Outreach Workshop for the USDA-NRI Wheat 3AS Chromosome Physical Map and BAC Sequencing Project

Rama Maiti

August 27, 2008

Funding: USDA-NRI

J. Craig Venter INSTITUTE

USDA

Sequence Annotation: Why?

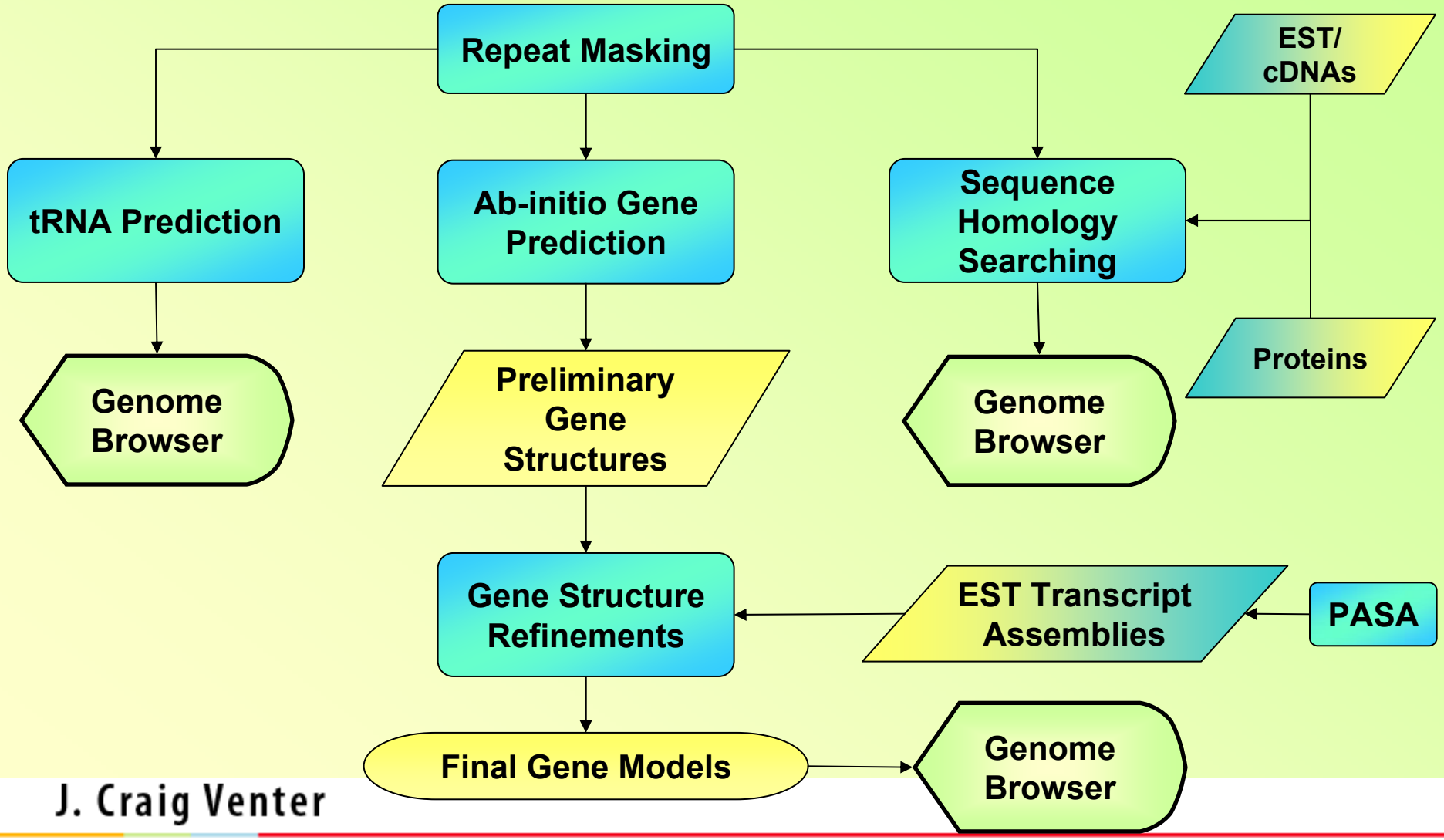
- Raw sequence data is not meaningful to biologists.
- To be meaningful/useful it has to be converted into biological significant knowledge : markers, genes, RNAs, protein sequences.
- Genome annotation is the first step toward this knowledge acquisition.

Annotation Process

The annotation mainly consists of these steps:

- **Structural annotation:** defining the structure of each transcript and structural element.
- **Functional annotation:** classifying each element, attributing, when possible, a specific function.
- **Display of the Annotation Data**: The data are uploaded to a genome browser for viewing and accessing the data.

Gene Finding: Flowchart



Repeat Masking

- Fragments of transposons are often confused for protein-coding exons of genes.
- Identification of repeats in the genome and masking the repeats from the sequence becomes necessary for good quality annotation.
- Prior to gene finding efforts, repetitive elements were identified and masked by Repeatmasker using JCVI plant repeat database.

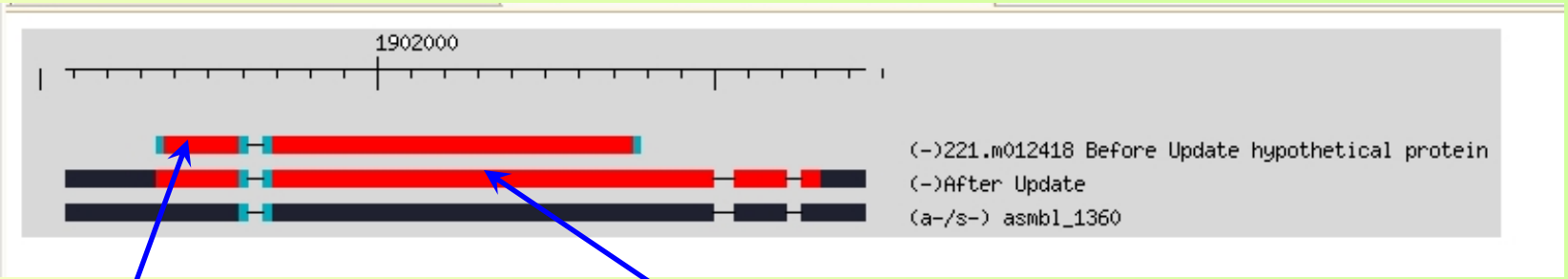
Homology Evidence

- AAT gapped alignments (Huang, 1997)
  - In house Non redundant protein database
  - Rice protein database
  - In house Plant Transcript Assemblies (TA)
- EST: PASA Spliced Transcript Assemblies (PASA -http://sourceforge.net/projects/pasa)
  - Wheat ESTs/cDNAs are aligned to the genome and assembled.

Ab-initio Gene Predictions

- Different gene prediction software uses different computational algorithms. It is important to use the most suitable ones.
- Following gene finders have been optimized and used for wheat BAC annotation in JCVI.
  - FGENESH (monocot)
  - Genscan (Maize)
  - Genscan+ (Arabidopsis)
  - GlimmerHMM (rice)
- FGENESH predictions are modeled as gene models.
- Other predictions are viewed as the evidence in the genome browser.

Gene Model Refinement with Experimental Evidence

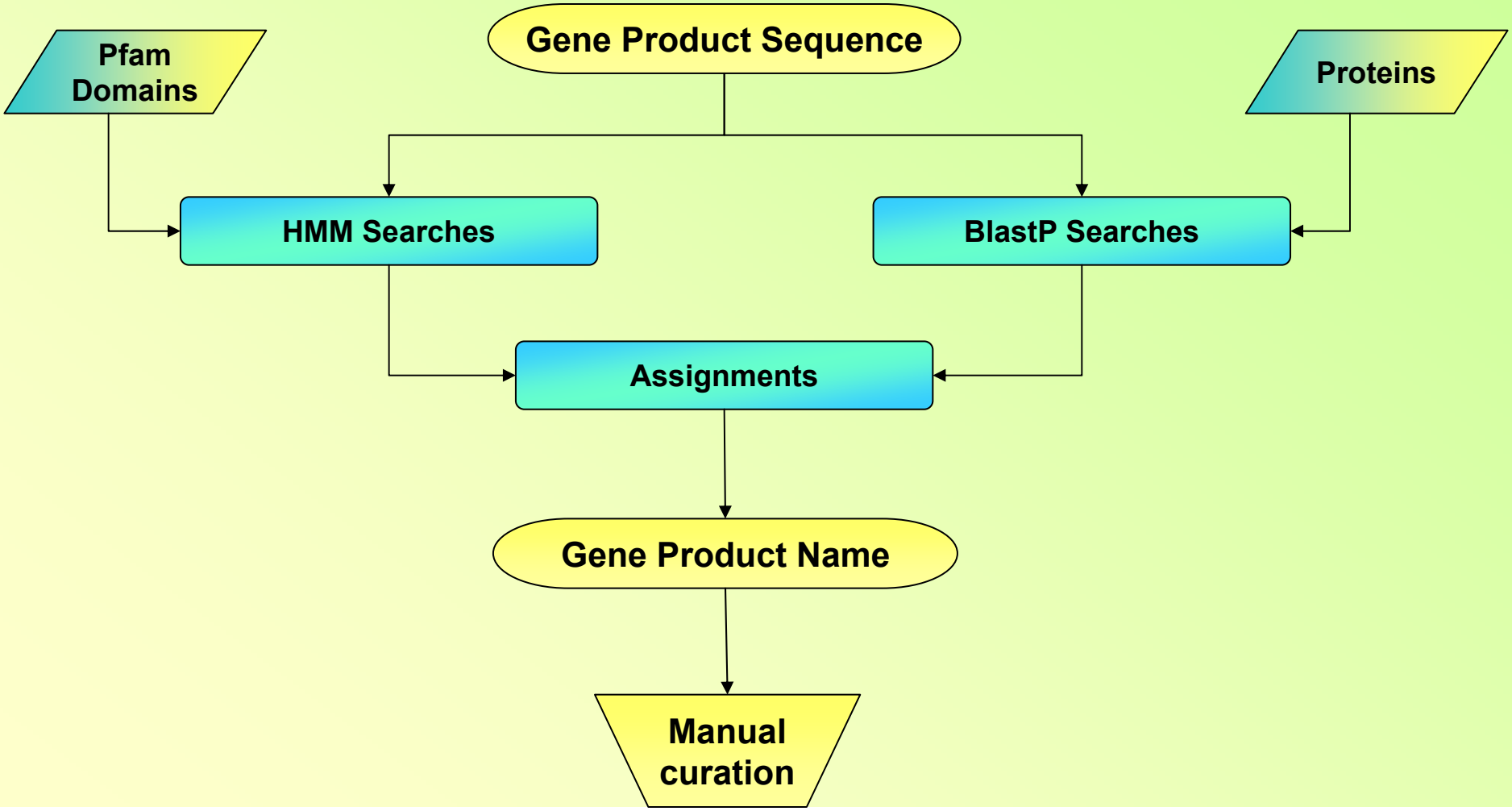


Preliminary Gene Structure

Refined Gene Structure

EST assemblies are best evidence for improving the gene structure with correct intron/exon boundaries.

Functional Annotation Flowchart



J. Craig Venter INSTITUTE

Functional Annotation: Naming the Gene Product

Gene products are named based on the following nomenclature.

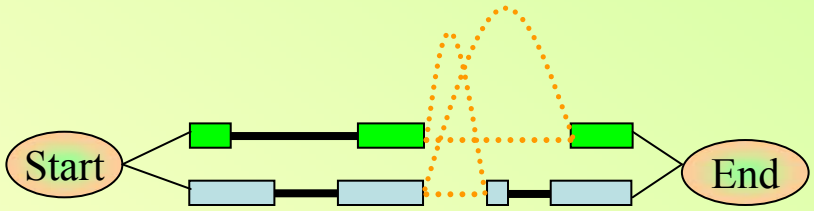
- Known, Putative: Identical or strong similarity to documented gene(s) in Genbank or has high similarity to a Pfam domain
- Domain Protein: Has high similarity to a Pfam domain
- Expressed protein: Only match is to an EST with an unknown function
- Hypothetical: Predicted solely by gene prediction programs

Display of Annotation Resources

- Wheat genomic or annotated data is accessible through JCVI wheat web site. (http://wheat.tigr.org)

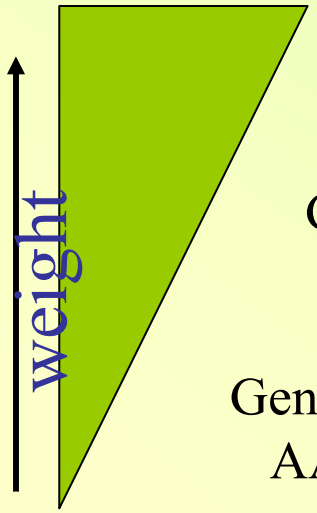
- The BAC sequence and the annotation features are displayed in the context of its supporting evidence through a genome browser.

Improvement in Annotation Pipeline: Combining the Evidences



EVidenceModeler (EVM)

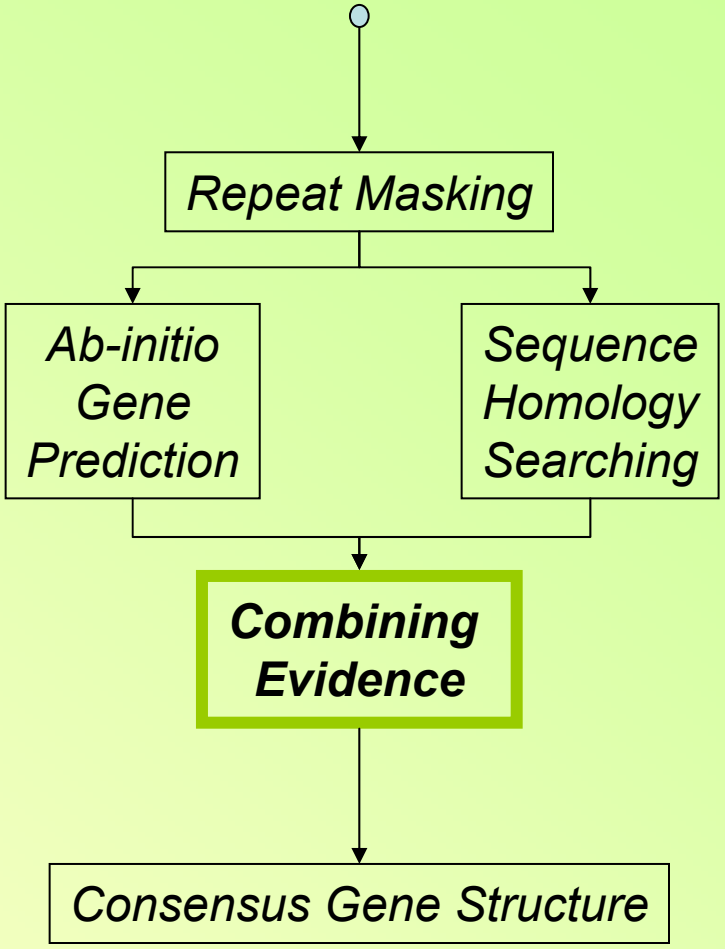
Combines predicted exons and alignments into weighted consensus gene structures



PASA transcript alignment assemblies

Genewise protein alignments

Gene Predictions, AAT alignments



Acknowledgment

Project Director:

- Bikram S. Gill, Kansas State University, USA

Co-Project Directors:

- Katrien M. Devos, University of Georgia
- Justin D. Faris, USDA–ARS
- Piotr Gornicki, University of Chicago
- Wanlong Li, Kansas State University
- Ming-Cheng Luo, University of California Davis, USA
- Pablo Rabinowicz, Institute for Genome Sciences, U. of Maryland USA
- Agnes Chan, J. Craig Venter Institute, USA

Collaborators:

- Rudi Appels, Murdoch University, Australia
- Jaroslav Dolezel, Institute of Botany, Czech Republic
- Jan Dvorak, University of California, USA
- Catherine Feuillet, INRA-UBP, Clermont-Ferrand, France
- C. Robin Buell, Michigan State University

- **JCVI Annotation Group**