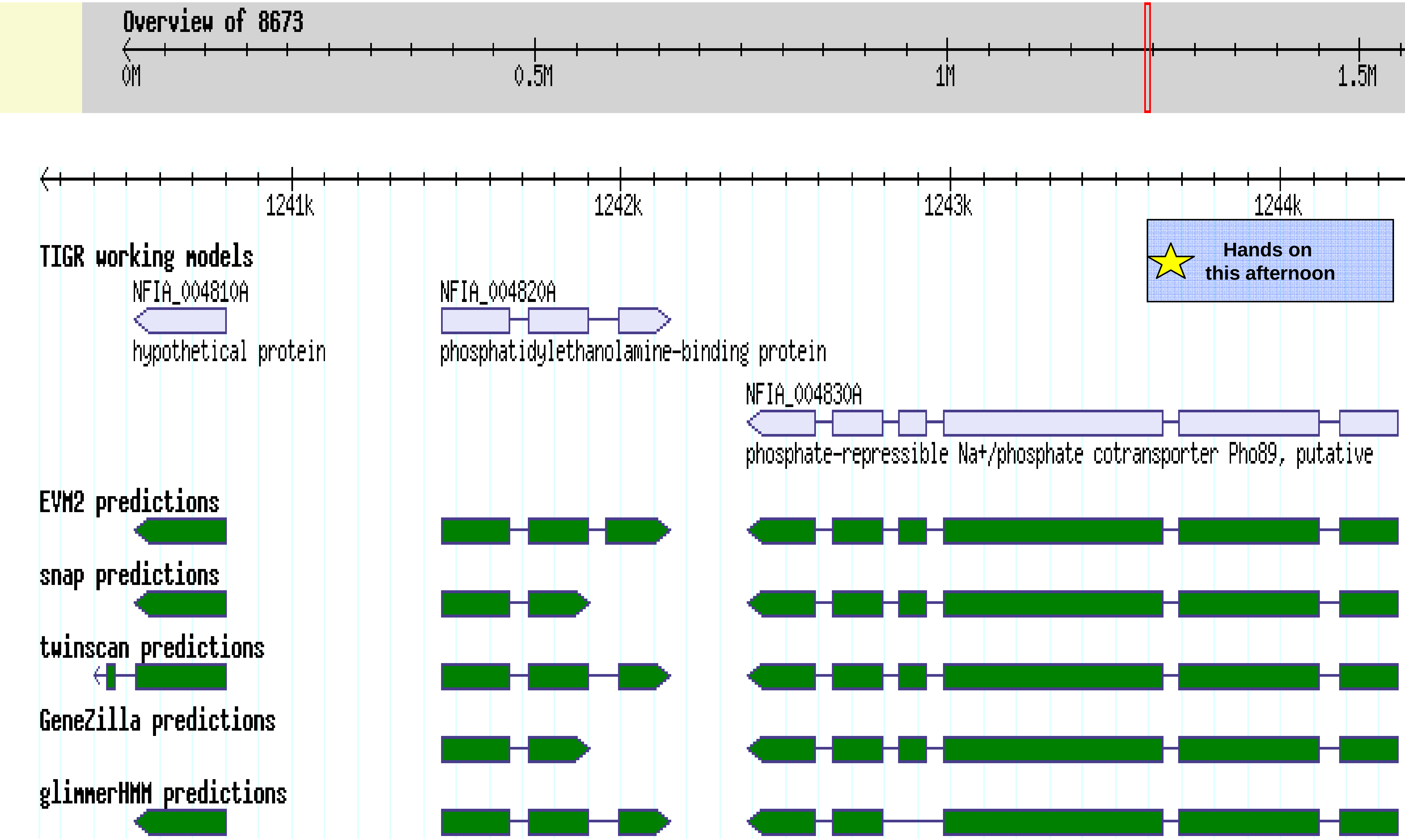
- ❖ **Train the gene finders based on the training set**

  Training set could be partitioned
  -Part of it to train gene finders
  -Part of it to test the prediction

- ❖ **Predict genes in the genome**

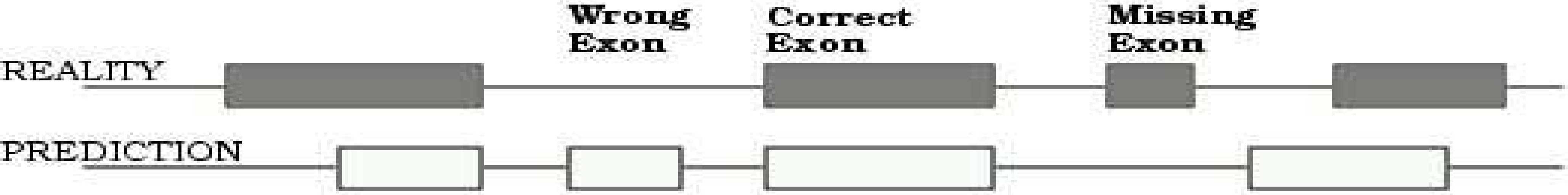
- ❖ **Evaluate the predictions**

Example of a gene prediction



Hands on this afternoon

Evaluating the prediction



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**

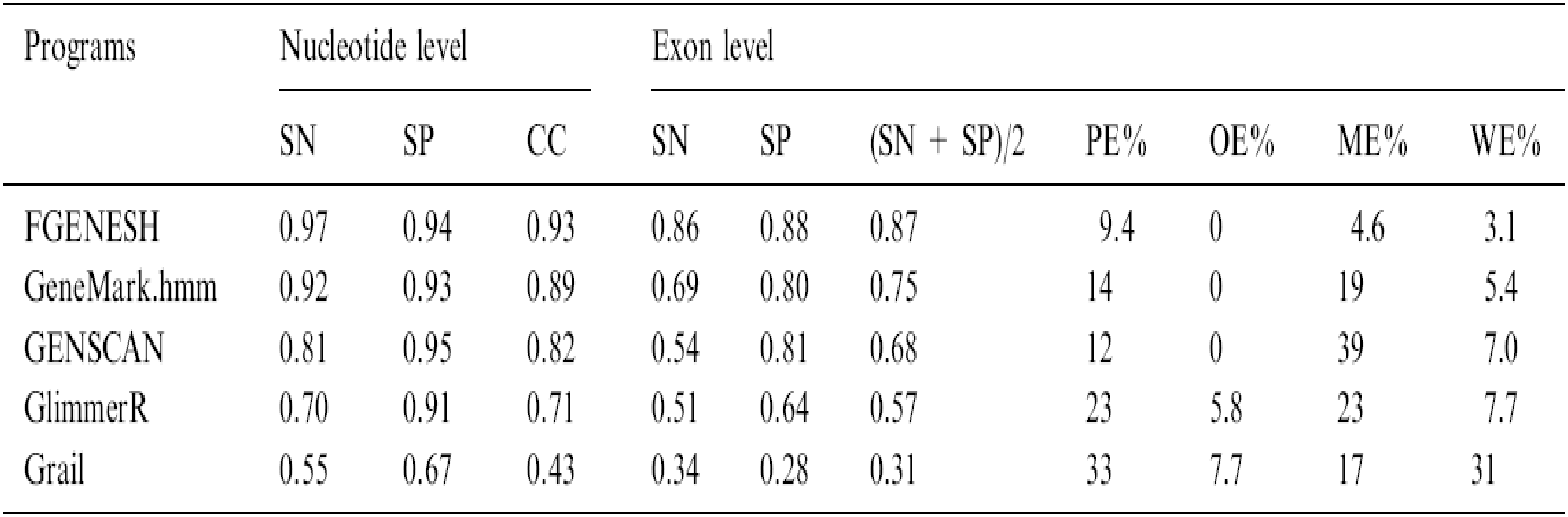
$$Sp = \frac{\text{number of Correct Exons}}{\text{number of Predicted Exons}}$$ **Specificity**

$$ME = \frac{\text{number of Missing Exons}}{\text{number of Actual Exons}}$$ **(Sensitivity)**

$$WE = \frac{\text{number of Wrong Exons}}{\text{number of Predicted Exons}}$$ **(Specificity)**

Specificity vs. Sensitivity

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 *ab initio* 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
- Repeats/transposons

Best Gene Finding Strategy?

- Use a combination of multiple gene prediction programs, including:

  Ab initio programs based on signals/content

  Alignment programs

  Programs that combine intrinsic and extrinsic data

- Automated or manual methods applied to combine the best features of each program

- Multiple iterations of training, parameterization

- Comparative genomics holds the best promise for automated methods

Performance of gene finders are dependent on…

- ❖ Benchmarks like training data and test data set
- ❖ Training gene finders with or without EST′s

www.genefinding.org

Contacting the author of the gene finder program always helped in our experience !

Database comparison and searches

- Sequence database searches: AAT (Analysis and Annotation Tool) : AAT searches are similar to the blast searches.

  Nucleotide vs. Protein databases : dps/nap

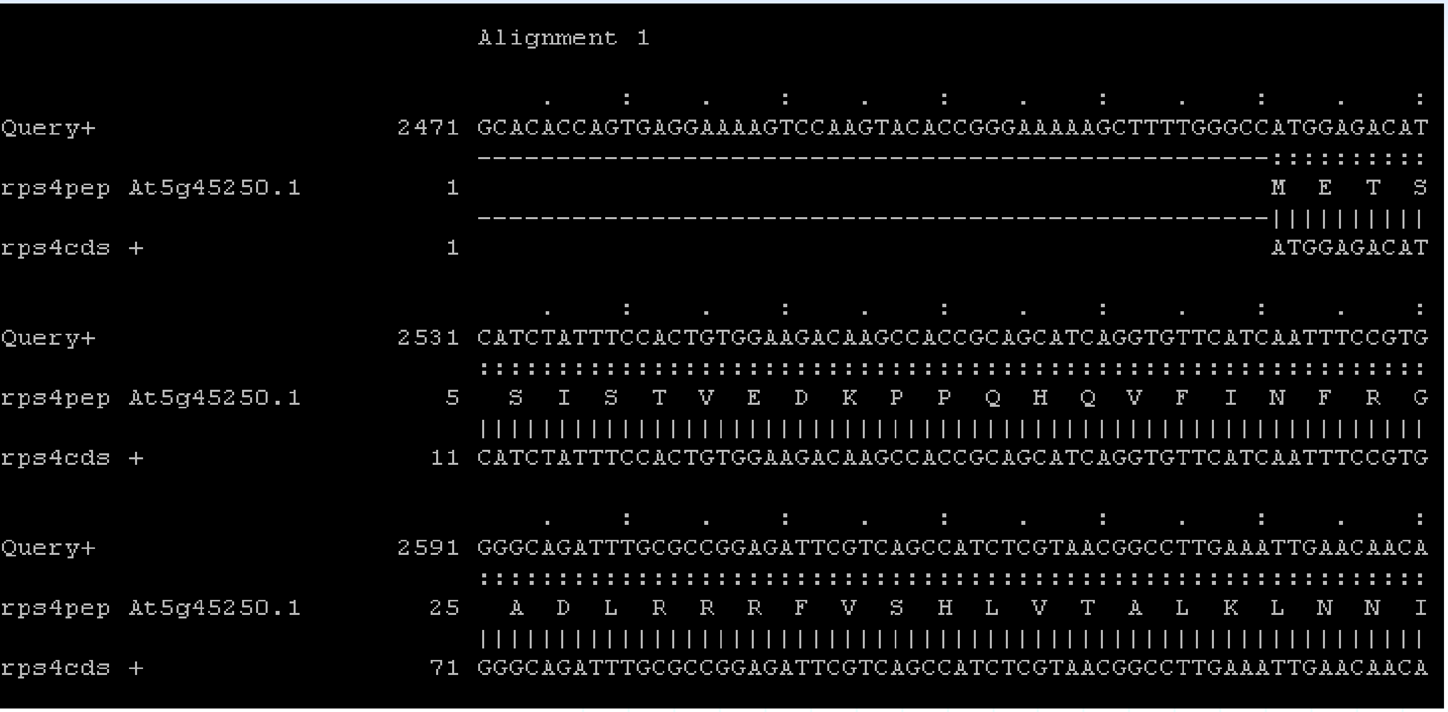
  Nucleotide vs. Nucleotide databases : dds/gap2

  - Protein database searches : nraa
  - Nucleotide database searches : EST databases

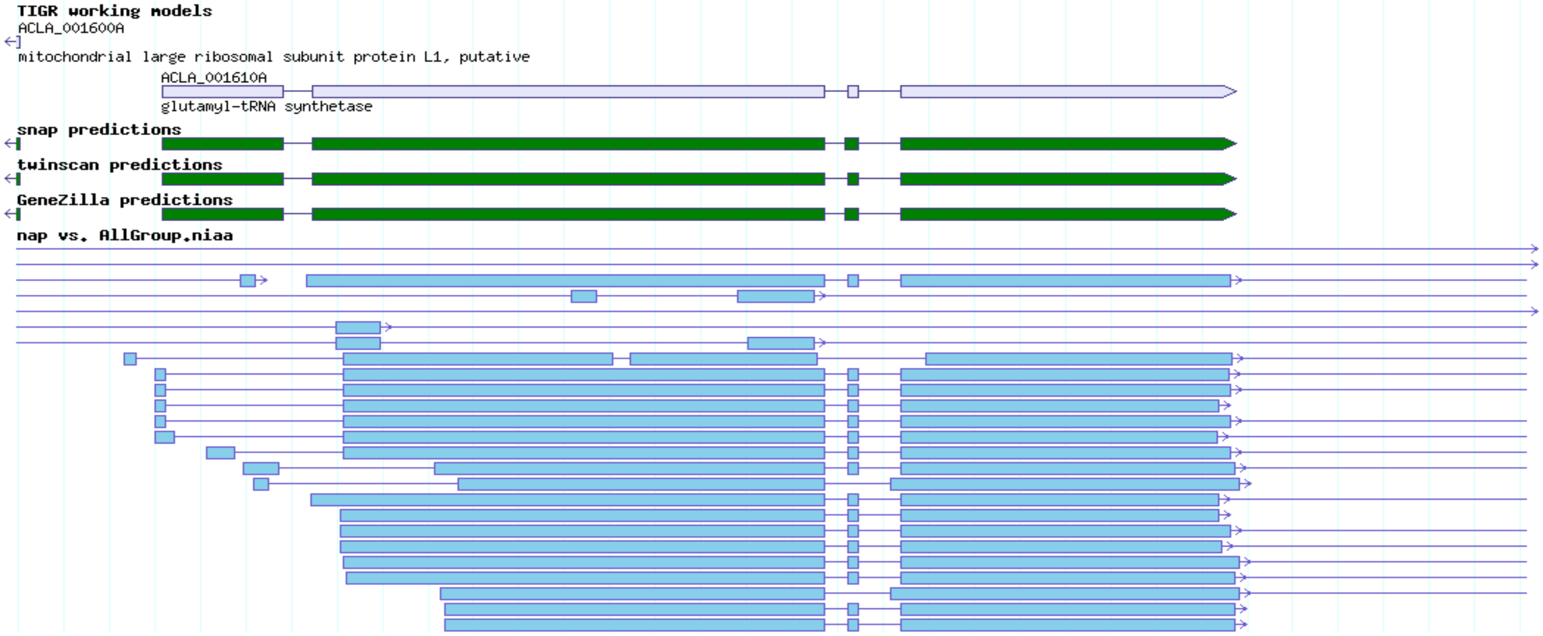
↓

Spliced genome alignments

AAT Alignment



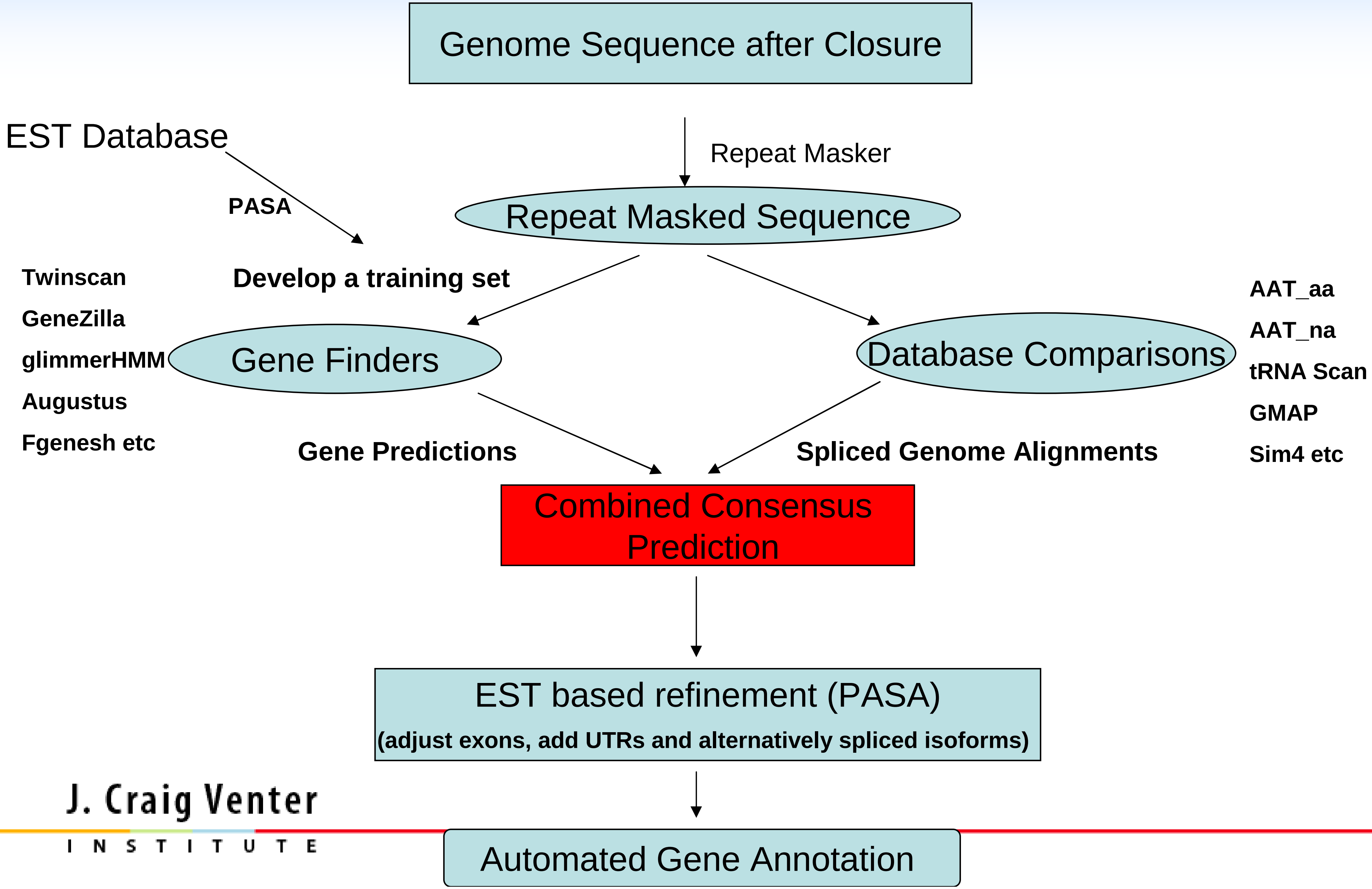
```
                              Alignment 1

                                  .    :    .    :    .    :    .    :    .    :    .    :
Query+                   2471 GCACACCAGTGAGGAAAAGTCCAAGTACACCGGGAAAAAGCTTTTGGGCCATGGAGACAT
                              --------------------------------------------------::::::::::
rps4pep At5g45250.1         1                                                   M  E  T  S
                              --------------------------------------------------||||||||||
rps4cds +                   1                                                   ATGGAGACAT

                                  .    :    .    :    .    :    .    :    .    :    .    :
Query+                   2531 CATCTATTTCCACTGTGGAAGACAAGCCACCGCAGCATCAGGTGTTCATCAATTTCCGTG
                              ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
rps4pep At5g45250.1         5   S  I  S  T  V  E  D  K  P  P  Q  H  Q  V  F  I  N  F  R  G
                              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
rps4cds +                  11 CATCTATTTCCACTGTGGAAGACAAGCCACCGCAGCATCAGGTGTTCATCAATTTCCGTG

                                  .    :    .    :    .    :    .    :    .    :    .    :
Query+                   2591 GGGCAGATTTGCGCCGGAGATTCGTCAGCCATCTCGTAACGGCCTTGAAATTGAACAACA
                              ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
rps4pep At5g45250.1        25   A  D  L  R  R  R  F  V  S  H  L  V  T  A  L  K  L  N  N  I
                              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
rps4cds +                  71 GGGCAGATTTGCGCCGGAGATTCGTCAGCCATCTCGTAACGGCCTTGAAATTGAACAACA
```



J. Craig Venter
INSTITUTE

Eukaryotic Structural Annotation pipeline

(Genome Level Computes)



Genome Sequence after Closure

Repeat Masker

Repeat Masked Sequence

EST Database

PASA

Develop a training set

Twinscan

GeneZilla

glimmerHMM

Augustus

Fgenesh etc

Gene Finders

Database Comparisons

AAT_aa

AAT_na

tRNA Scan

GMAP

Sim4 etc

Gene Predictions

Spliced Genome Alignments

Combined Consensus Prediction

EST based refinement (PASA)

(adjust exons, add UTRs and alternatively spliced isoforms)

Automated Gene Annotation

Consensus Prediction

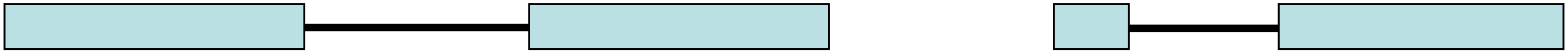
A program that combines diverse evidence (like gene predictions, sequence alignments) into a weighted **consensus** gene prediction.

- EVidence Modeler (B. Haas)
  http://evidencemodeler.sf.net

- JIGSAW (J. Allen)
  http://cbcb.umd.edu/software/jigsaw

Combining Predictions

Decomposing a single set of gene predictions:



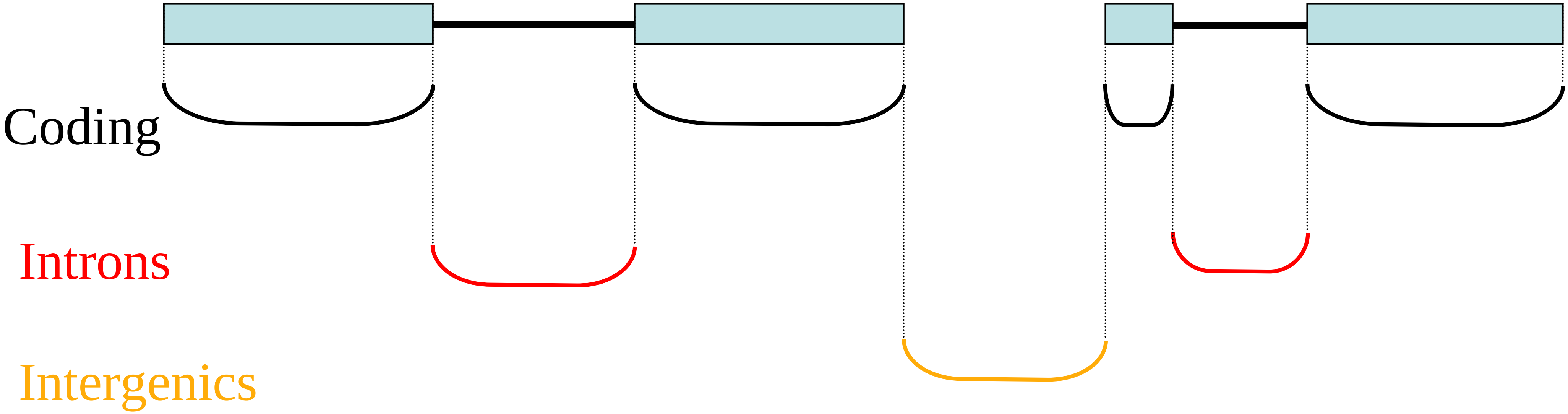
Coding

Introns

Intergenics

Combining Predictions

Decomposing a single prediction:



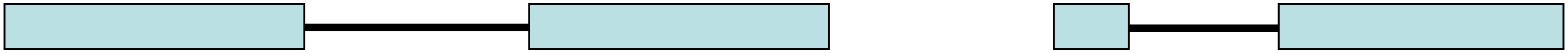
Coding

Introns

Intergenics

Combining Predictions

Scoring at the basepair level

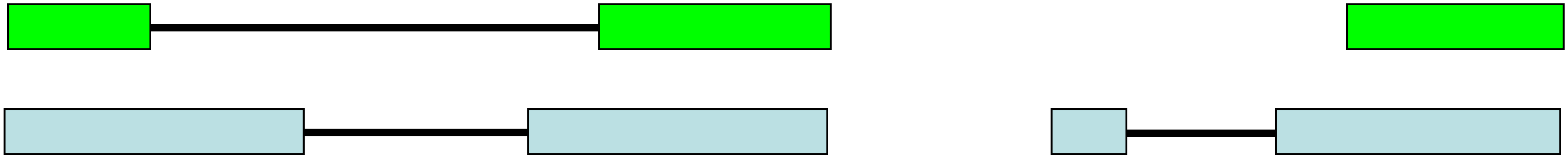


Coding	1	1	1	1				1	1	1	1				1			1	1	1	1
Intron					1	1	1									1	1				
Intergenic												1	1	1							

For example purposes, each prediction type has weight = 1

Combining Predictions

Scoring basepairs, multiple predictions



Coding	2	2	1	1				1	2	2	2				1			1	2	2	2
Intron			1	1	2	2	2	1								1	1				
Intergenic												2	2	2	1	1	1	1			

Scores are summed up for highest scoring path.

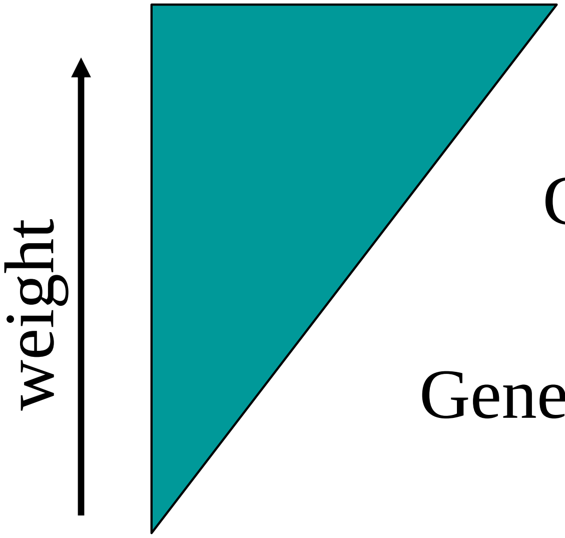
Including Other Evidence

Protein and Transcript Alignments:

- contribute to Coding and Intron scores
- contribute to Feature sets:
  - introns (require consensus splice sites)
  - internal exons (must encode uninterrupted ORF)
  - initial, terminal, and single exons via special options.
- Scores are additive: every alignment counts.

Setting Evidence Weights

An intuitive approach would weight transcript alignments and protein homology more than computationally predicted exons.



weight

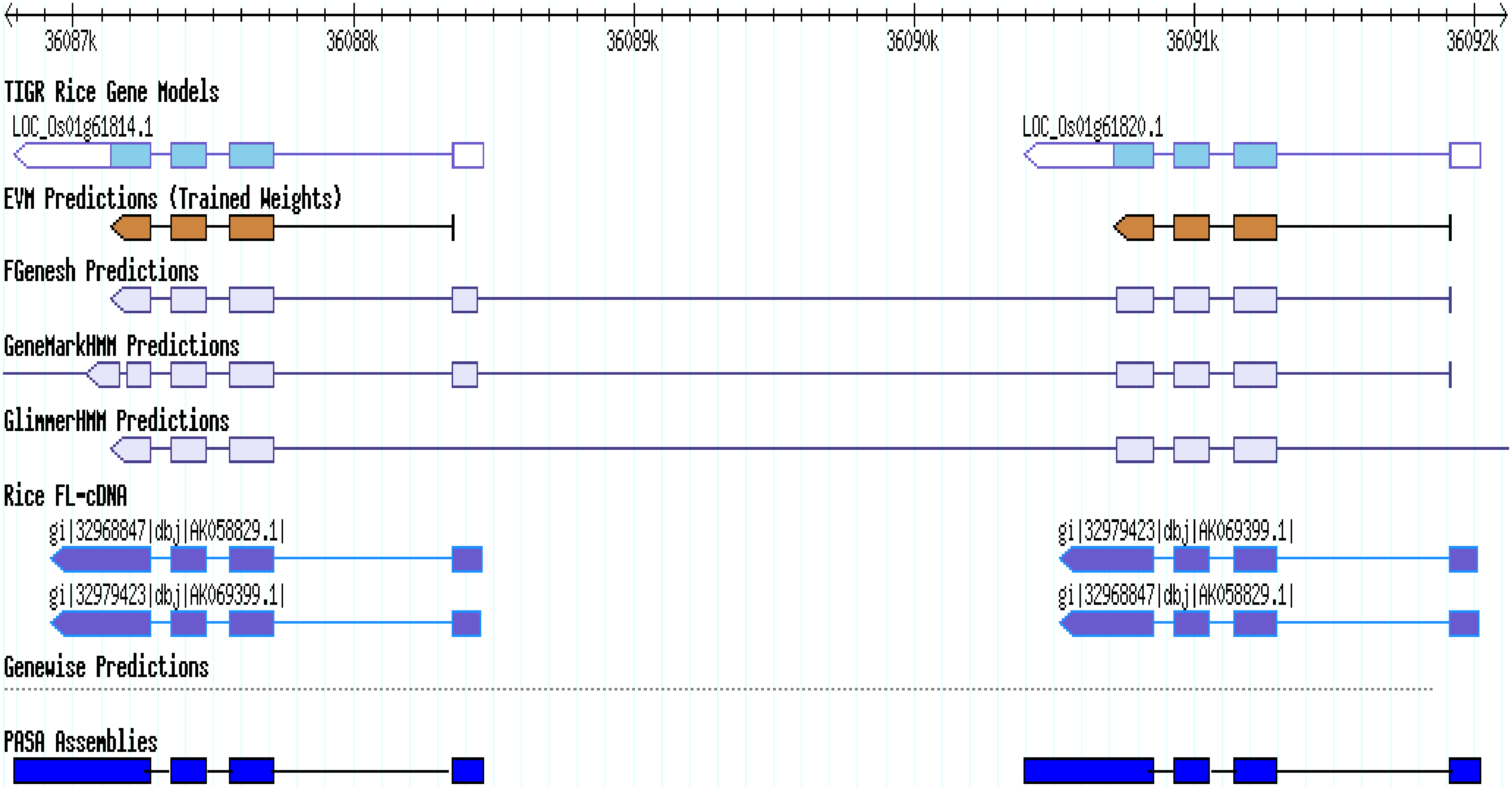
PASA transcript alignment assemblies

Genewise protein alignments

Gene Predictions, AAT alignments

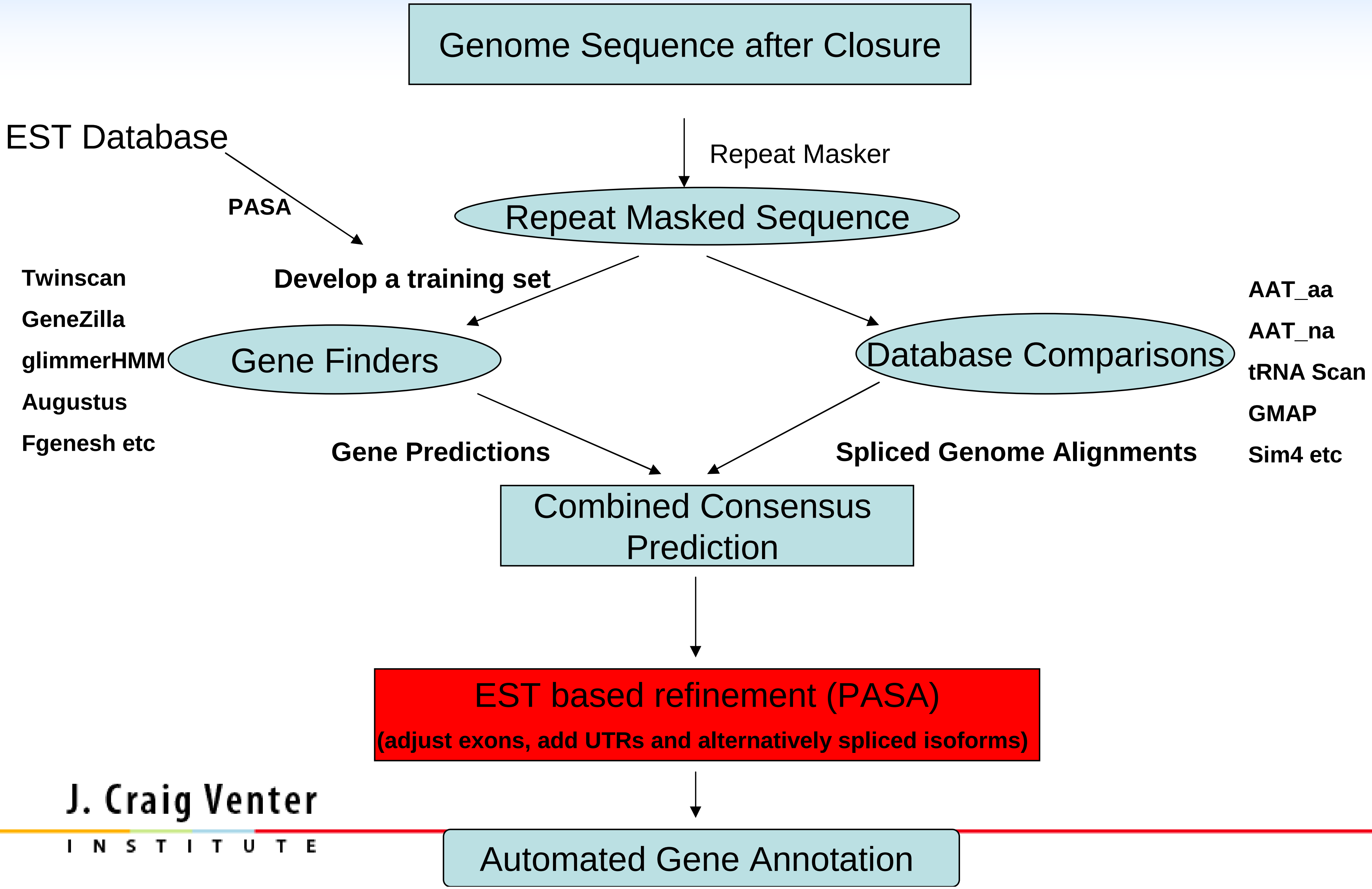
Evidence weights are manually configured or trained

An example : EVM



Eukaryotic Structural Annotation pipeline

(Genome Level Computes)



Genome Sequence after Closure

Repeat Masker

EST Database

PASA

Repeat Masked Sequence

Develop a training set

Twinscan

GeneZilla

glimmerHMM

Augustus

Fgenesh etc

Gene Finders

Database Comparisons

AAT_aa

AAT_na

tRNA Scan

GMAP

Sim4 etc

Gene Predictions

Spliced Genome Alignments

Combined Consensus Prediction

EST based refinement (PASA)

(adjust exons, add UTRs and alternatively spliced isoforms)

Automated Gene Annotation

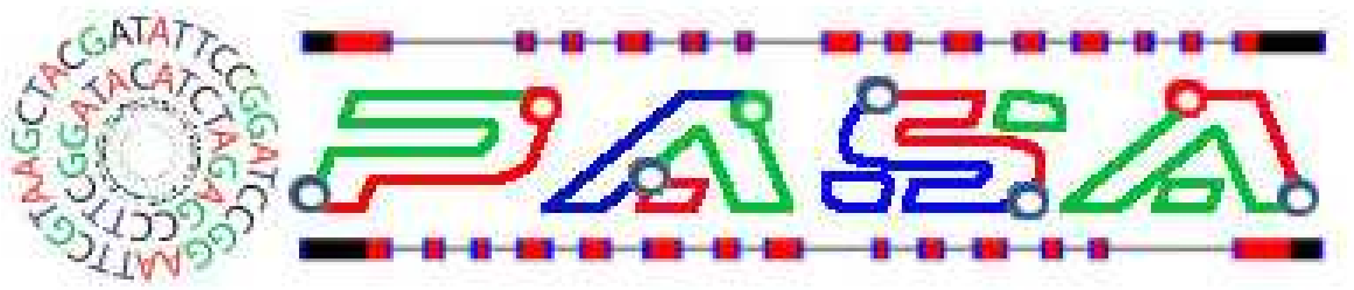
J. Craig Venter INSTITUTE

ESTs and Full-length cDNAs for Genome Annotation

PASA - **P**rogram to **A**ssemble **S**pliced **A**lignments

- “**Gold standard**” for gene structure resolution
  - Introns and exons via spliced alignment

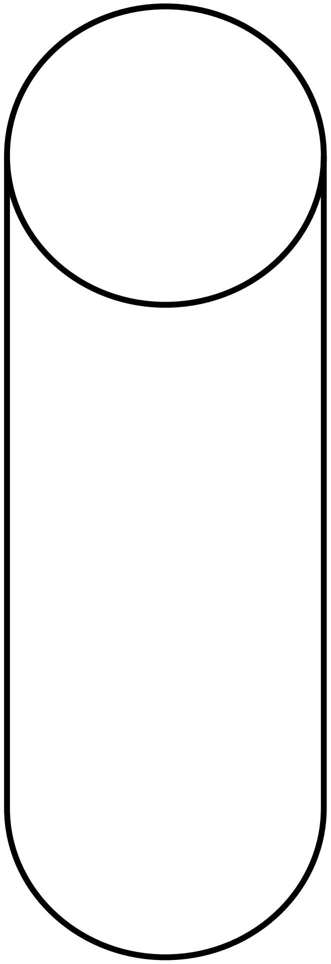
- Direct evidence for:
  - Alternative splicing
  - Untranslated regions (UTRs)
  - Polyadenylation sites



http://pasa.sf.net

Brian Haas

The PASA Pipeline [at a glance]

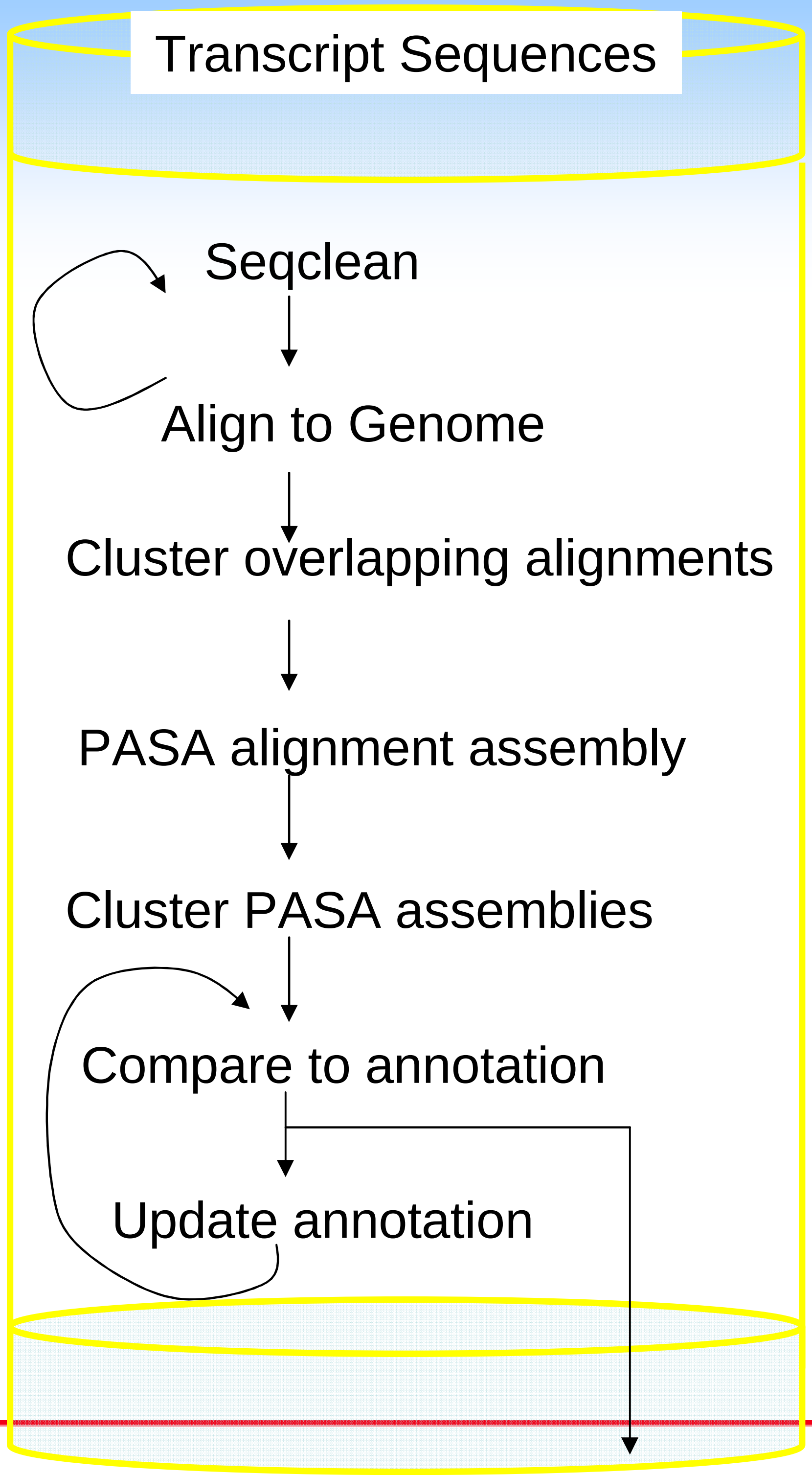


- Align transcripts to genome
- Assemble the alignments

  PASA: **P**rogram to **A**ssemble **S**pliced **A**lignments
- Compare alignment assemblies to existing annotations, suggest updates

PASA Pipeline

Genome Based Alignment and Assembly of cDNA and EST Sequences



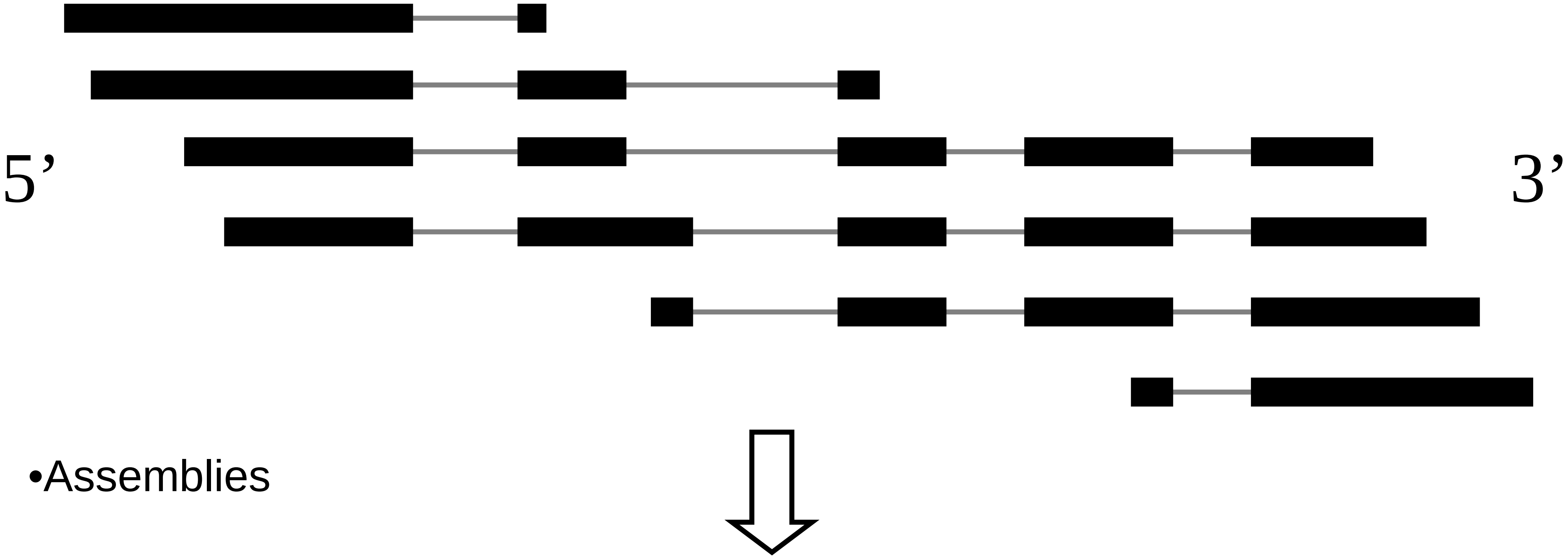
J. Craig Venter INSTITUTE

Alignment Assembly

- Consolidate overlapping alignments into gene structures with maximal evidence support.
- (Maximum evidence) ~ (Maximum # alignments)
- Goal: find maximal assembly of compatible alignments.

Alignment Assembly using PASA:

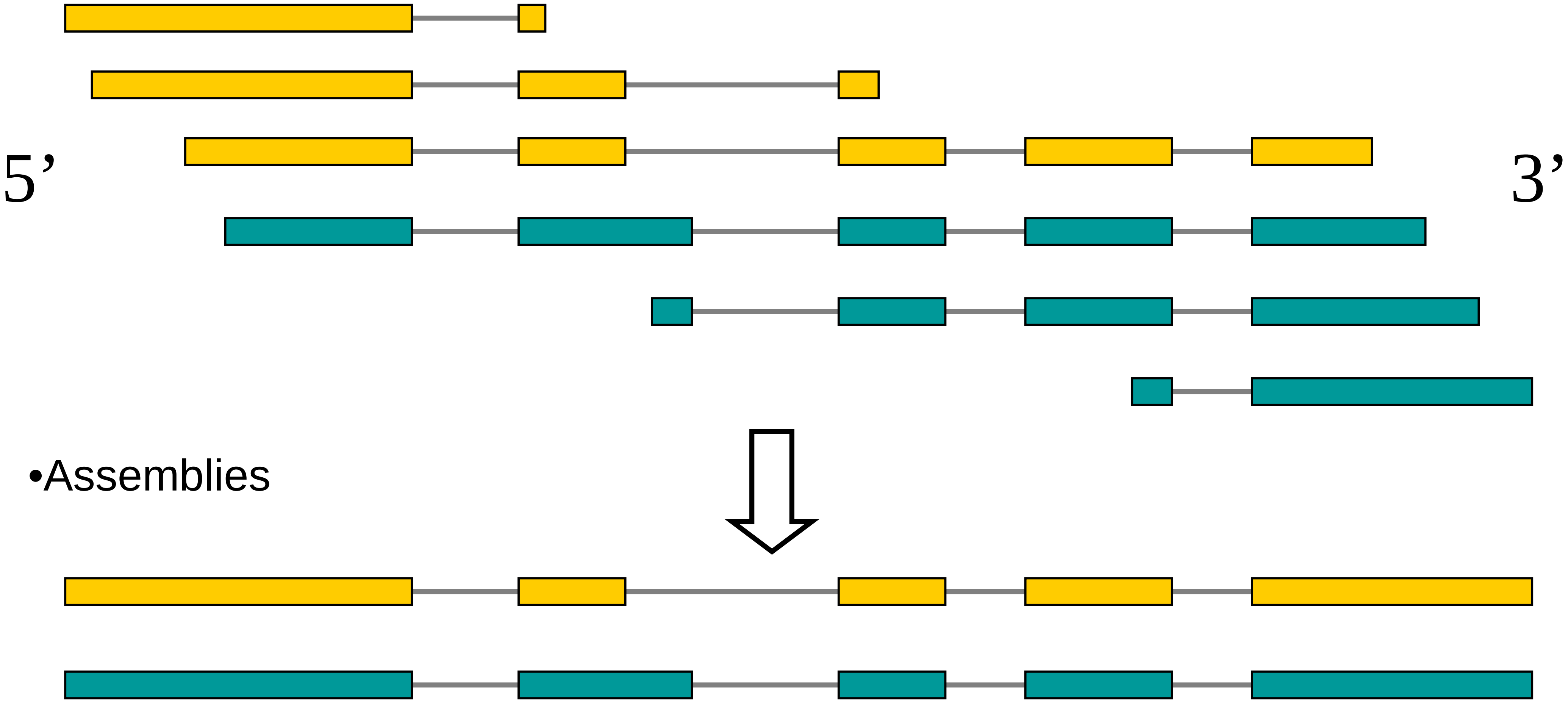
Program to Assemble Spliced Alignments



- Assemblies

Maximally Assemble Compatible Alignments

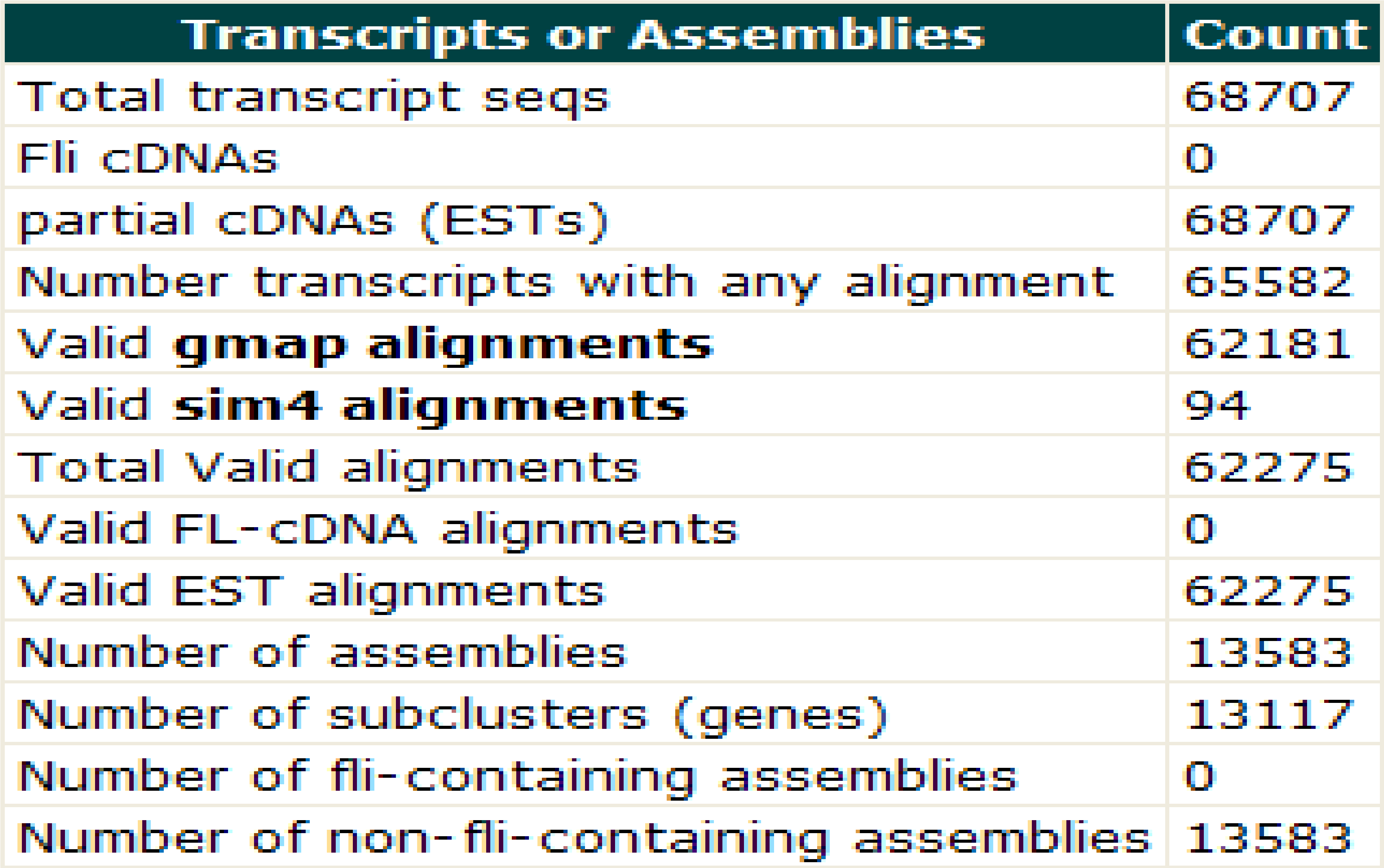
Alignment Assembly using PASA:
Program to Assemble Spliced Alignments



- Assemblies

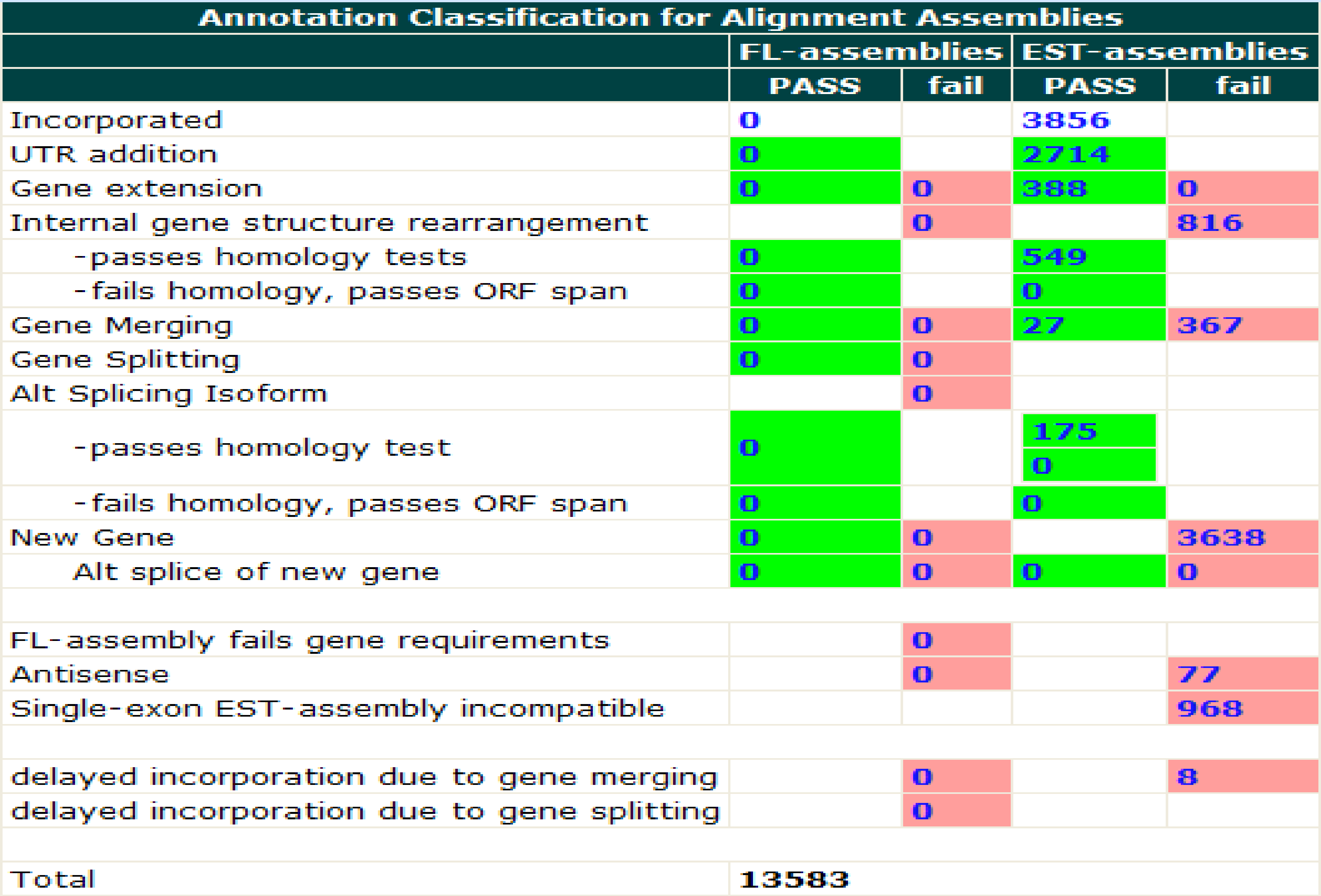
J. Craig Venter INSTITUTE Maximally Assemble Compatible Alignments

PASA Main Page : An Example



Transcripts or Assemblies	Count
Total transcript seqs	68707
Fli cDNAs	0
partial cDNAs (ESTs)	68707
Number transcripts with any alignment	65582
Valid **gmap alignments**	62181
Valid **sim4 alignments**	94
Total Valid alignments	62275
Valid FL-cDNA alignments	0
Valid EST alignments	62275
Number of assemblies	13583
Number of subclusters (genes)	13117
Number of fli-containing assemblies	0
Number of non-fli-containing assemblies	13583

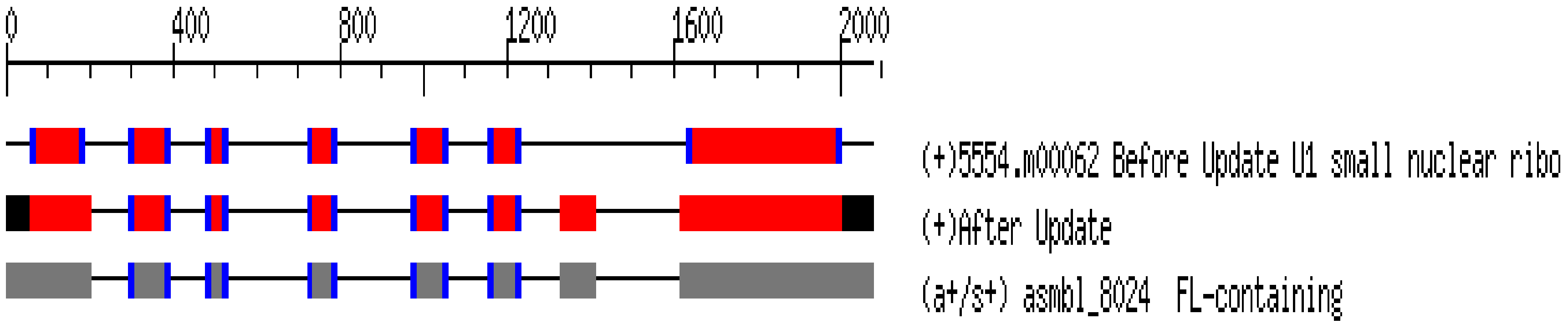
Annotation classification



Annotation Classification for Alignment Assemblies				
	FL-assemblies		EST-assemblies	
	PASS	fail	PASS	fail
Incorporated	0		3856	
UTR addition	0		2714	
Gene extension	0	0	388	0
Internal gene structure rearrangement		0		816
-passes homology tests	0		549	
-fails homology, passes ORF span	0		0	
Gene Merging	0	0	27	367
Gene Splitting	0	0		
Alt Splicing Isoform		0		
-passes homology test	0		175 0	
-fails homology, passes ORF span	0		0	
New Gene	0	0		3638
Alt splice of new gene	0	0	0	0
FL-assembly fails gene requirements		0		
Antisense		0		77
Single-exon EST-assembly incompatible				968
delayed incorporation due to gene merging		0		8
delayed incorporation due to gene splitting		0		
Total	13583			

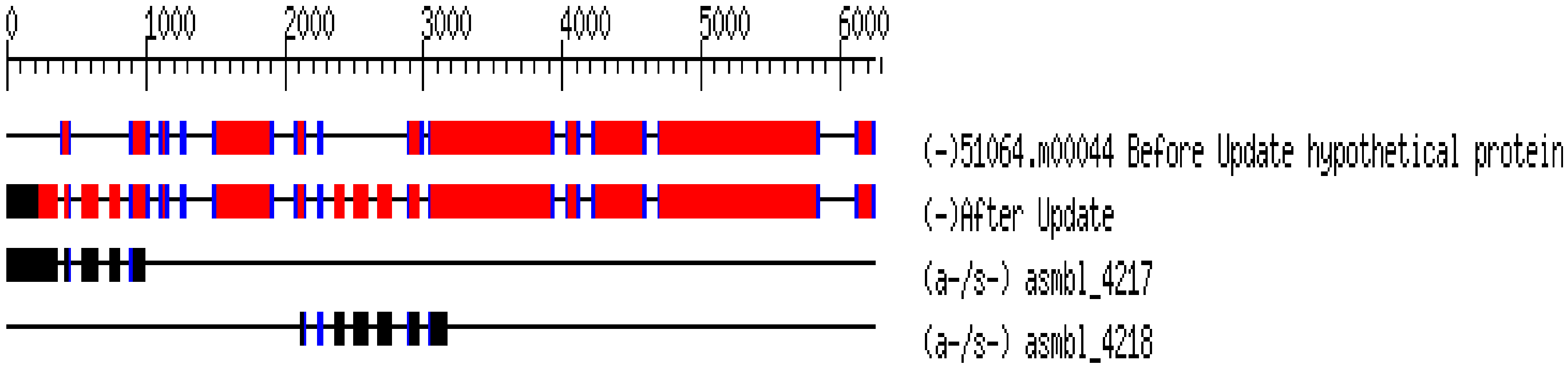
Structural refinement

FL-assembly


0
400
800
1200
1600
2000
(+)5554.m00062 Before Update U1 small nuclear ribo
(+)After Update
(a+/s+) asmbl_8024 FL-containing

Replace existing gene model with FL-assembly

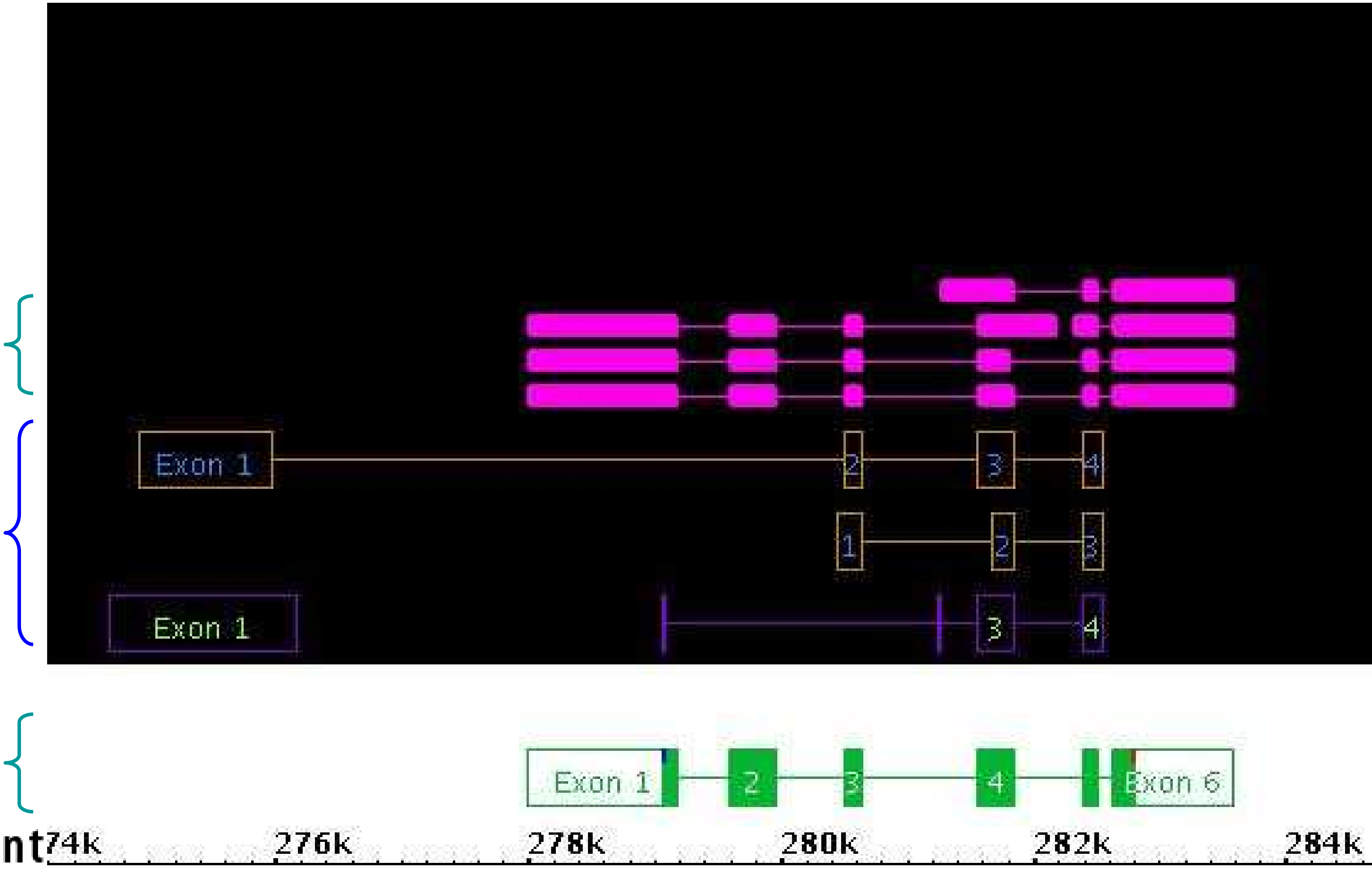
non-FL-assembly



‘Stitch’ non-FL-assembly into gene model

An example

Exons Supported by ESTs not Predicted by Gene finders



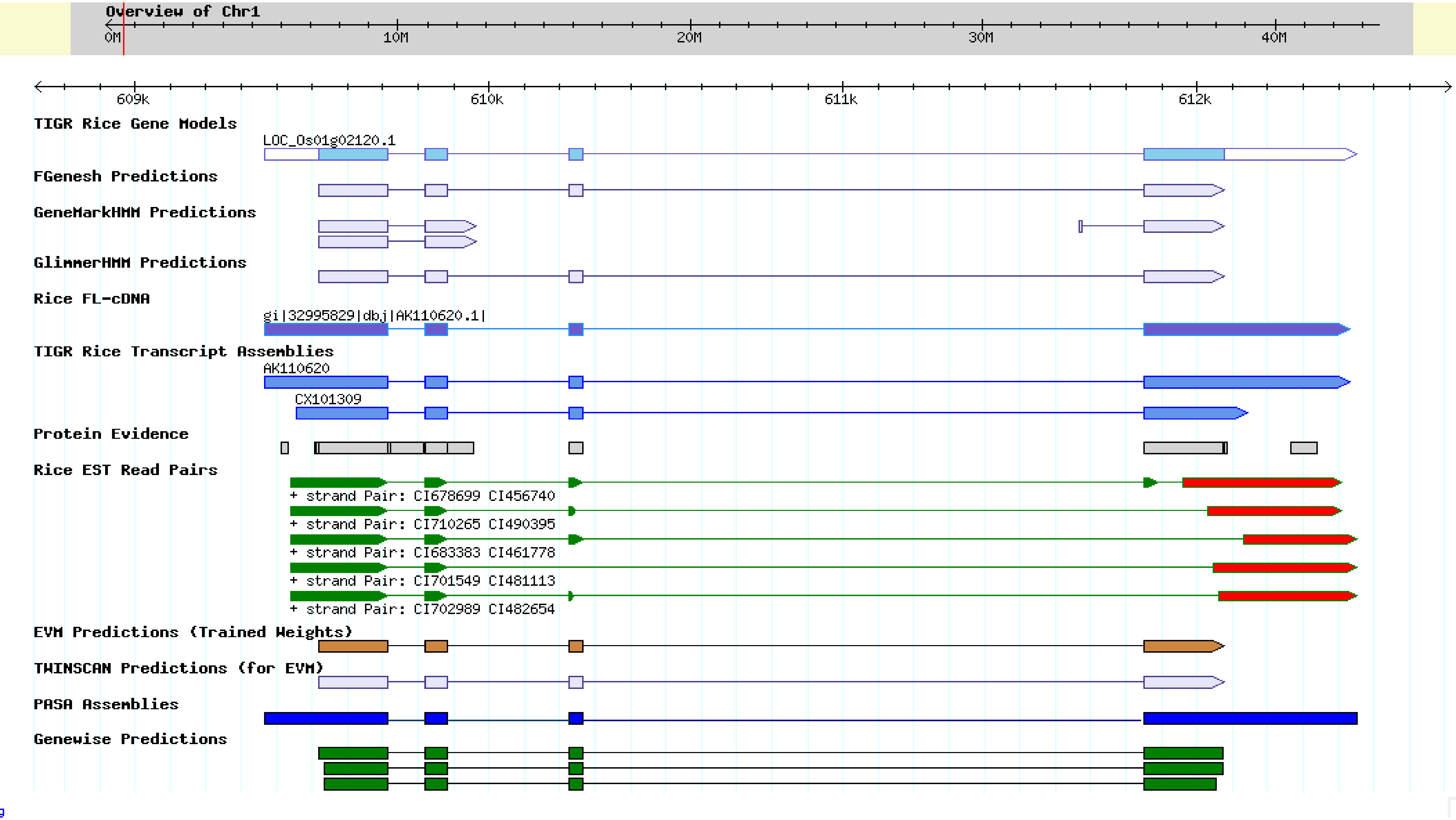
EST Evidence (PASA)

Gene Predictions

Gene Model

Load and view automated gene structures

☆ Hands on this afternoon


Overview of Chr1
0M
10M
20M
30M
40M
609k
610k
611k
612k
TIGR Rice Gene Models
LOC_Os01g02120.1
FGenesh Predictions
GeneMarkHMM Predictions
GlimmerHMM Predictions
Rice FL-cDNA
gi|32995829|dbj|AK110620.1|
TIGR Rice Transcript Assemblies
AK110620
CX101309
Protein Evidence
Rice EST Read Pairs
+ strand Pair: CI678699 CI456740
+ strand Pair: CI710265 CI490395
+ strand Pair: CI683383 CI461778
+ strand Pair: CI701549 CI481113
+ strand Pair: CI702989 CI482654
EVM Predictions (Trained Weights)
TWINSCAN Predictions (for EVM)
PASA Assemblies
Genewise Predictions

How good are automated gene predictions?

Automated annotation identifies a vast majority of genes but accuracy may be limited

Eukaryotic Gene Prediction is not a Solved Problem

What you are getting is a prediction…

- ❖ Manual curation is often used to assess various types of evidence and improve upon automated gene predictions
- ❖ Ultimately experimental verification is the only way to be sure that a gene structure is correct