-2

Transcript Alignment Assembly and Automated Gene Structure Improvements Using PASA-2

Mathangi Thiagarajan

mathangi@jcvi.org

Rice Genome Annotation Workshop

May 23rd, 2007

J. Craig Venter
INSTITUTE

About PASA

- PASA is an open source free to download software program written by **Brian Haas** (bhaas@jcvi.org)

- Reference :Its original application is described in:

  Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res, 31, 5654-5666.

Topics Outline

- Overview of the PASA Pipeline
- Alignment Assembly Algorithm
- Annotation comparison

FL-cDNAs and ESTs

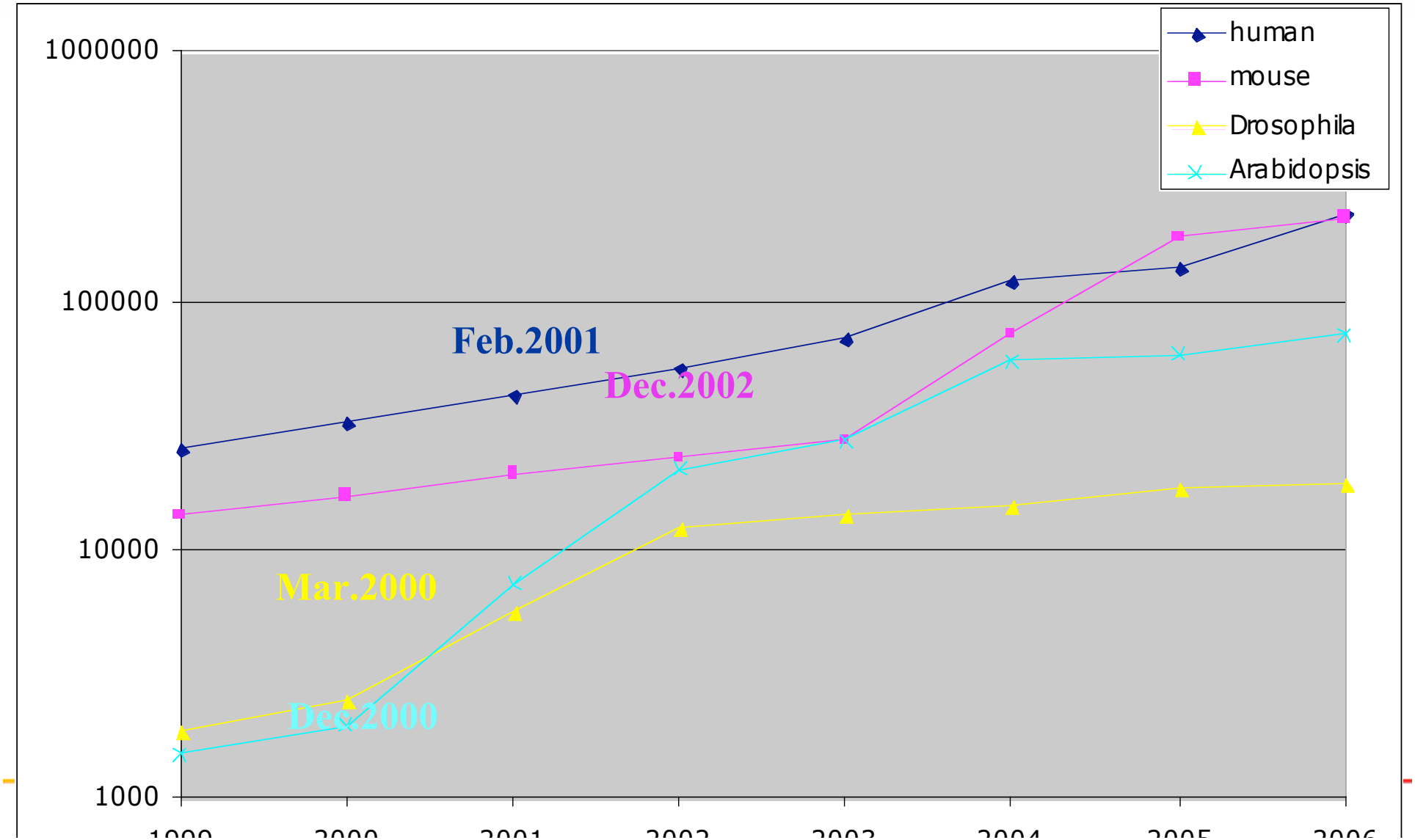
- "**Gold standard**" for gene structure resolution
  - Introns and exons via spliced alignment
- Direct evidence for:
  - Alternative splicing
  - Untranslated regions (UTRs)
  - Polyadenylation sites

The PASA Pipeline

- Automate incorporation of transcript alignments into gene structure annotations

- It was originally developed to refine gene structures in Arabidopsis as part of our Arabidopsis re-annotation effort.
- Since that time, we've expanded the pipeline and applied it to a range of other organisms at TIGR, now with a special focus on Rice.

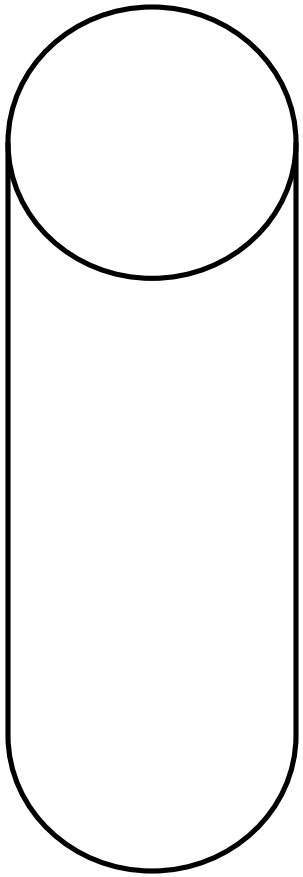
Influxes of mRNA Sequences After Initial Genome Releases



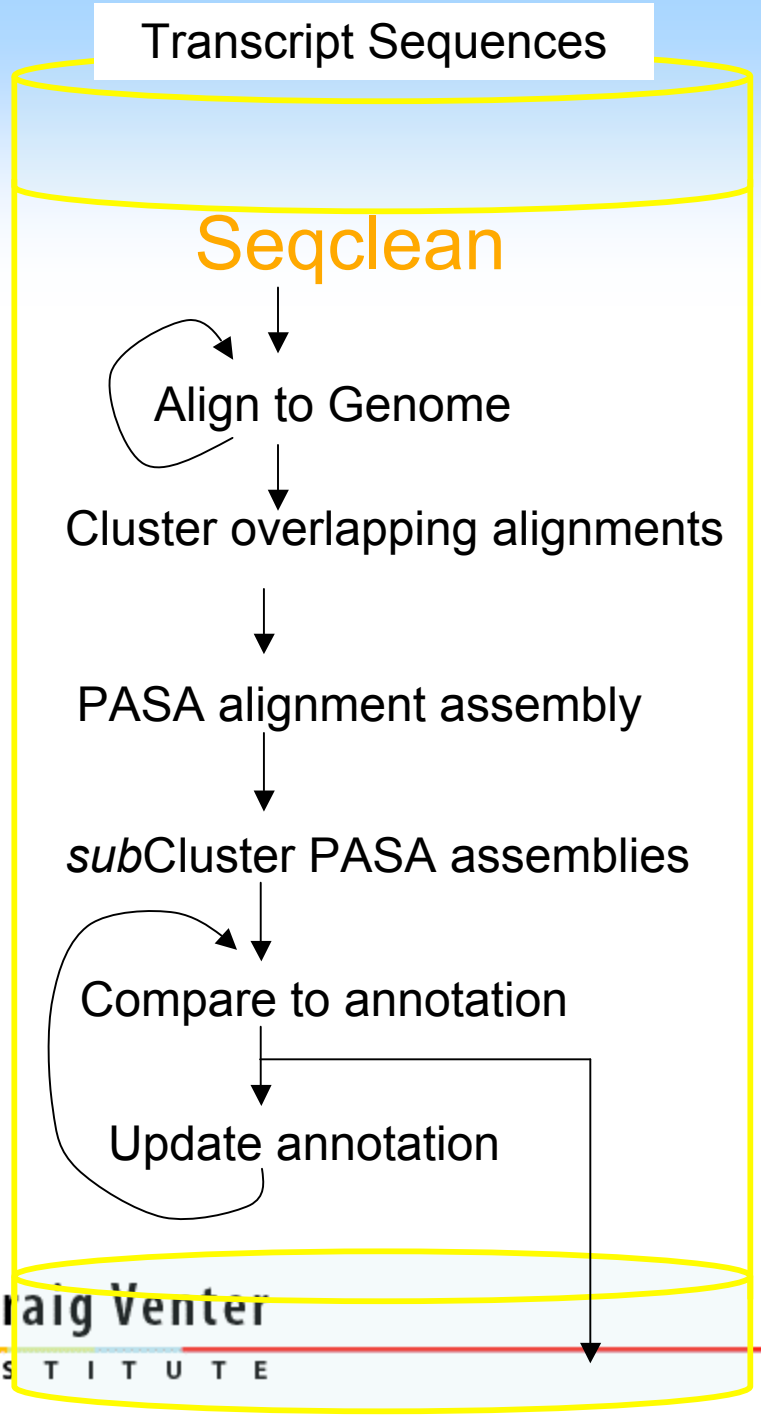
Additionally Found Uses of PASA

- Automated generation of training sets for Gene Finders (Aedes, Aspergillus, Tetrahymena)
- Evaluation of EST libraries (Tetrahymena)
  - examine redundancy within EST library
  - selection of clones for full-length sequencing
- Transitive gene structure annotation for closely related species (Aspergillus sp.)
- Comparing different annotation methods on the same contigs (Plasmodium vivax)
- Cataloging polyA sites for more detailed studies (Arabidopsis, Rice)

The PASA Pipeline [at a glance]

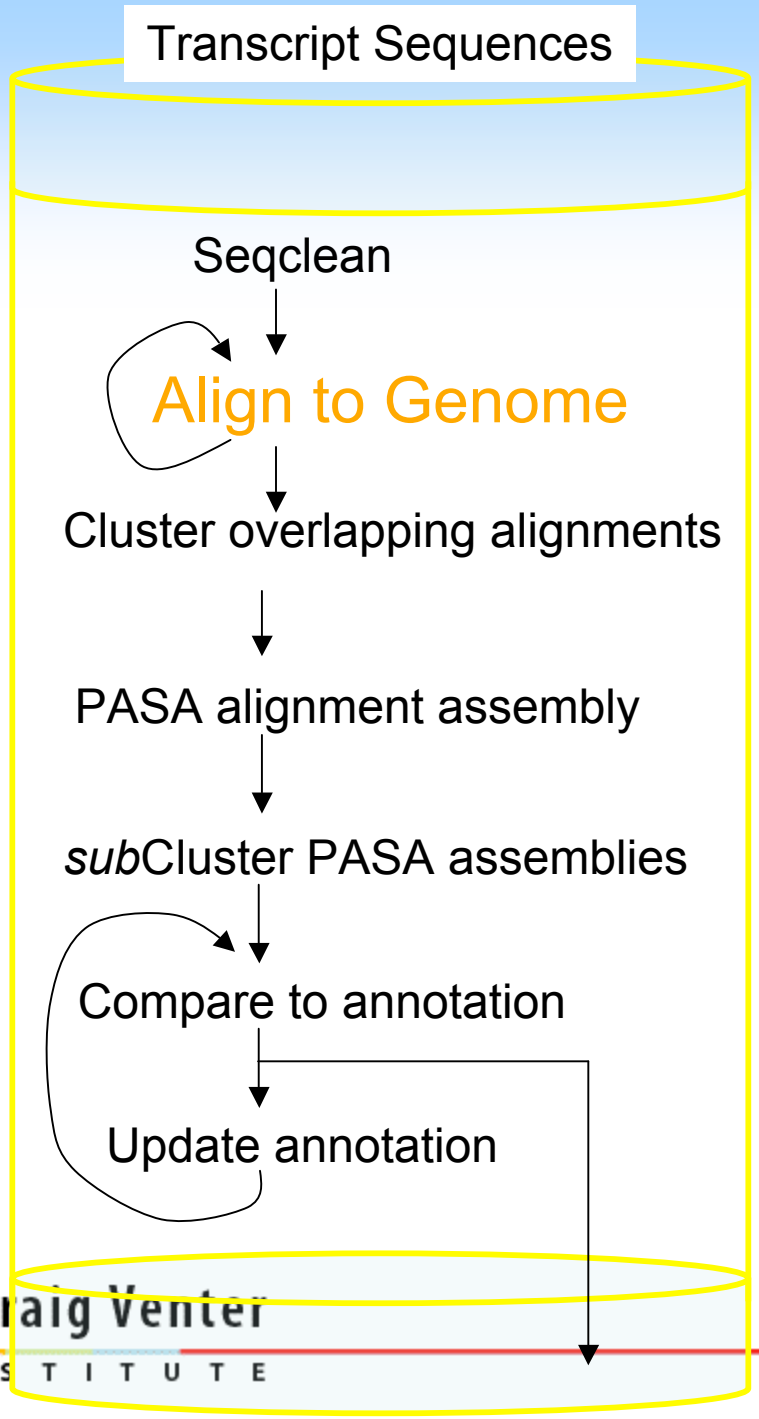


- Align transcripts to genome
- Assemble the alignments
  PASA: Program to Assemble Spliced Alignments
- Compare alignment assemblies to existing annotations, suggest updates



PASA Pipeline

Seqclean (TIGR Gene Indices)
- vector removal
- poly-A identification, stripping
- trash low quality seqs



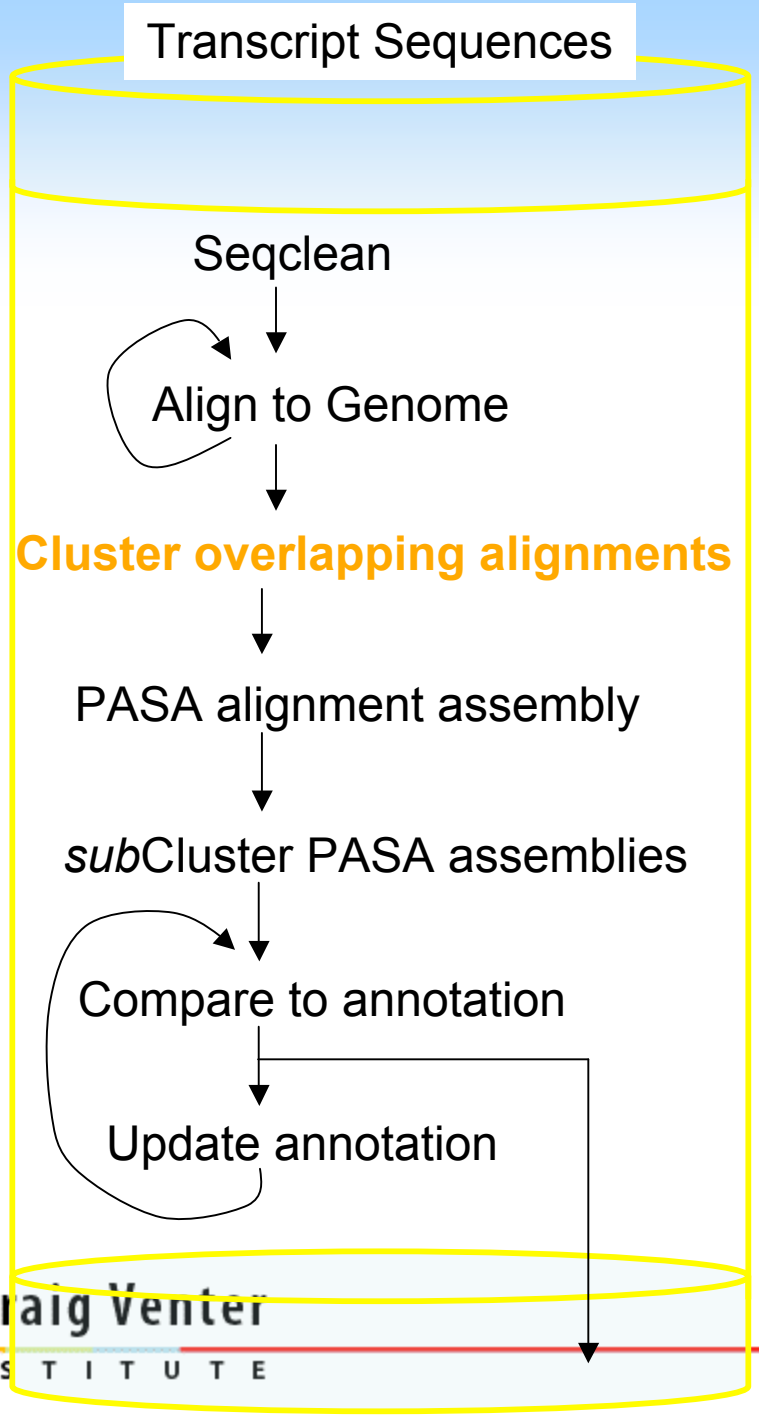
PASA Pipeline

BLAT and sim4 spliced alignments

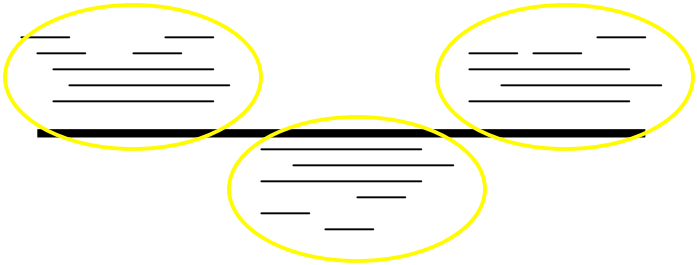
Valid alignment criteria:

- min *95%* Identity
  min *90%* transcript length aligned
  (both configurable parameters)
- consensus splice sites
  - (GT,GC) **donors**
  - AG acceptor
- Assign Transcribed Orientations
- Splice sites
- Polyadenylation sites

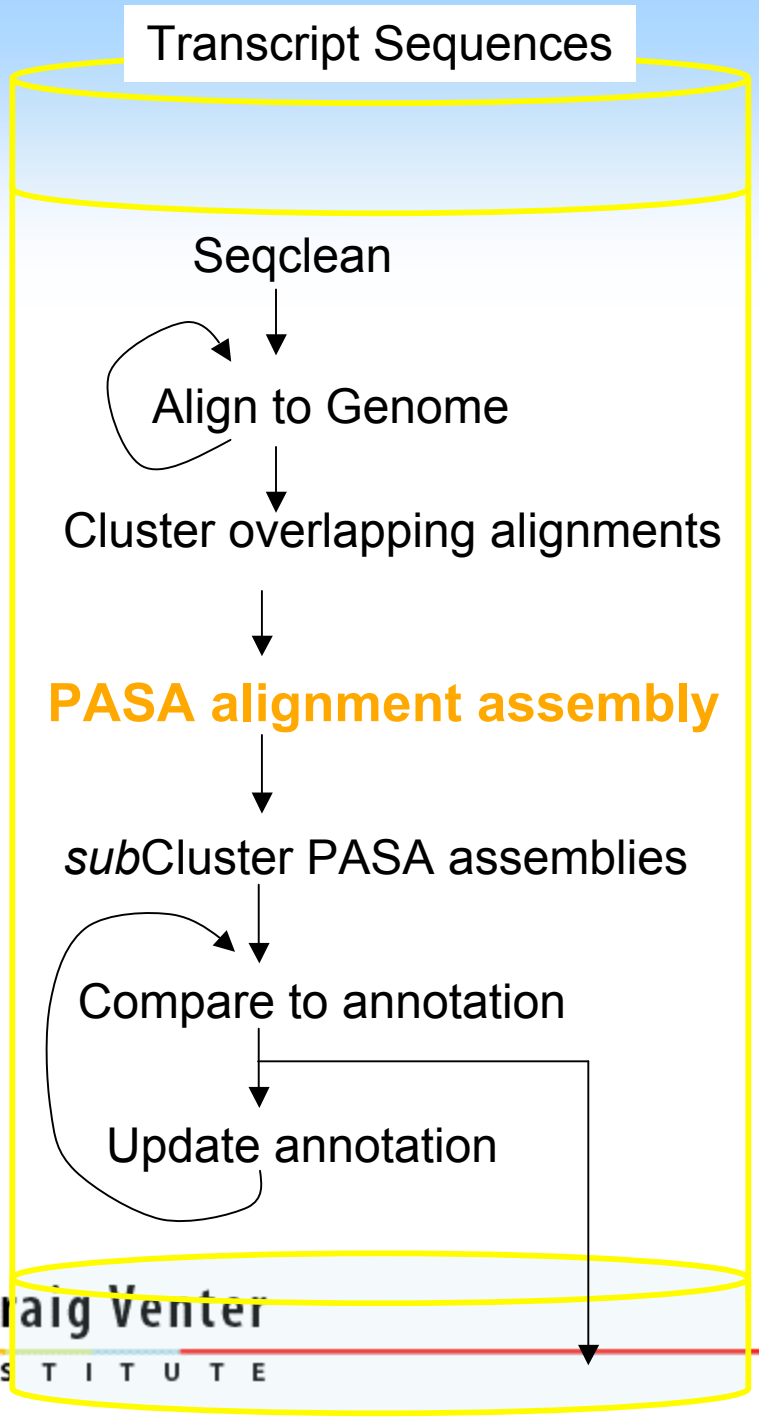
PASA Pipeline



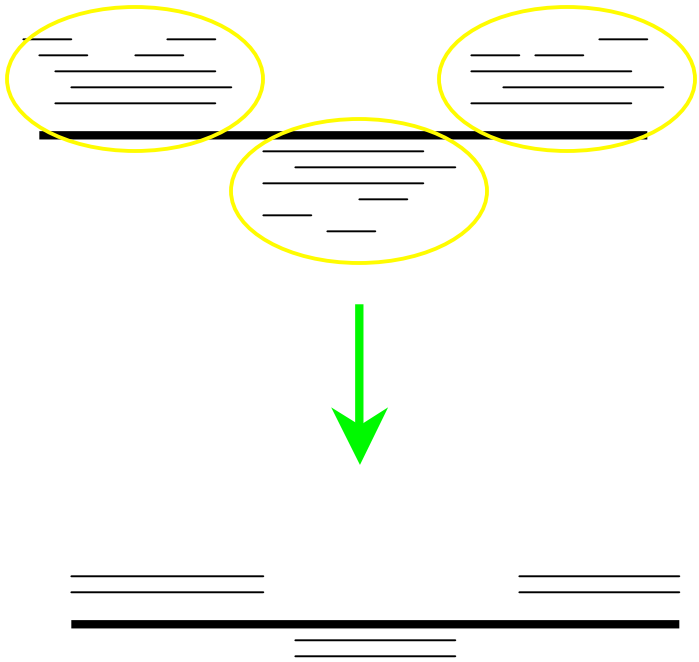
BLAT and sim4 spliced alignments



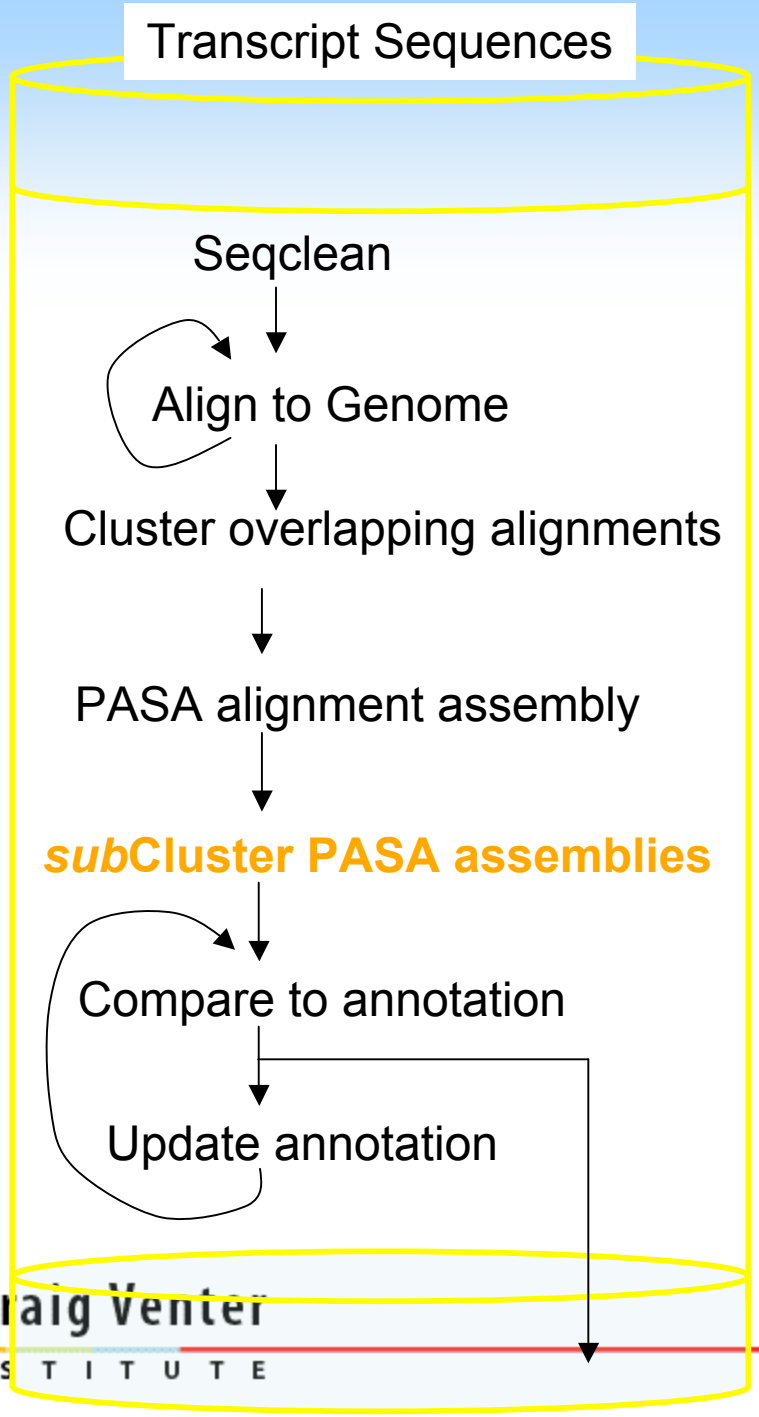
PASA Pipeline



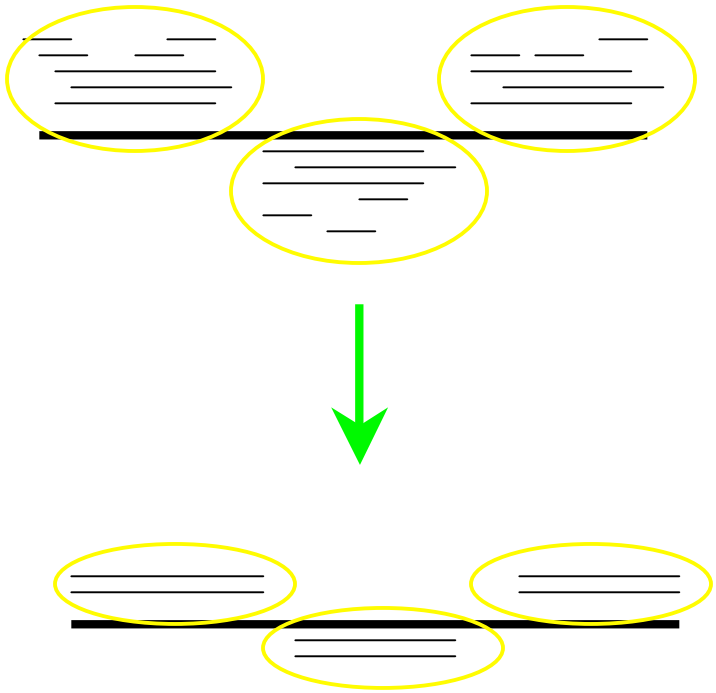
BLAT and sim4 spliced alignments

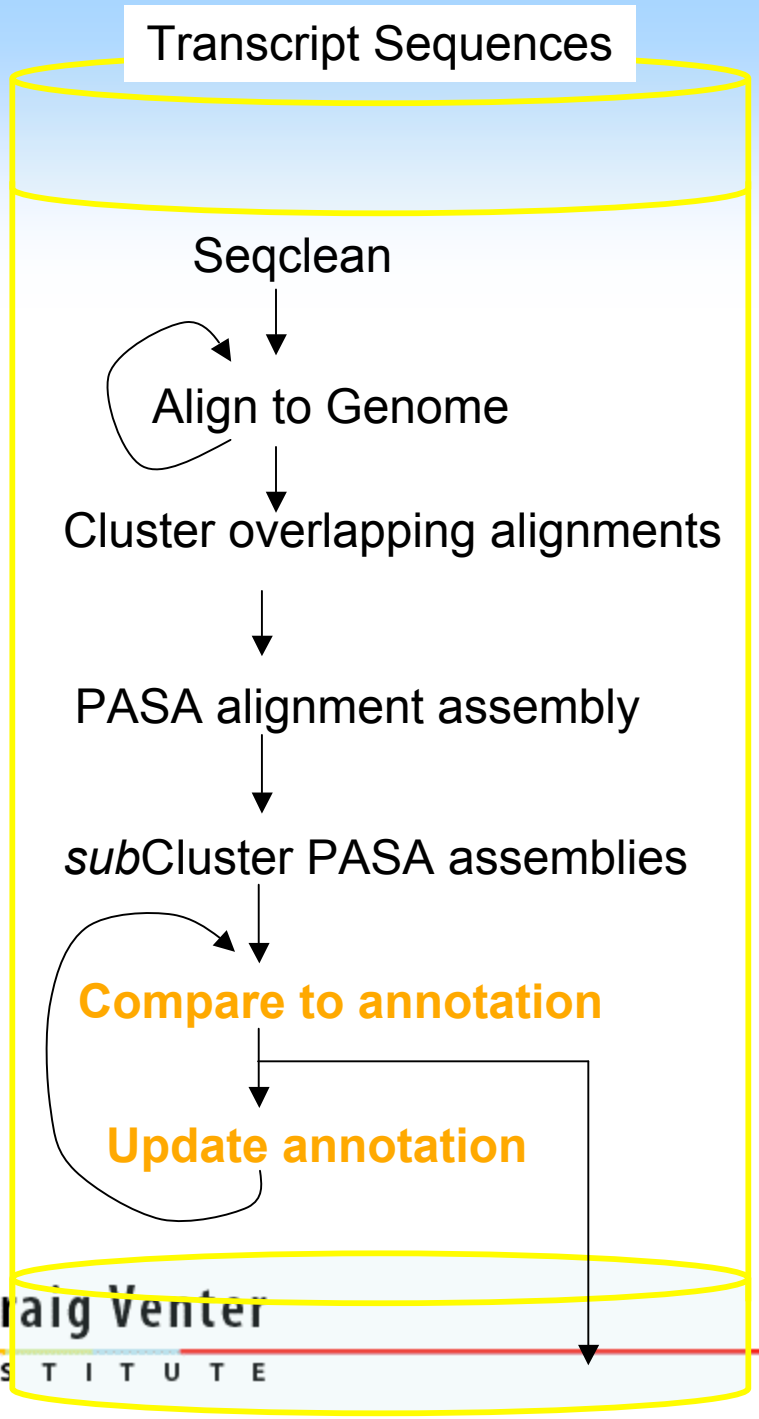


PASA Pipeline



BLAT and sim4 spliced alignments





PASA Pipeline

- **Annotation Comparison**

  FL-cDNAs and ESTs treated separately with different rules for incorporation

- **Annotation Updates**

  -exon modifications

  -alt splice isoform additions

  -gene merges

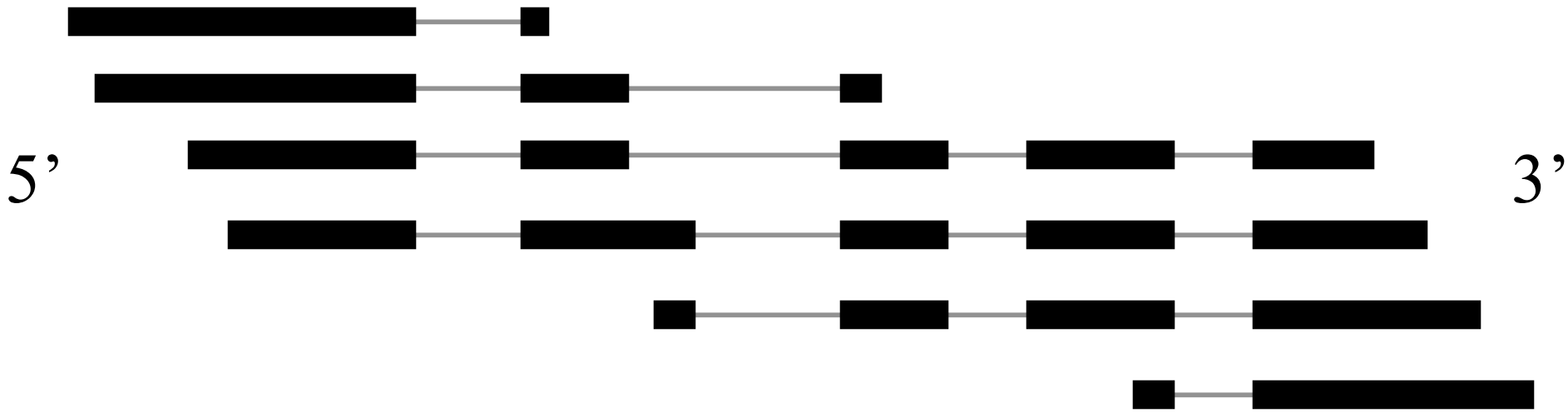
  -gene splits

  -new genes

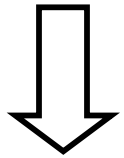
Alignment Assembly

- Maximize evidence supporting gene structures.
- (Maximum evidence) ~ (Maximum # alignments)
- Goal: find maximal assembly of compatible alignments.

Alignment Assembly using PASA:
Program to Assemble Spliced Alignments

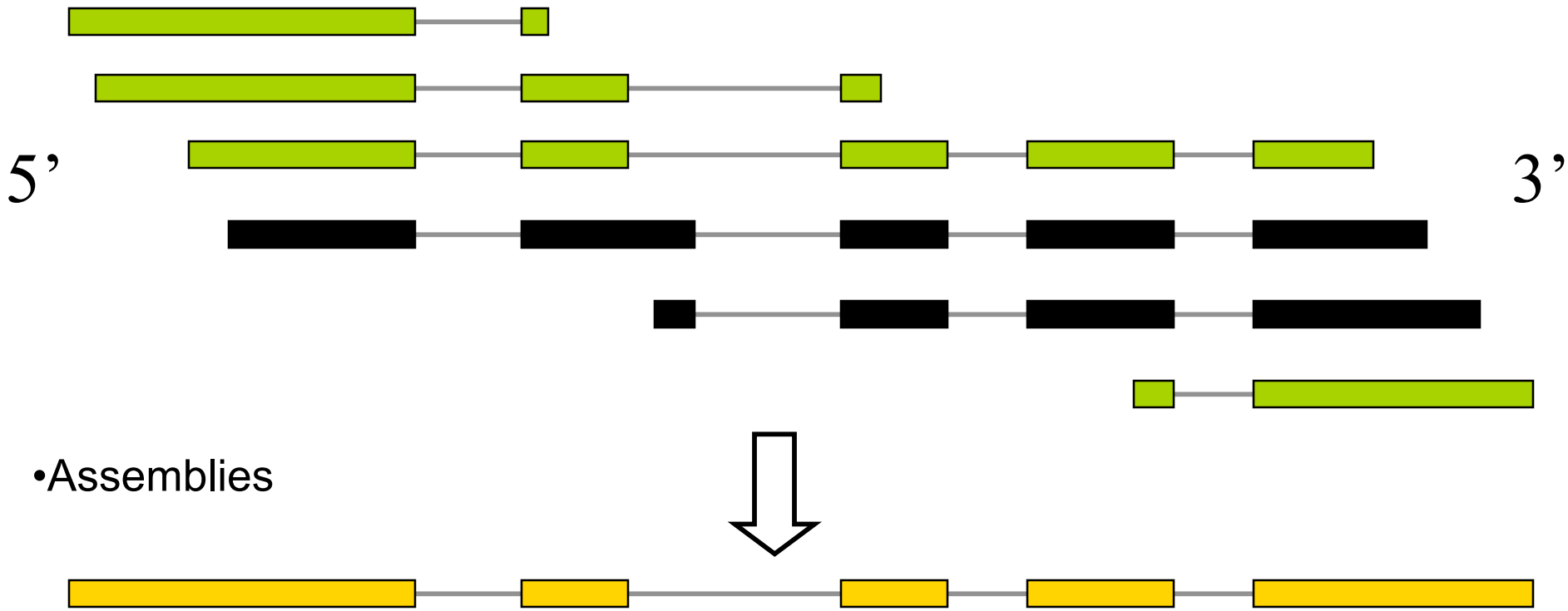


•Assemblies



Maximally Assemble Compatible Alignments

Alignment Assembly using PASA:
Program to Assemble Spliced Alignments

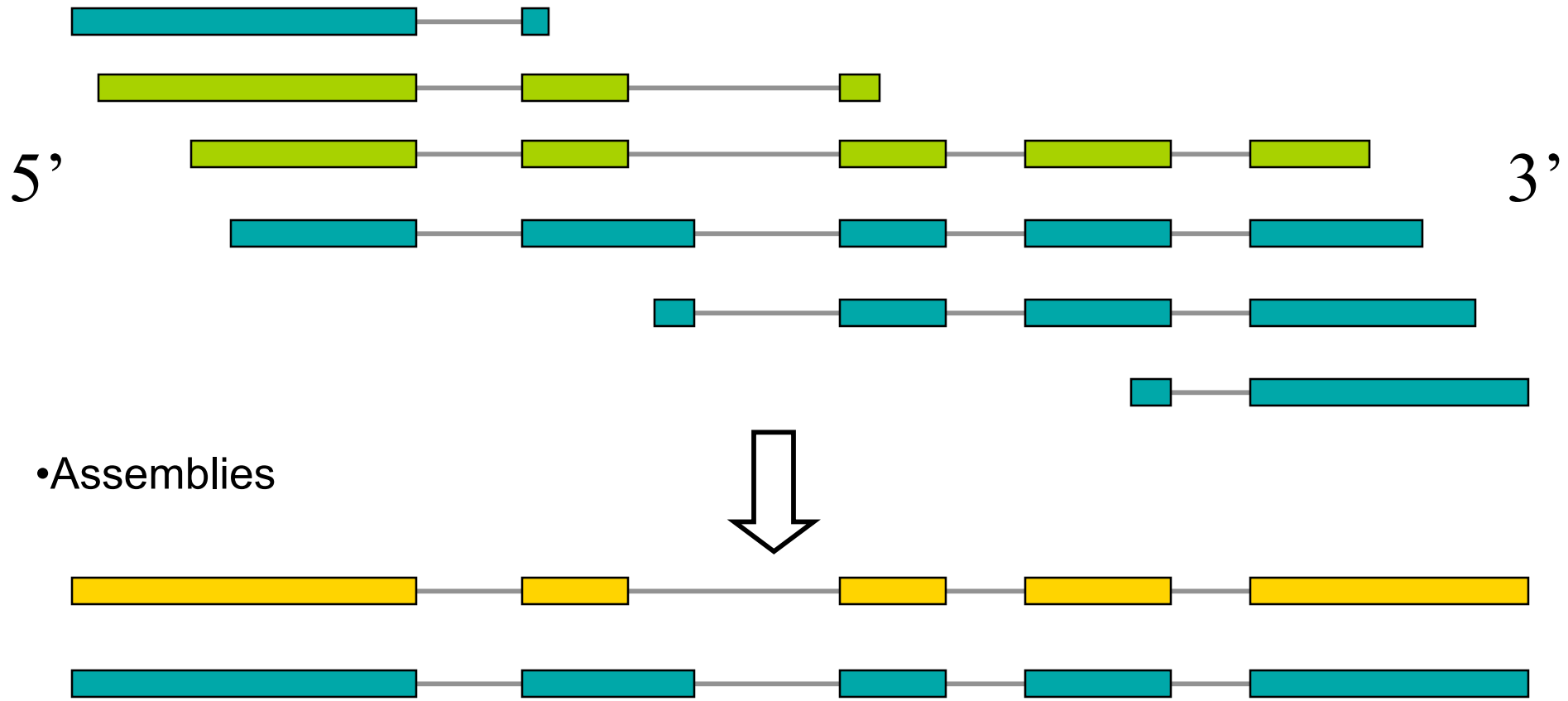


•Assemblies

Maximally Assemble Compatible Alignments

J. Craig Venter
INSTITUTE

Alignment Assembly using PASA:
Program to Assemble Spliced Alignments



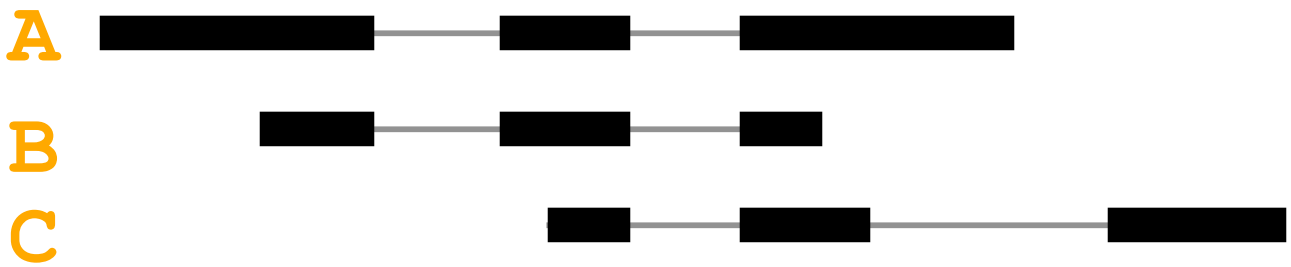
•Assemblies

Maximally Assemble Compatible Alignments

J. Craig Venter
INSTITUTE

PASA Algorithm

- Containments preclude the simple chaining of compatible alignments (B is contained within A)



A ~ B

B ~ C

A !~ C

~ :compatible

!~ :not compatible

PASA Algorithm
Finding the Single Maximal Assembly

- Sort list of alignments by left-most coordinate
- Determine pairwise containments

  C_a = # alignments contained in a, including a

- Determine pairwise compatibilities
- Chain compatible alignments, summing unique containments.

  {Create Left Path Graph, chain compatible alignments from left to right}

  L_a = maximal chain of alignments originating from the left of alignment **a** and ending at **a**.

$$L_a = \max_b \left\{ C_a,\ L_b + C_{a \setminus b} \;\middle|\; \begin{array}{c} b \text{ is compatible with } a, \\ b \text{ is strictly left of } a, \\ a \text{ is not contained within } b \end{array} \right\}$$

Solve by dynamic programming

- Find maximal assembly

$$M_a = \max_b \{L_b\}$$

as the chain with maximal # alignments.

PASA Algorithm
Find Maximal Assemblies for Missing Alignments (Alt Spliced Isoforms)

- Create reciprocal {right path} graph
  {chain compatible alignments from right to left}

  R_a = maximal chain of alignments originating from the Right of alignment **a** and ending at **a**.

$$R_a = \max_b \left\{ C_a,\ R_b + C_{a \setminus b} \;\middle|\; \begin{array}{c} b \text{ is compatible with } a, \\ b \text{ is strictly right of } a, \\ a \text{ is not contained within } b \end{array} \right\}$$

- For each missing alignment **a**, find the maximal assembly containing **a**

$$M_a = \max_b \{ L_b + R_b - C_b \mid b \text{ contains } a \}$$ (restated as sum of left and right paths)

Annotation Comparison

The PASA Pipeline [Capabilities]

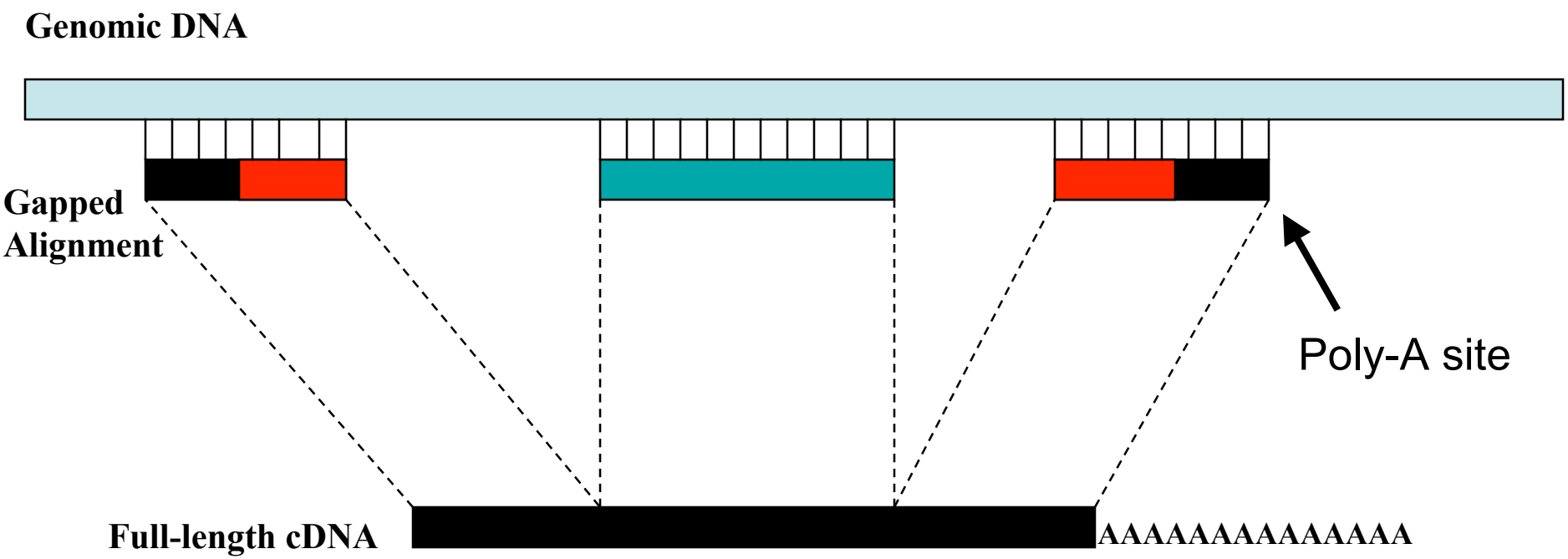
- Then (NAR, 2003) :
  - Update gene structures:
    - Changes in introns and exons
    - UTR additions
  - Model additional gene structures
    - Alternative splicing isoforms
    - New gene models

- Now, PASA-2 (above plus following enhancements) :
  - Gene merging
  - Gene splitting
  - Antisense classification
  - Polyadenylation sites

Incorporation of PASA assemblies into the annotation

- FL-assemblies
  - contain at least one FL-cDNA, expected to encode all exons, complete protein, possibly UTRs.
- non-FL-assemblies
  - encode part of a gene:
    - part of one or more exons
    - potentially UTRs.

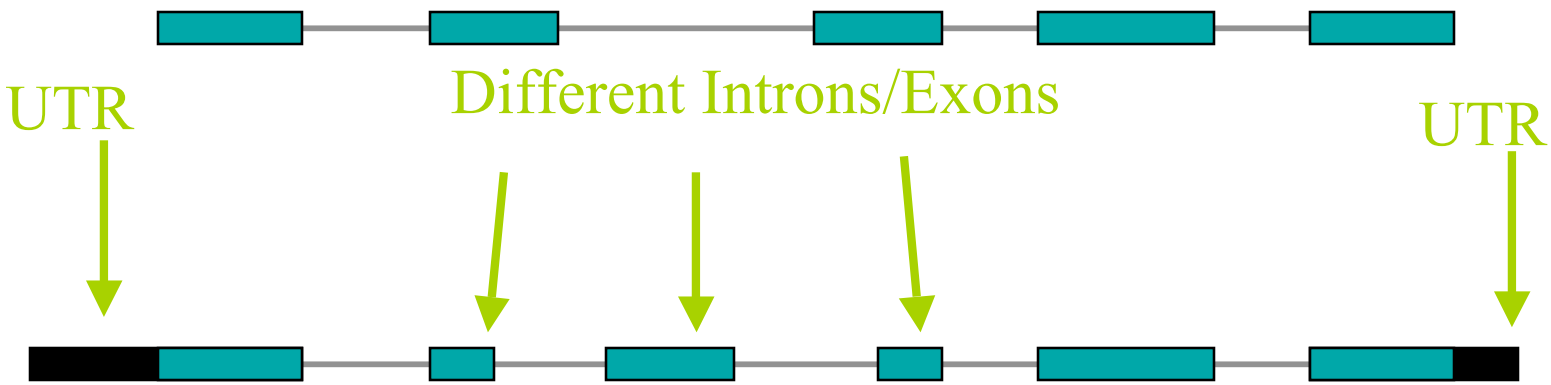
Full-length cDNAs Provide Complete Gene Structures (hence, full-length Assemblies too!)



- cDNA-genome spliced alignment
- ORF reconstruction based on the joined exons.
- UTRs identified.
- Automated process

FL-assembly-based updates

Existing model:



FL-assembly-based model:

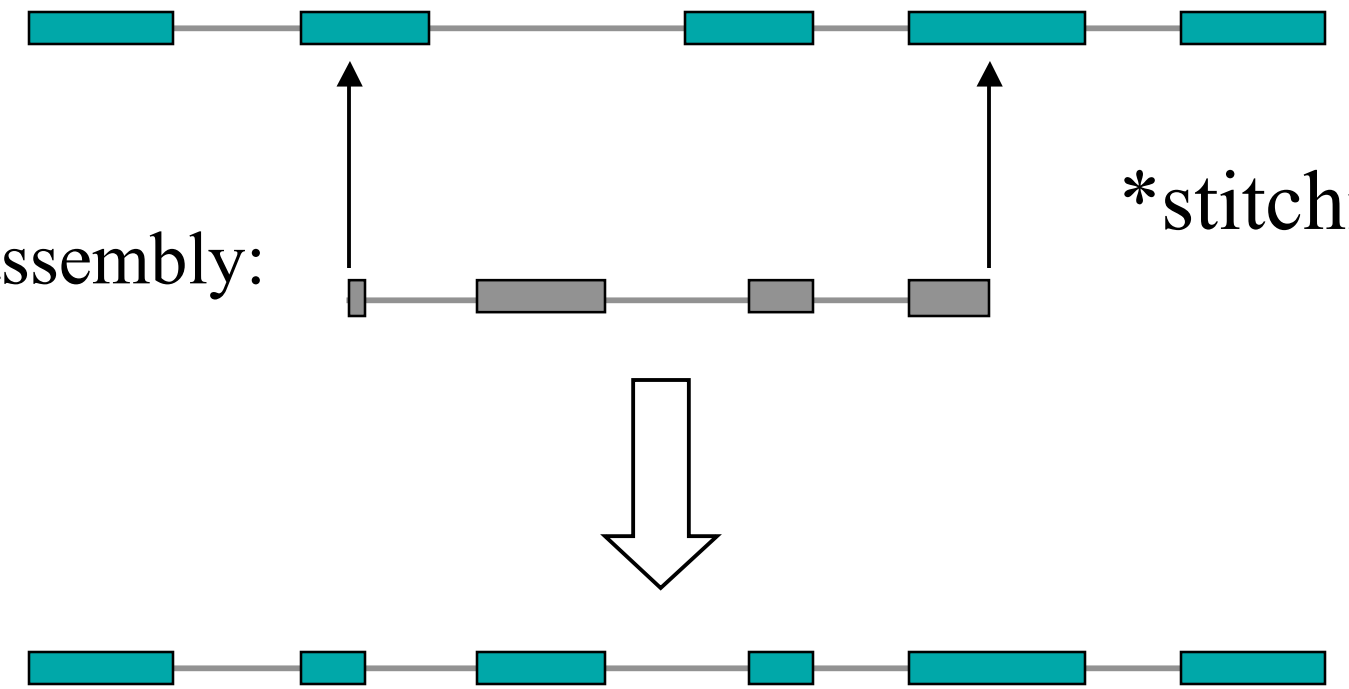
::FL-assembly-based model replaces the existing model

= CDS

= cDNA

Non-FL-assembly-based updates

Existing model:

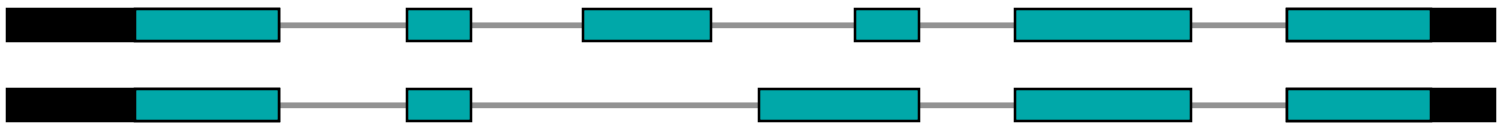


Non-FL-assembly:

stitching

Stitched product replaces existing model

Alternative Splicing
(incompatible alignment assemblies)



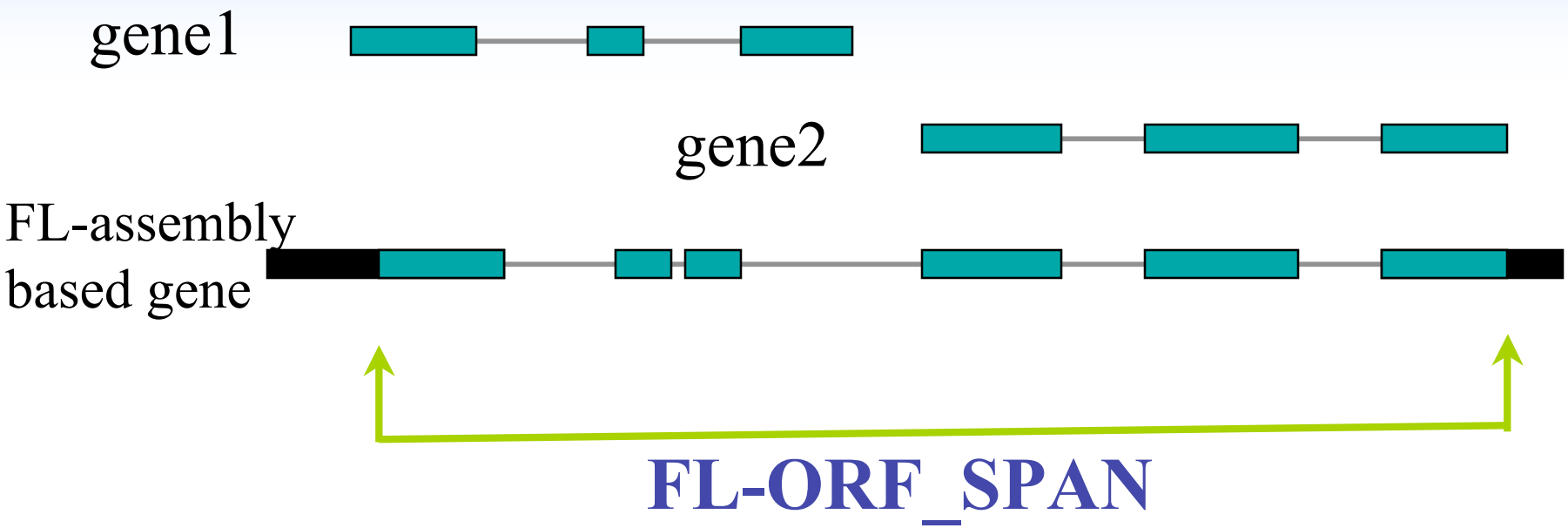
- Sets of mutually incompatible alignment assemblies
  - Multiple FL-assemblies
  - FL-assembly(s) and non-FL-assembly(s)
  - Non-FL-assemblies (*pre-existing gene model required)

Minimize Corruption or Pollution of Existing Annotations

- Requirements of a FL-assembly
  - Min ORF size requirement
    - *MIN_PERCENT_PROT_CODING* (ie. 40%)
    - *MIN_FL_ORF_SIZE* (ie. 100 aa)
  - Max # UTR exons (ie. 2 or 3)
    - *MAX_UTR_EXONS*
- Requirements of an annotation update
  - Compared to existing model, must pass validation tests:
    - **Length test** (ie. must encode a protein at least 70% the length of the current one)
      - *Maybe trust FL assemblies more than ESTs; can set stringencies separately:
      - MIN_PERCENT_LENGTH_FL_COMPARE (involving FL-assemblies)
      - MIN_PERCENT_LENGTH_NONFL_COMPARE (involving non-FL assemblies)
    - **Homology test** [Fasta Alignment] (ie. 70% identity, 70% length)
      - MIN_PERID_PROT_COMPARE (ie. 70% identity)
      - MIN_PERCENT_ALIGN_LENGTH (ie. 70% of the shorter protein length)

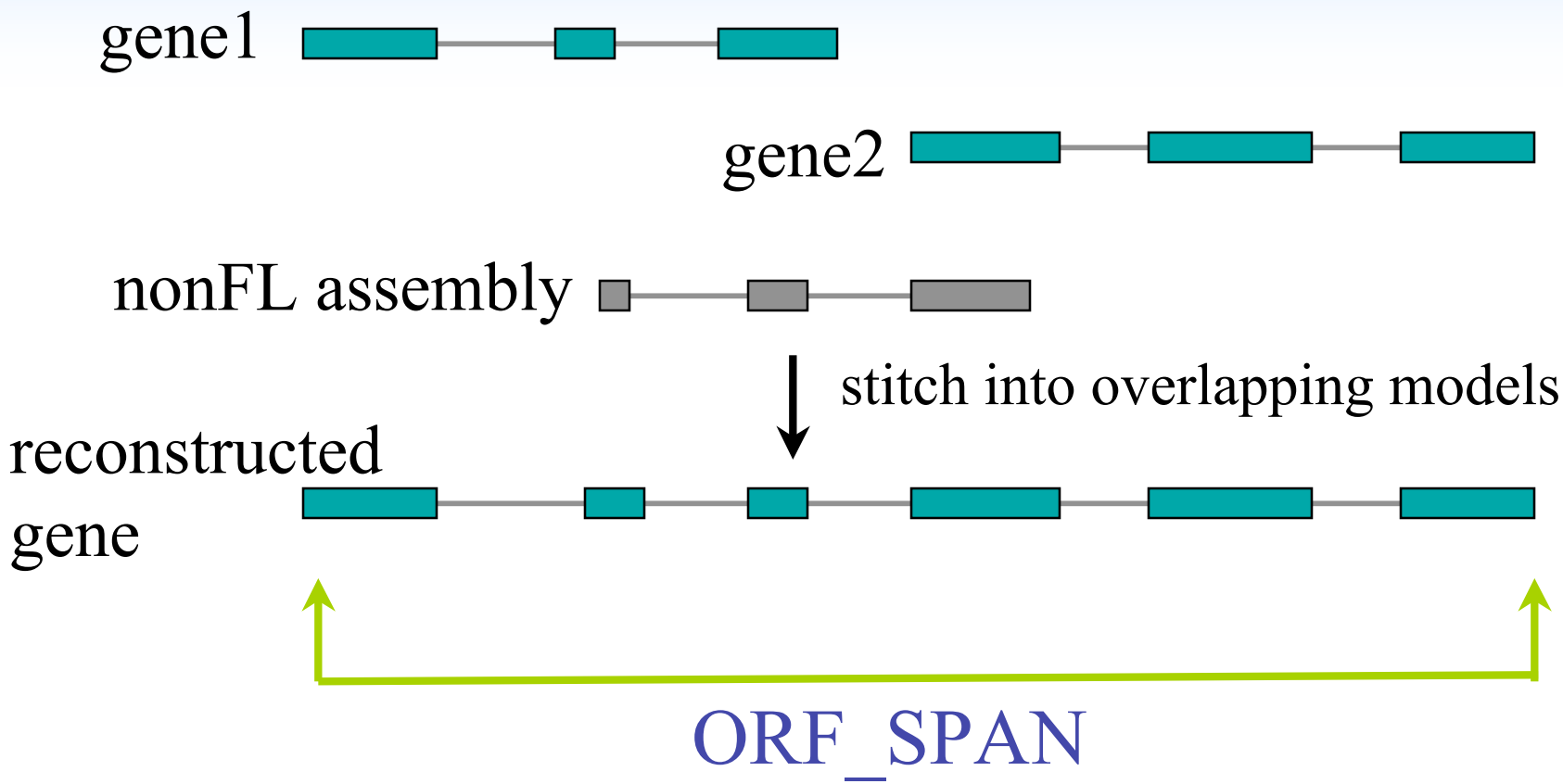
* all user-configurable parameters, option names shown in italics

Enhancements: Gene Merging (FL-cDNA)



- If FL-ORF_SPAN overlaps both gene1 and gene2 [, ... geneX] by at least M*IN_PERCENT_OVERLAP_GENE_REPLACE*, gene1 and gene2 [,... geneX] are to be merged and replaced by the FL-assembly based gene.

Enhancements: Gene Merging (non-FL)



- Same rule as before, using ORF_SPAN and MIN_PERCENT_OVERLAP_GENE_REPLACE

Enhancements: Gene Splitting

Existing gene

FL-assemblies

- Requires
  - multiple FL-assemblies from distinct sub clusters map to the same gene
  - have the same transcribed orientation,
  - and the min and max of the new ORFs must cover at least *MIN_PERCENT_OVERLAP_GENE_REPLACE* of the gene to be split.

Gene Merging and Gene Splitting

- Homology (used loosely) between the existing gene and the replacement is not required.
- Only require that the locus of interest continues to be covered by ORFs.
- Why?
  - Merged and split genes may appear very different from the existing [predicted] gene.
  - One of the split products may look quite similar to the preexisting gene, but the other may not.
  - Our experience is that the existing methodology of splitting and merging works quite well, and we haven't needed to explore additional methods.

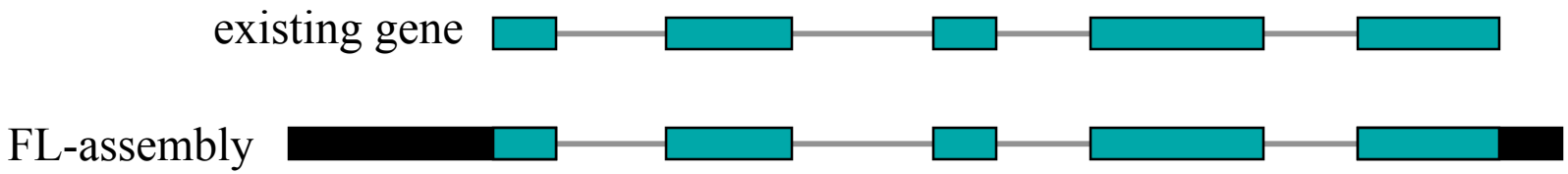
Want more aggressive updates?

- Besides merging and splitting, individual gene updates must pass the homology test. Failures require manual inspection.
- But, many that fail homology may still provide reasonable, and improved gene structure updates.
- Option (flag):
  - *STOMP_HIGH_PERCENTAGE_OVERLAPPING_GENE*
  - If update fails the homology test, consider the ORF_SPAN alone.
  - if ORF_SPAN > MIN_PERCENT_OVERLAP_GENE_REPLACE, allow update to occur.

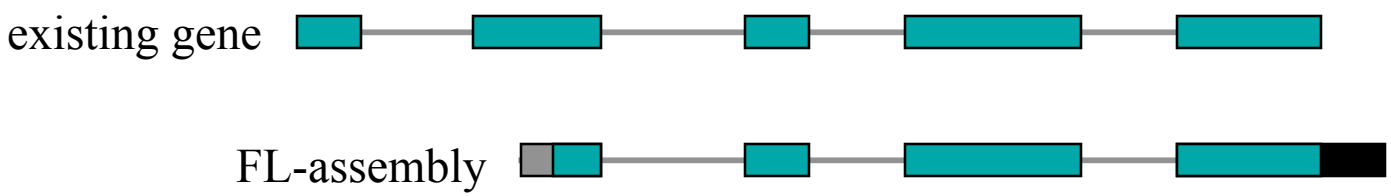
Called STOMPing

Trusting the FL-Status

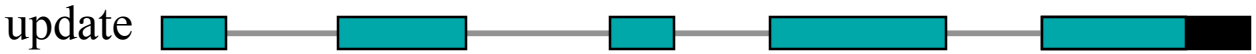
- Ideally, FL-transcripts are full length!



- But, often:



  - Solution: If a FL-assembly is compatible with an existing gene annotation, treat it as non-FL



* option *TRUST_FL_STATUS*, by default, disabled.

Example Application of PASA to Rice

Database rice_genome_annot_06192005 Contents

Total cDNAs/ESTs (mapped to genome using blat)	412689
Fli cDNAs (mapped)	32874
non-Fli cDNAs (ESTs) (mapped)	379815
Valid **Blat alignments**	330835
Valid **Sim4** alignments	24119
Total Valid alignments	354954
Valid FL-cDNA alignments	30929
Valid EST alignments	324025
Number of assemblies	48641
Number of subclusters (genes)	34694
Number of fli-containing assemblies	23633
Number of non-fli-containing assemblies	25008

Describe alignment assemblies

Describe subclusters of assemblies

Retrieve alignment assembly tentative cDNA sequences

Click here to search the database.

Construct customized URLs linked from PASA assembly report pages.

Results from Annotation Comparison

(Counting PASA assemblies)

	FL-assemblies		EST-assemblies	
	PASS	fail	PASS	fail
Incorporated	9843		5634	
UTR addition	3921		2432	
Gene extension	324	108	153	0
Internal gene structure rearrangement		1899		4690
-passes homology tests	1570		748	
-fails homology, passes ORF span	105		78	
Gene Merging	238	455	67	498
Gene Splitting	119	44		
Alt Splicing Isoform		413		
-passes homology test	510		565 1544	
-fails homology, passes ORF span	47		187	
New Gene	102	0		3631
Alt splice of new gene	6		20	49
FL-assembly fails gene requirements		2519		
Antisense		1401		894
Single-exon EST-assembly incompatible				3755
delayed incorporation due to gene merging		9		47
delayed incorporation due to gene splitting		16		
Total	48641			

cgi-bin/status_report.cgi

J. Craig Venter INSTITUTE

Gene Comparison Summary

(Counting Genes)

cgi-bin/status_report.cgi

Gene Comparison Overview

A total of 25523 genes mapped to PASA assemblies	
23633 FL-assemblies mapped to **19350** genes	
Successful Incorporations by Genes	
Genes supported by PASA assemblies	**20970**
Genes supported by **16785** FL assemblies	**15199**
Genes supported by **11428** EST assemblies only *(no FL assemblies)*	**5771**
Failed Incorporations by Genes	
Genes failing PASA assembly incorporation	**11161**
*of these, **6608** genes successfully incorporate other PASA assemblies*	
Genes failing FL assembly incorporation	**5612**
*but **1461** of these genes incorporate other FL assemblies successfully.*	
Genes failing EST assemblies only	**5549**

The same gene may be counted in multiple categories.

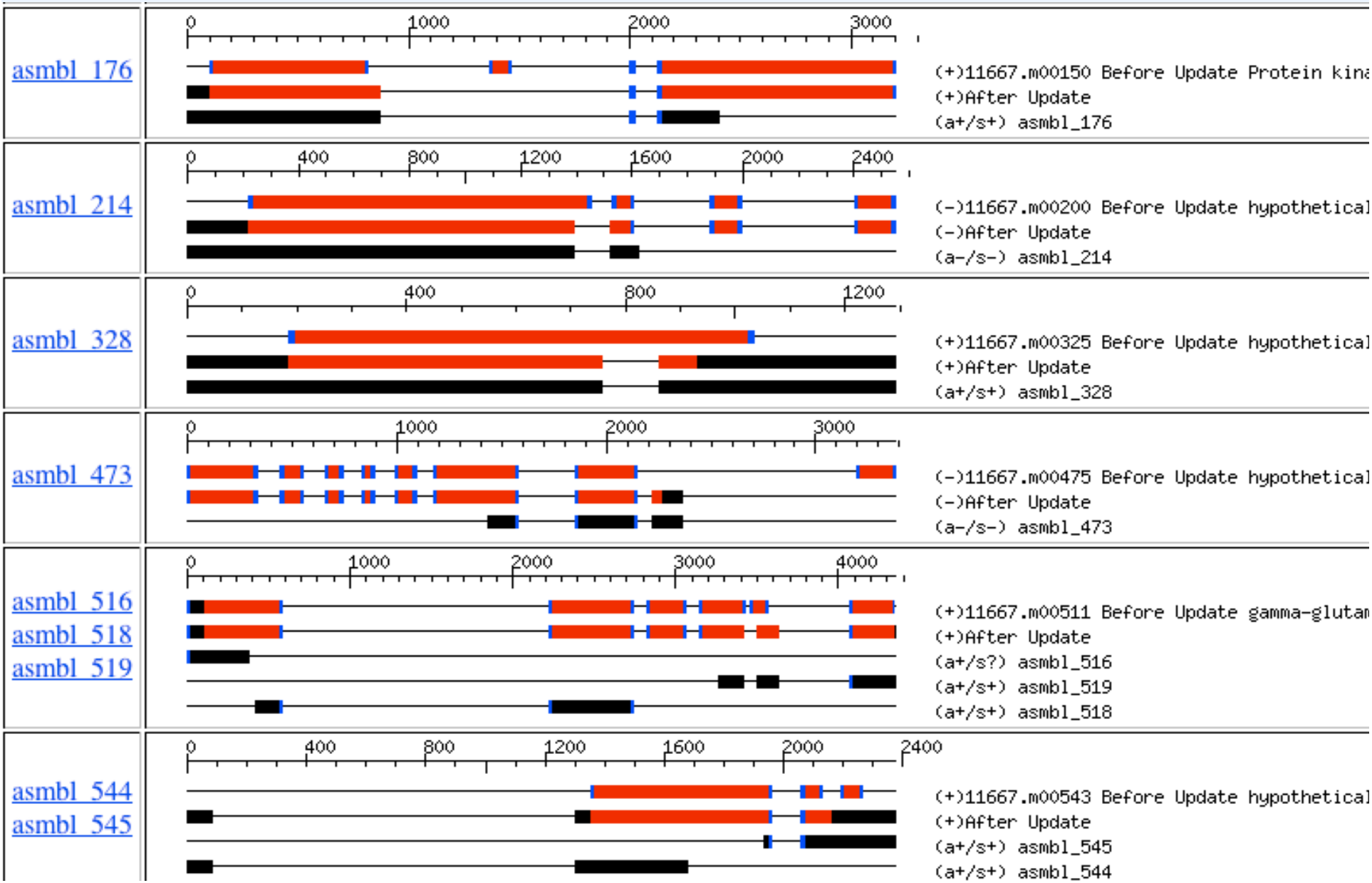
Retrieve Lists of Genes Corresponding to Pass and Fail Categories.

Gene Structure Updates Summary

Proposed Annotation Updates *[counts as unique models, not genes]*

status_id	Description	Num Gene Model Updates	Num Alt Splice isoforms to Add	Num Novel Genes to Add
4	gene-compatible fl-cdna assembly alters UTRs.	3919	0	0
6	gene-compatible fl-cdna assembly alters protein, passed validation.	324	0	0
8	incompatible fl-cdna assembly alignment updates gene structure.	1570	0	0
9	incompatible fl-cdna assembly provides alternative splicing isoform, passes validation.	0	510	0
10	fl-cdna assembly provides a novel gene.	0	0	102
24	FL-cDNA assembly stitched by EST assembly to provide alt splicing isoform.	0	1544	0
26	FL-cDNA spans single gene and allowed to STOMP it.	105	0	0
29	FL-cDNA found capable of merging multiple genes	238	0	0
33	FL-assembly STOMPS new splice isoform	0	47	0
40	FL-cDNAs split single gene into multiple genes	119	0	0
44	FL-cDNA provides alt splicing isoform of a novel gene	0	6	0
13	EST assembly extends UTRs.	2291	0	0
14	EST assembly alters protein sequence, passes validation.	153	0	0
16	EST assembly properly stitched into gene structure.	732	0	0
17	EST assembly stitched into Gene model requires alternative splicing isoform. (deprecated, see status_ids: 24,25)	0	0	0
25	EST assembly stitched into Gene model requires alternative splicing isoform.	0	565	0
27	EST-assembly stitched into a FL-alignment providing new alt splice isoform.	0	20	0
31	EST-stitched assembly STOMPS model lacking transcript support.	78	0	0
32	EST-stitched gene w/preexisting transcript support STOMPS a new alt splicing variation.	0	187	0
36	EST-assembly found capable of merging multiple genes.	67	0	0
	Totals *(some models in multiple classes)*	**9489**	**2879**	**102**

Examining Updates
(clicking any link in the previous report)

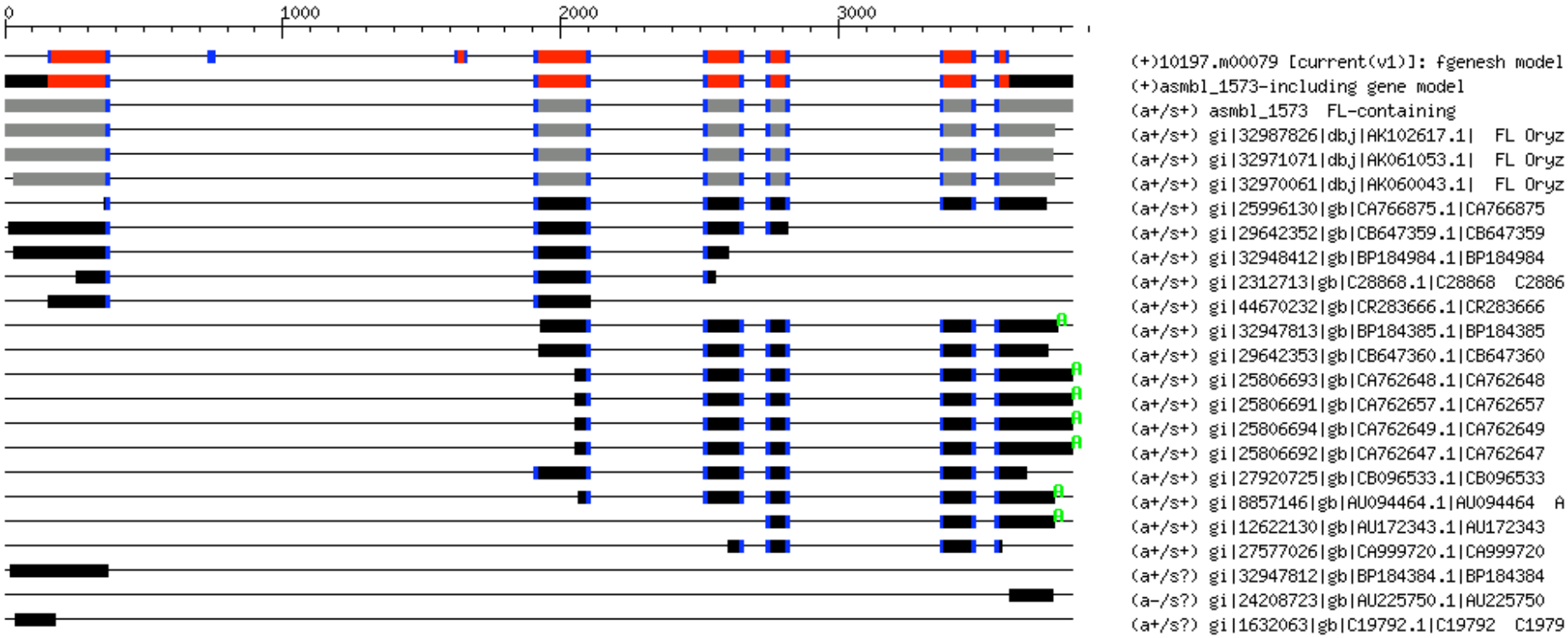


Assembly Report Page

Report for cDNA subcluster: 1258

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

Subcluster view.

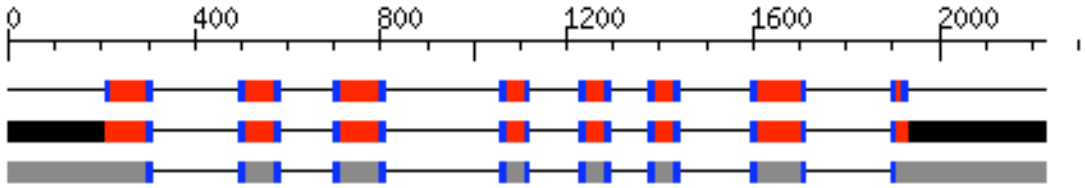


Assembly description

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

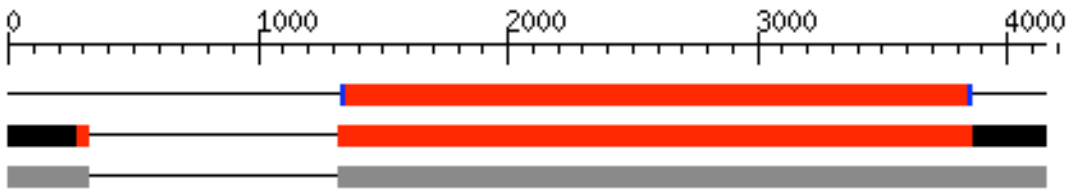
Examples of Classified Updates

FL adds/extends UTRs



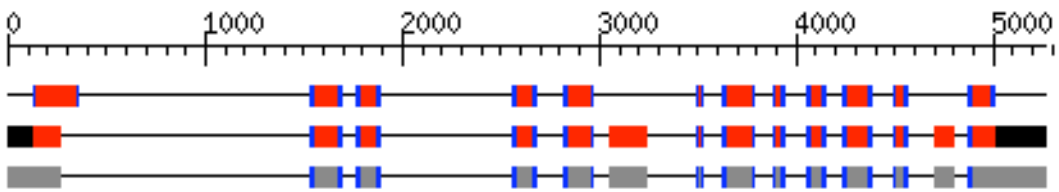
(+)10000.m00089 Before Update fgenesh model 10000.
(+)After Update
(a+/s+) asmbl_1 FL-containing

FL extends protein



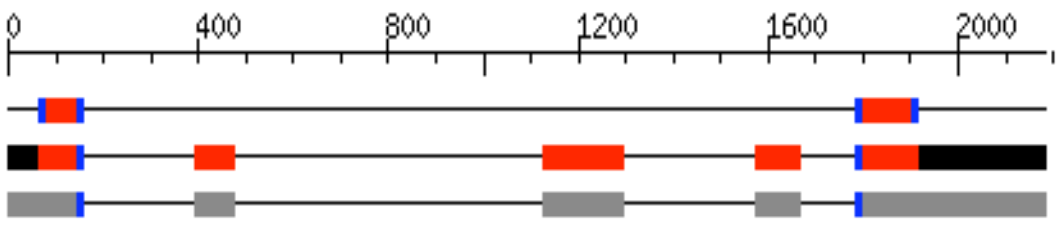
(+)10021.m00075 Before Update fgenesh model 10021.
(+)After Update
(a+/s+) asmbl_164 FL-containing

FL updates structure (passes homology test)



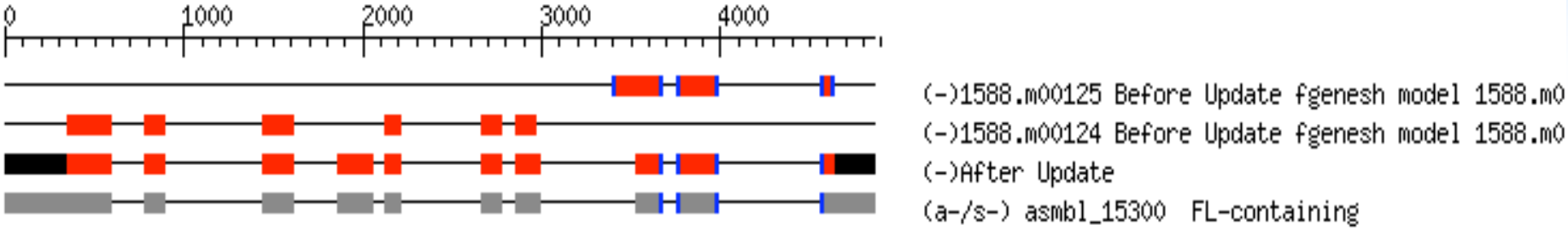
(+)10039.m00085 Before Update fgenesh model 10039.
(+)After Update
(a+/s+) asmbl_282 FL-containing

FL updates structure (fails homology, passes ORF span)

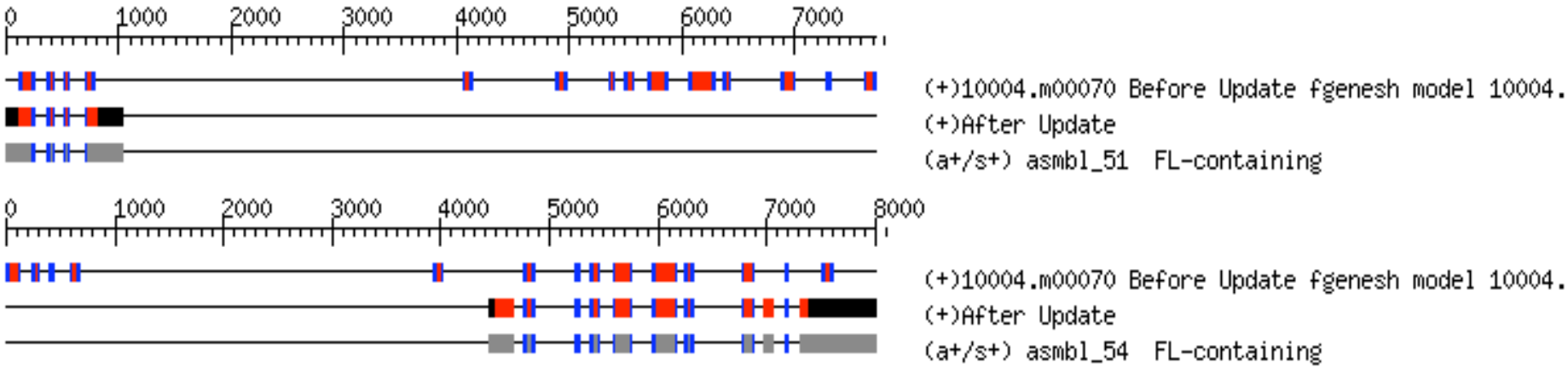


(+)10177.m00095 Before Update fgenesh model 10177.
(+)After Update
(a+/s+) asmbl_1240 FL-containing

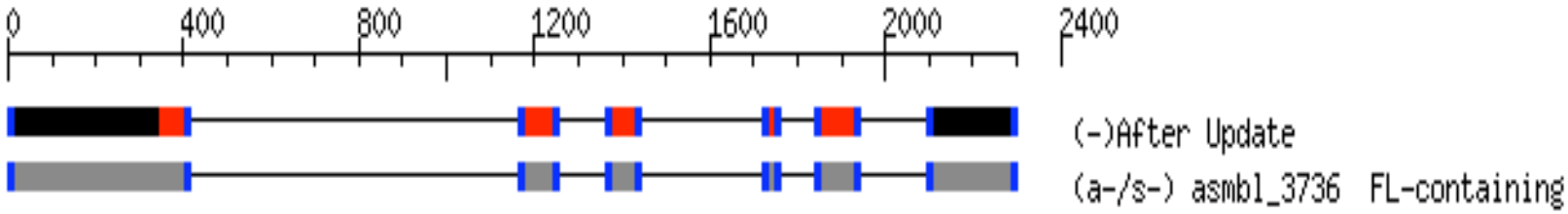
FL merges genes



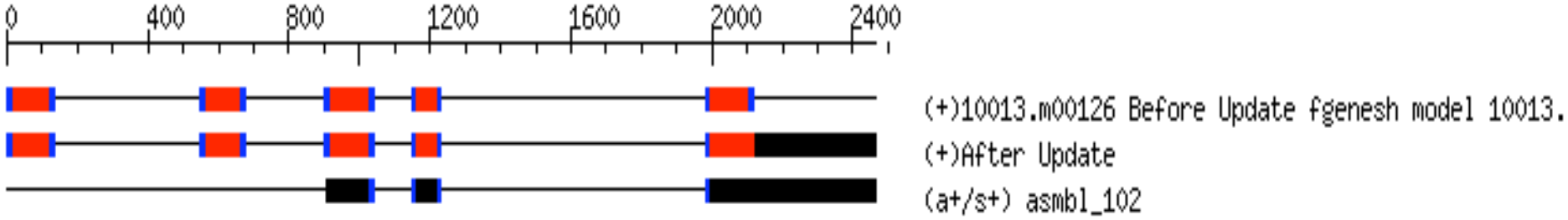
FL split gene



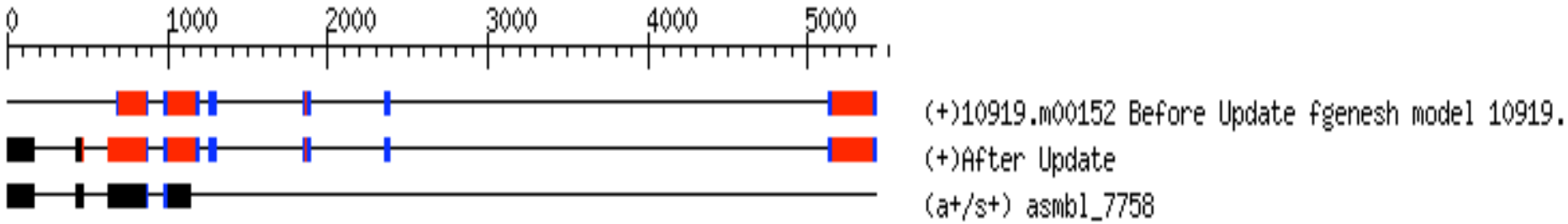
FL novel gene



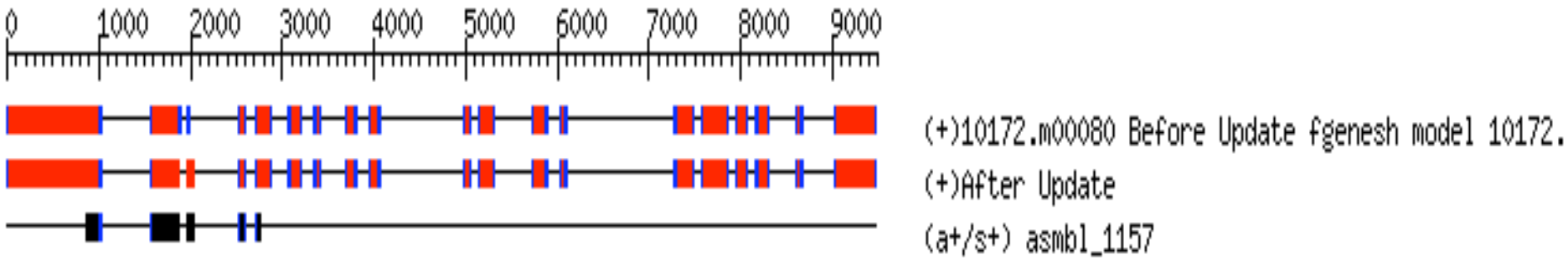
EST extends UTRs



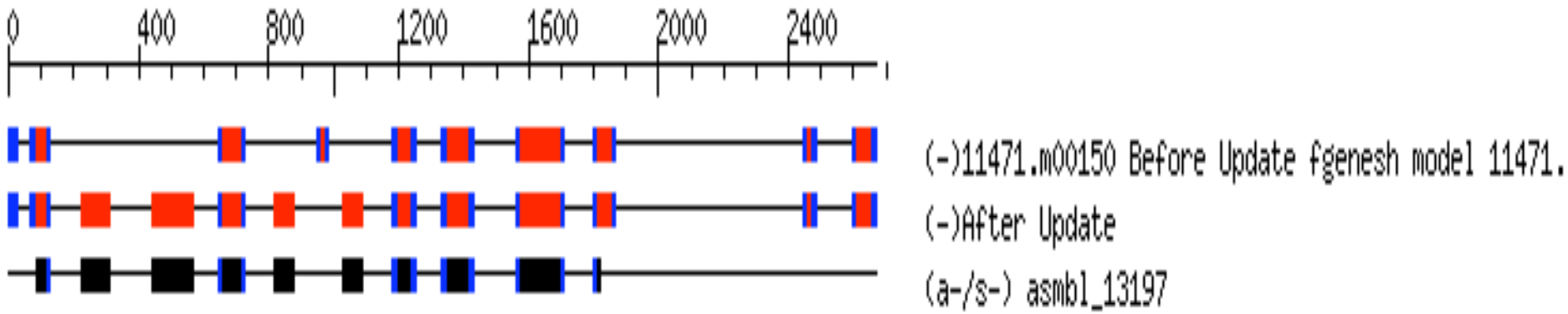
EST extends protein



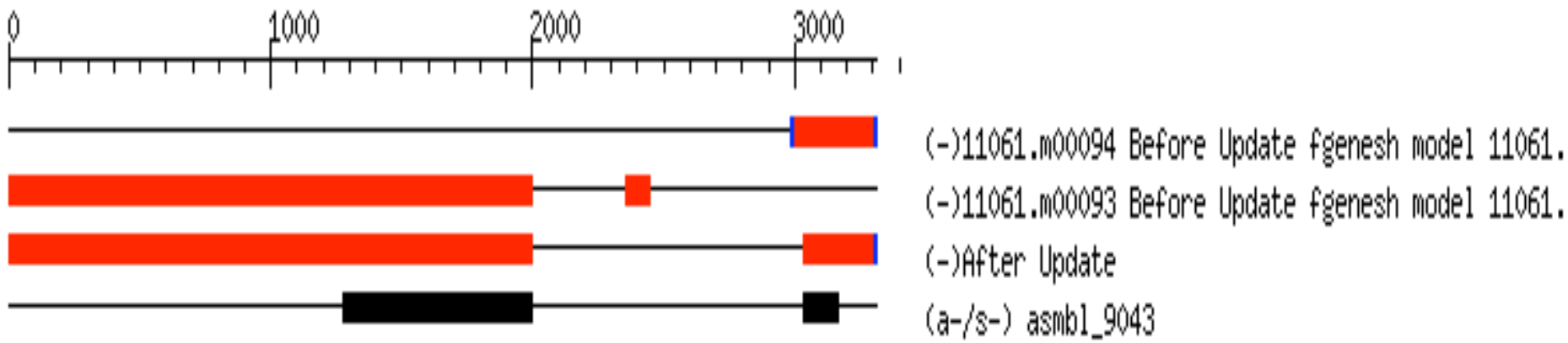
EST updates structure (passes homology test)



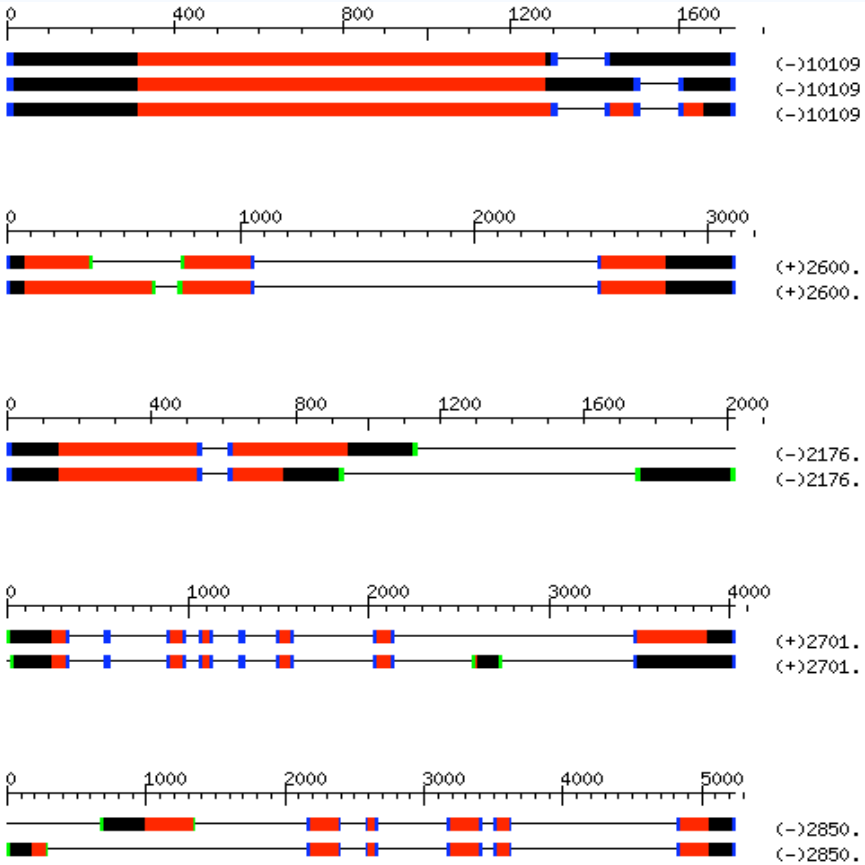
EST updates structure (fails homology test, passes ORF span)



EST merges multiple genes



A tool for Studying Alternative Splicing
Evidence for >5000 genes alternatively spliced

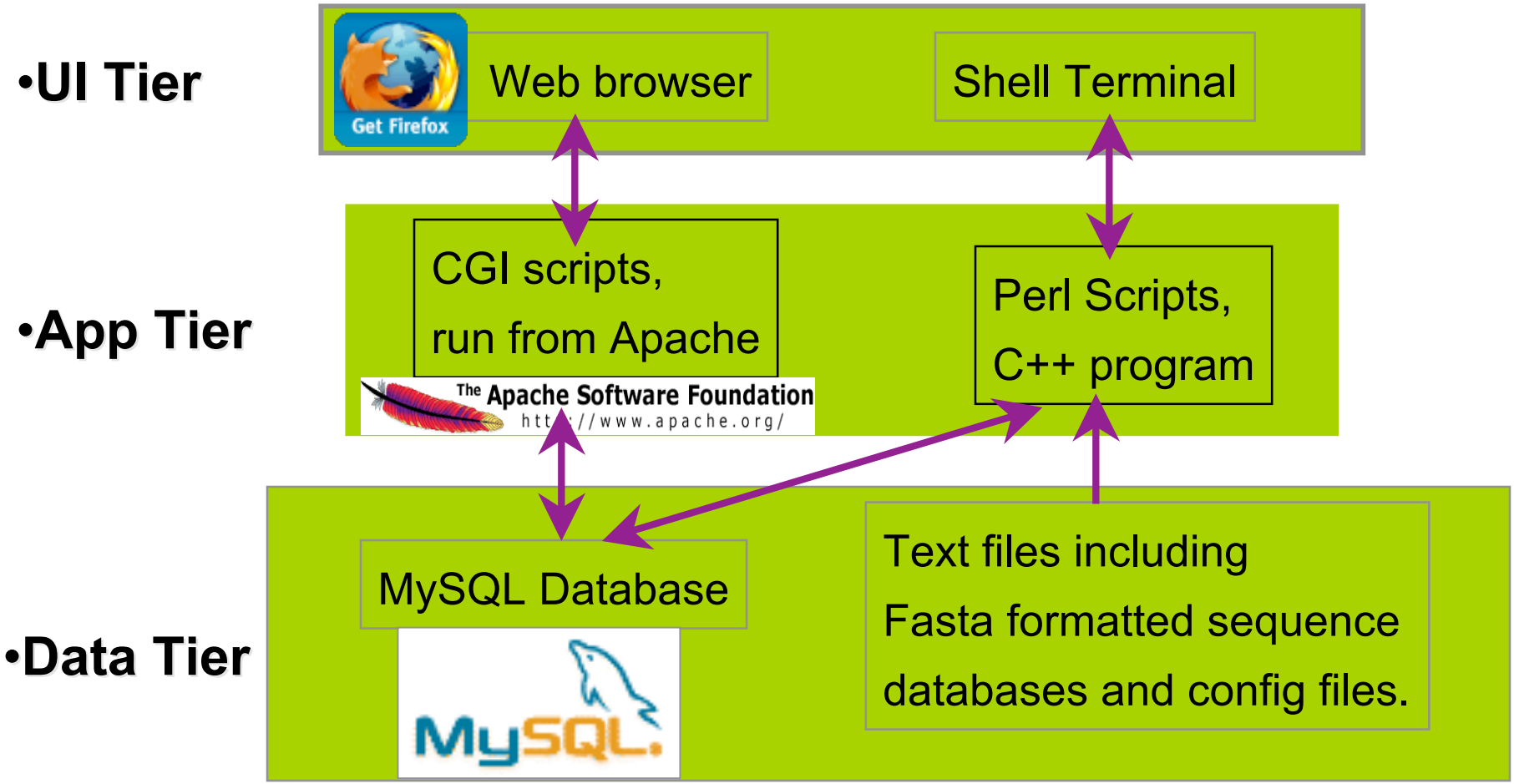


- Unspliced Introns: 45%
- Alt donor/acceptor: 32%
- Start/end in intron: 34%
- Exon skipping: 8.4%
- Alternate exon: 7.4%

*categories overlap due to combinations

Distribution of splicing variations is similar to those described in Arabidopsis.

PASA Pipeline Application Framework



PASA Documentation

http://pasa.sf.net

Gene Structure Annotation and Analysis Using PASA

Brian Haas

<bhaas@tigr.org>

PASA, acronym for Program to Assemble Spliced Alignments, is a Eukaryotic genome annotation tool that exploits spliced alignments of expressed transcript sequences to automatically model gene structures, and to maintain gene structure annotation consistent with the most recently available experimental sequence data. PASA also identifies and classifies all splicing variations supported by the transcript alignments.

Table of Contents

TIGR THE INSTITUTE FOR GENOMIC RESEARCH

Home > Software

Home | What's New | Search | About TIGR | Careers | Site Map | Contact Us | FAQ

- Genome Databases
  - Comprehensive Microbial Resource
  - Unfinished Genomes
  - Eukaryotic Projects
  - Gene Indices
  - more >>
- Functional Genomics
  - PFGRC, PGA, AT Arrays, Cancer Arrays
  - more >>
- Microbial Sequencing Center
- Tree of Life
- Software
  - Glimmer
  - MUMmer
  - more >>
- Conferences
  - GSAC 16
  - Computational Genomics
  - Genomes 2004
  - more >>
- Scientific Publications
- Faculty
- Education & Training
- Genome News Network
- Related Links
- Privacy Statement

Software Tools

TIGR has many software systems available for free download. All of them are OSI Certified Open Source Software. Click on the links below to get more information about the various packages and to download the source code.

OSI certified

Gene Finding/Annotation | Alignment | Sequencing/Finishing
Microarray | Grid Computing | Other

Gene Finding/Annotation

MANATEE

MANATEE is a web-based gene evaluation and genome annotation tool. Manatee can store and view annotation for prokaryotic and eukaryotic genomes. The Manatee interface allows biologists to quickly identify genes and make high quality functional assignments, such as GO classifications, using search data, paralogous families, and annotation suggestions generated from automated analysis.

PIRATE (Prediction Informatics Resources at TIGR & Elsewhere) is a central repository of open-source bioinformatics prediction programs and reusable software components, documentation, training data, experimental results, tips and tricks, and external links. Updated often.

PASA

PASA The PASA pipeline: Includes PASA (Program to Assemble Spliced Alignments) as well as the pipeline to generate transcript alignments, compare alignment assemblies to existing gene model annotations, update gene structure annotations based on transcript alignments, and automatically model new genes based on full-length cDNA containing alignment assemblies. This system as well as its original application is described in: Haas, et al. Nucleic Acids Res. 2003 Oct 1;31(19):5654-66

Obtaining PASA

http://www.tigr.org/software

QUESTIONS?