Introduction to Genome Annotation: Overview of What You Will Learn This Week

C. Robin Buell

May 21, 2007

Types of Annotation

Structural Annotation: Defining genes, boundaries, sequence motifs

e.g. ORF, exon, intron, promoter

Functional Annotation: what a gene or sequence does

e.g. Rubisco, expressed in leaves, functions in photosynthesis, found in chloroplasts

Comparative Genomics: Similarities/differences between organisms

e.g., collinearity (synteny) of gene order between human, rat, mouse

How to do annotation?

Automated: computationally generated via algorithms, highly dependent on transitive events

Manual: a trained human “annotator” views the annotation data types and interpets the data

Due to the VOLUMES of genome data today, most genome projects are annotated primarily using automated methods with limited manual annotation

Pros and Cons?

Automated: Fast, cheap, accuracy can be compromised, rapidly updated

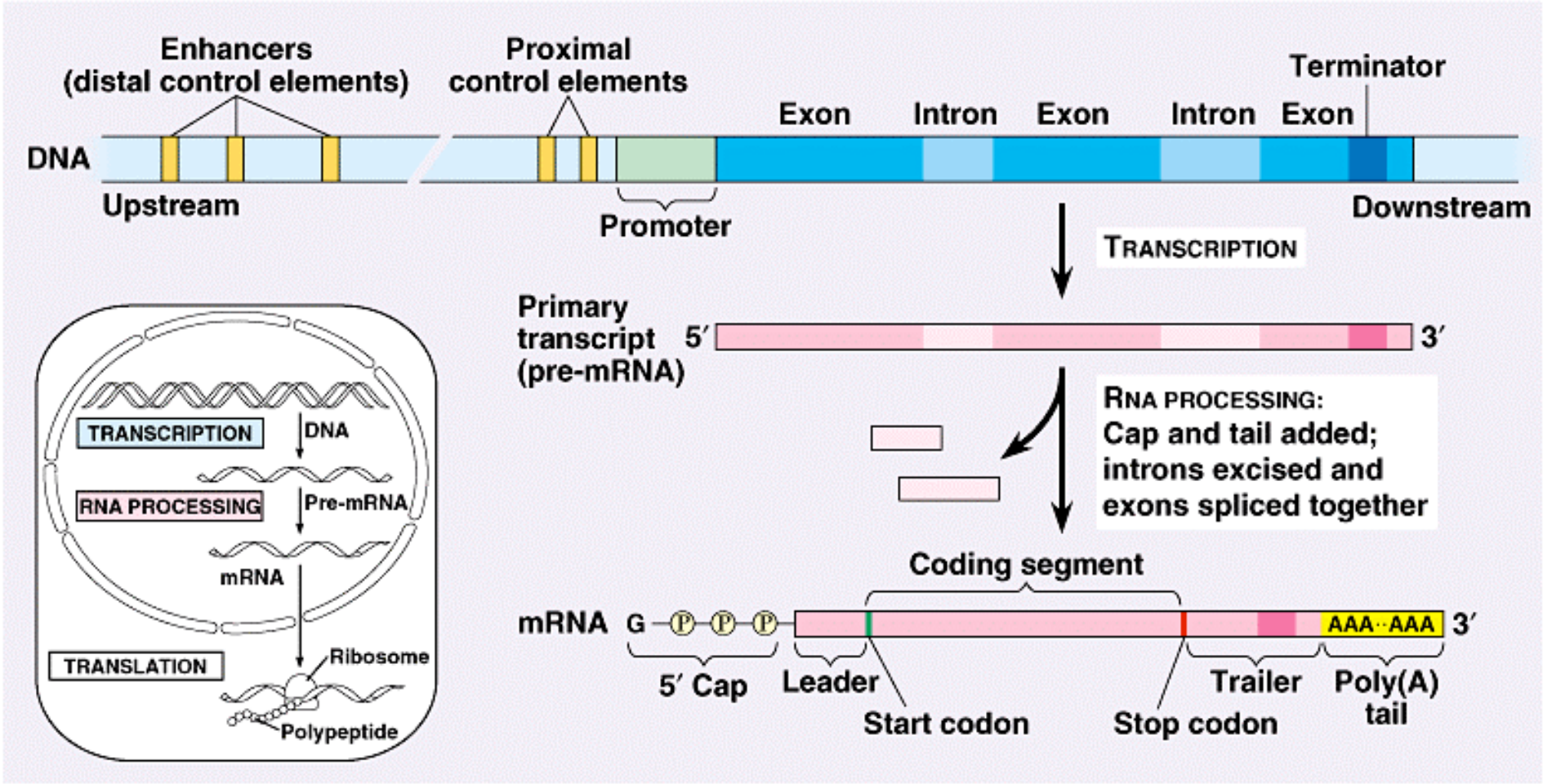
Manual: slow, expensive, accurate, slowly updated

$$$$$ and interest determine the blend of automated versus manual annotation for a genome

Eukaryotic Genome Structure

Gene | Intergenic Region | Gene | Intergenic Region | Gene

Eukaryotic Gene Structure and Transcript Processing



Structural Annotation: Finding the Genes in Genomic DNA

Two main types of data used in defining gene structure:

Prediction based: algorithms designed to find genes/gene structures based on nucleotide sequence and composition

Sequence similarity (DNA and protein): alignment to mRNA sequences (ESTs) and proteins from the same species or related species; identification of domains and motifs

Eukaryotic Gene Finding

Identifying the protein coding region of eukaryotic genes

AAAGCATGCATTTAACGAGTGCATCAGGACTCCATACGTAATGCCG

Gene finder (many different programs)

Wei Zhu will give a talk on gene finders on Tuesday

AAAGC ATG CAT TTA ACG A GT GCATC AG GA CTC CAT ACG TAA TGCCG

TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

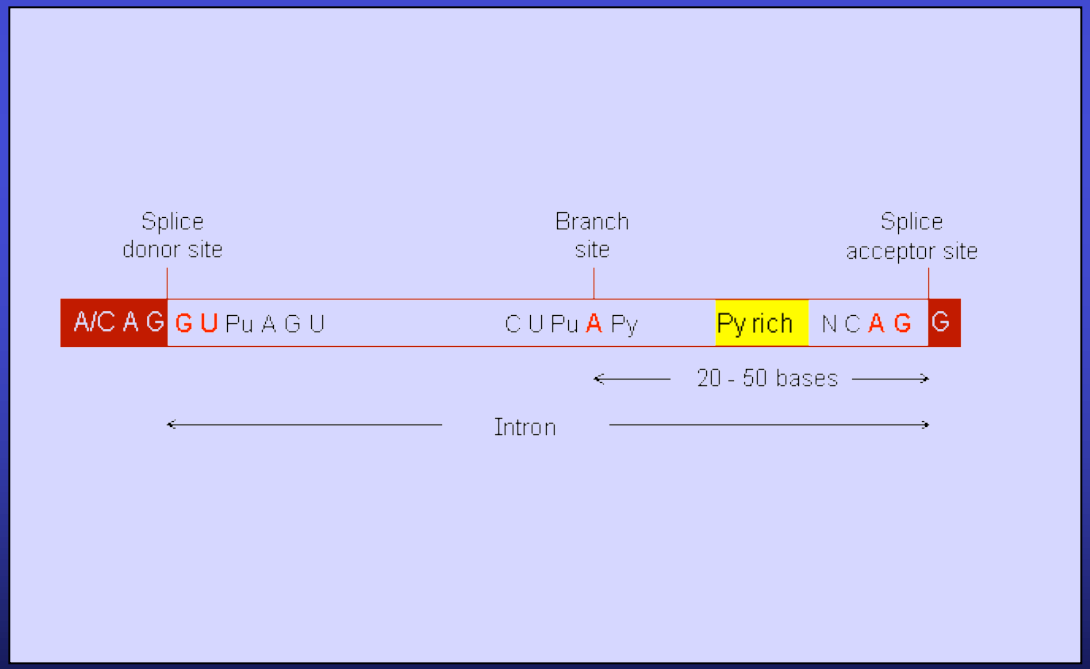
Types of (Coding) Exons

- Single exon gene
  - Start codon to stop codon
- Initial exon
  - Start codon to donor site
- Internal exon
  - Acceptor site to donor site
- Terminal exon
  - Acceptor site to stop codon

***Initial and terminal exon most difficult to identify

Signals Within DNA

- Splice sites to identify intron/exon junctions
- Transcription start and stop codons
- Promoter regions
- PolyA signals



Experimental Evidence

DNA sequence evidence:

Transcript sequence (EST, full length cDNA, other expression types); more restrictive in evolutionary terms

Protein Evidence: alignment to protein that suggests structural similarity at the amino acid level; can be more distant evolutionarily

Experimental Evidence

Transcript evidence:

-Demonstrates gene is transcribed

-Delineates exon boundaries

-Defines splice sites and alternative transcripts

-If EST based, indicates expression patterns

Experimental Evidence

Protein evidence via alignment to non-redundant amino acid databases:

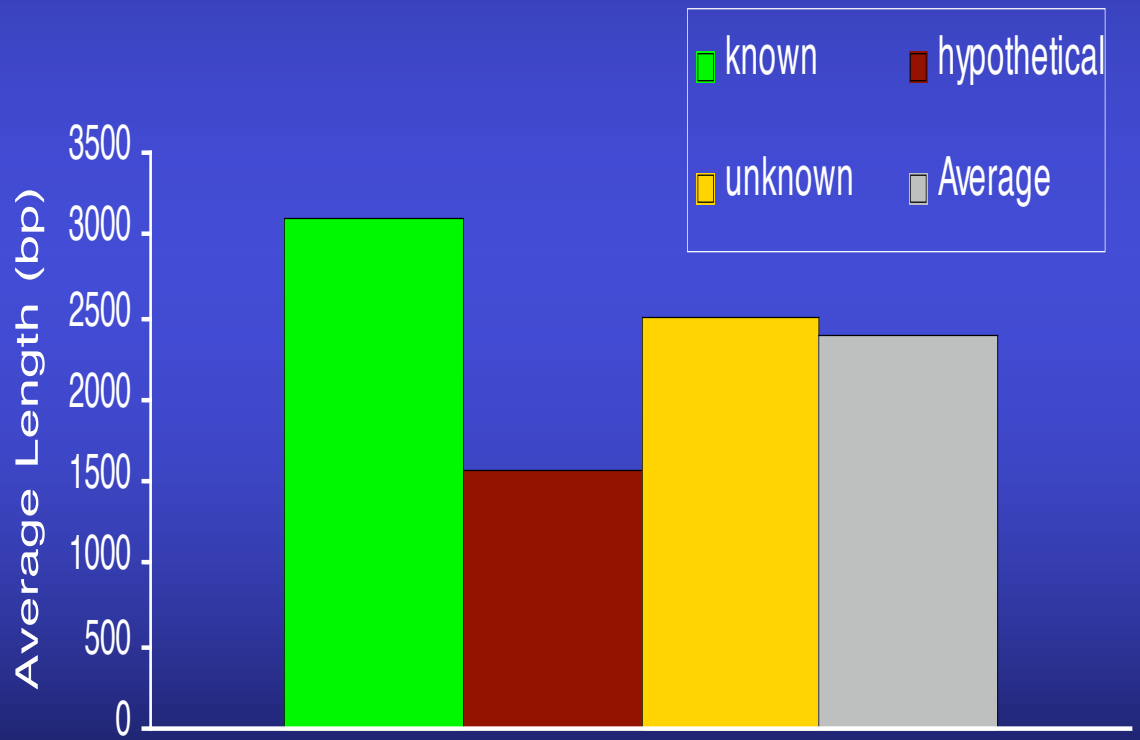
-Shows conservation across species and within species of proteins (orthologs and paralogs)

-Delineates exon structure to some extent

-Provides functional annotation as to what gene encodes

Experimental Evidence

Both protein and transcript evidence are important as they allow for extension of the gene model



Experimental Evidence

Protein Evidence: Domain and Motifs

Domains: conserved amino acid sequence that confers a function (Pfam); determined differently from simple alignment methods

Motifs: amino acid sequence with known function (signal peptide, transmembrane domain)

Eukaryotic Automated Annotation is Not a Solved Problem

What you are getting is output from a series of prediction tools or alignment programs

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

Features Typically Resolved During Manual Annotation

- incorrect exon boundaries
- merged, split, missing genes
- missing untranslated regions (UTRs)
- missing alternative splicing isoform annotations
- degenerate transposons annotated as protein-coding genes

Manual Annotation of Gene Models

Use of a graphic viewer facilitates interpreting the myriad of computational and experimental evidence.

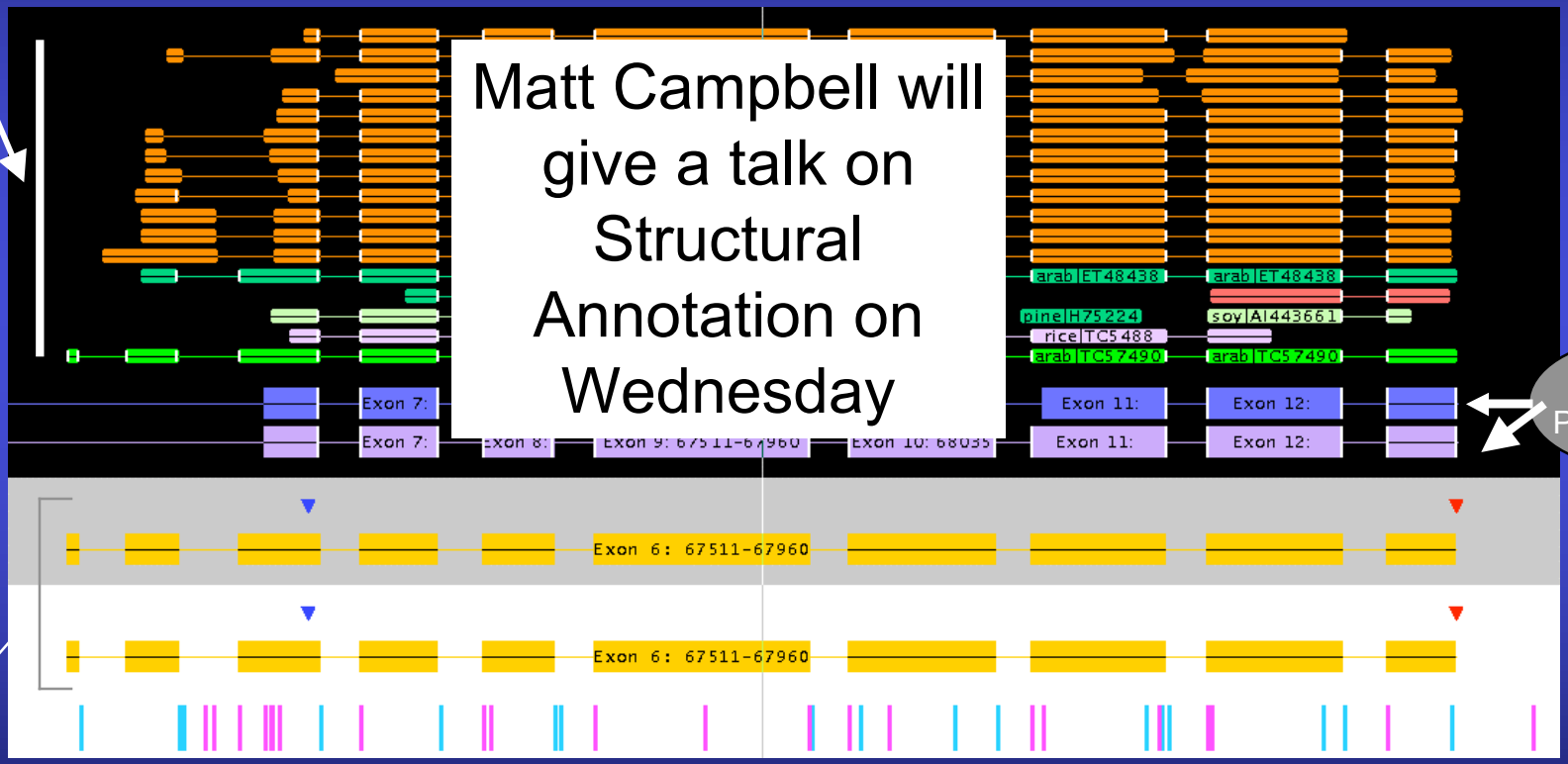
At TIGR, we use Annotation Station as a tool to:

- Examine the results of the automated pipeline
- Edit the predicted structures of genes
- Identify new genes

Structural Annotation: Graphic Viewer Annotation Station

Sequence Database Hits
Top: Protein matches
Bottom: EST matches

Not shown graphically: gene name, nucleotide and protein sequence, MW, pI, organellar targeting sequence, membrane spanning regions, other domains.



Gene Predictions

Annotated Gene
Top: editing panel
Bottom: final curation

Splice site predictions:
red: acceptor sites
blue: donor sites

Screenshot of a component within Neomorphic's annotation station: www.neomorphic.com

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Functional Annotation

- Gene product names
- Gene symbols
- EC numbers
- Expression pattern
- Gene Ontology assignments
- Other fundamental features of gene relevant to organism

There will be a talk on functional annotation on Wednesday

Functional Annotation: Gene Product Names

Gene Name Assignment: Based on similarity to known proteins in nraa database

Categories:

Known or Putative: Identical or strong similarity to documented gene(s) in Genbank or has high similarity to a Pfam domain; e.g. kinase, Rubisco

Expressed Protein: Only match is to an EST with an unknown function; thus have confirmation that the gene is expressed but still do not know what the gene does

Hypothetical Protein: Predicted solely by gene prediction programs; no database match except possibly other hypotheticals (some annotation centers would call this a conserved hypothetical protein)

Functional Annotation: EC number

EC number: Enzyme Commission; standardized nomenclature for enzymes; e.g. EC 5.3.3.2 (isopentenyl-diphosphate isomerase)

Annotator or automated annotation can add EC number and gene symbol based on sequence similarity or publication reports of gene function

Functional Annotation: Expression Pattern

Expression patterns:

-Used to deduce function based on correlative evidence

-Obtained from EST frequency, microarrays, MPSS, etc

-Used to identify regulatory motifs of co-regulated genes

-Limited in application to date; will be greatly expanding in the future

Shu Ouyang will talk further on expression data on Thursday

The Three Ontologies

Gene Ontology: Unified ontologies that categorize genes into 3 categories (Molecular function, biological process and cellular component). Will allow for querying across genomes

- Molecular Function
  - *what the gene product does*
  - *think 'activity'*
- Biological Process
  - *a biological objective*
  - *must have more than one distinct step*
- Cellular Component
  - *location in the cell (or smaller unit)*
  - *or part of a complex*

There will be a talk on gene ontologies on Wednesday

Other Functional Annotation Data

-Genetic marker

-Mutation data; is a mutant available, what is its phenotype

-Tagged lines available (TDNA, Tos17)

-Position on chromosome (locus number/name)

-Ortholog/paralog information

So how many genes are there and how fast can I annotate them?

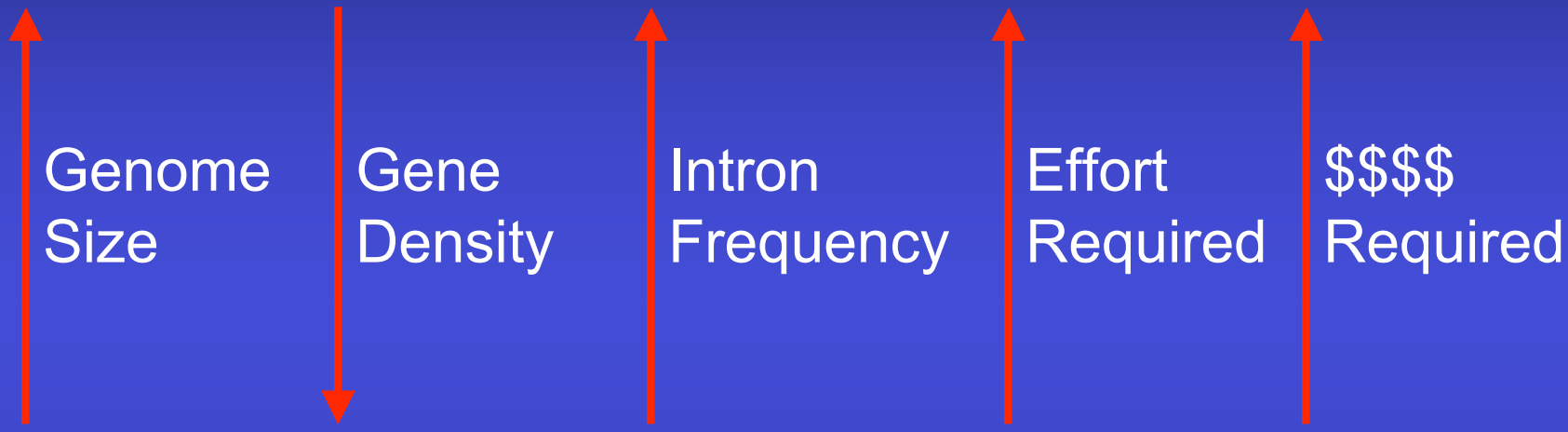
Mycoplasma pulmonis	780	964,000
Saccharomyces cerevisiae	6,300	12,100,000
Plasmodium falciparum	5,400	22,850,000
Caenorhabditis elegans	19,000	97,000,000
Drosophila melanogaster	16,000	120,000,000
Arabidopsis thaliana	25,000	115,400,000
Fugu rubripes	35,000	365,000,000
Oryza sativa	45,000 (60,000)	430,000,000
Homo sapiens	30,000	2,910,000,000

TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

So how many genes are there and how fast can I annotate them?

General Principal: Increase in genome size corresponds to a a decrease in gene density and the presence of more, larger introns



Annotator Speed:

Prokaryote (no introns): high level of annotation: 15 ORFs per day

Eukaryote (introns): lower level of annotation: 20-30 ORFs per day

Improving the Annotation

-Automated annotation is not bad, but manual annotation is much better

-Problem for manual annotation is time consuming and goes “stale” quickly

-Thus, how does a community update the annotation

Three models:

-Don’t update annotation

-Update through community efforts (highly focused, no mechanism to address whole genome, quality can be variable)

-Official annotation project (requires continual funding, restricted to large communities)

Annotation is Never Final, Even Manual Annotation

In 2000, Arabidopsis when the genome was finished:
25,498 Gene models (manual annotation by sequencing centers)

In 2004, TIGR Re-annotation of the genome:
26,207 Genes
3,786 Pseudogenes (includes the TE-related models)

<u>Major Contributions to the Re-annotation Effort:</u>
Full-length cDNA sequences
Automated & semi-automated update of gene model structure
Better annotation of TE gene models
Annotation via paralogous families