How to Add Content to a Reference Assembly

Use
"**Extrinsic**" (RNA-seq) or
"**Intrinsic**" (gene structure) data

1. Identify a Genomic Region

Pv02: 30,623,533..30,624,876

```
AAGGCCAAGGAGAGGAATTGGGTGTTCAGGAAGTTAACGTTCAAGAGTTGAATAGTTGAAGTGAAGTAAA
AAGCGTTGAATTGGAGAGATGACGGGTATGGTGGCTCAAAGAGTTGAAAGCTTGGCTAGCAGTGGGATGA
AGCATATCCCGAAGGAGTACGTGAGGCCGCAAGAGGAGTTGGACAACATAGGGAACGTCTTCGAGGAGGA
GAAGAAGGAAGGGCCTCAGGTTCCAACCATTGACCTGGCAGAGATAGATTCCCCCTCCGAGGTTGTTCGA
GGGAAGTGTCGGGAGAATCTTAAGAAAGCGGCGGAGGAATGGGGCGTCATGAACTTGGTCAACCATGGCA
TCCCTGAGGACCTCTTGAATCGGCTGCGTAAAGCAGGGGAAACCTTCTTCTCTCTTCCCATTGAGGAGAA
GGAGAACTACGCCAACGACCAAGCCTCTGGGAAGATTCAGGGCTATGGGAGCAAGCTTGCTAACAATGCC
AGTGGCCAATTGGAGTGGGAAGATTACTTCTTCCACCTTGTTTATCCCGAGGAGAAGCGTGACCTCTCCA
TCTGGCCAACCAAACCTTCTGATTATACGTGAGCATTCCTCATCCCTTTTCTTTCTTCTTTCACTCTTTT
TTTATTTATAAAACTTTCTTAATTATCAACAAATTTGACATCTCAGTGAGGCTACAAGCGAATATGCAAG
GCGATTGAGGAAGCTTGCGACGAAGATACTAGAGGCACTTTCTGTTGGATTGGGGTTGGAAGGTGGAAGA
CTAGAGAAGGAAGTTGGTGGAATGGAGGAGCTTTTGCTTCAATTGAAGATAAACTACTACCCAATTTGTC
CCCAGCCAGAATTGGCTCTGGGAGTTGAAGCTCACACGGATATAAGTTCACTCACCTTCCTCCTCCACAA
CATGGTGCCAGGTCTGCAACTTTTCTACGAGGGCAAATGGATCACAGCAAAGTGTGTGCCTAATTCCATT
TTGATGCACATTGGGGACACCATTGAGATCCTGAGTAACGGCAAGTACAAGAGTATTCTCCACAGGGGAT
TGGTGAACAAAGAAAAGGTTCGAATATCATGGGCAGTGTTCTGTGAACCACCCAAGGAGAAGATAATCTT
GCAGCCACTTCCTGAACTTGTGACTGAGAAAGACCCAGCTCGTTTTCCTCCTCGCACTTTTGCTCAACAT
ATTCACCACAAACTTTTCAGGAAGGACGAGGAAAGTCTCCCAAAATGAGTCTGTGTCTCCTCCTTCAATG
CCTTCTCTTCTGCACTTCTTAGTTCTTATGGCTTGTACCAATAAAATGACCATTCATGTGGTCTCCTTCT
CATTCTCATGTTAA
```

5'-UTR
(**U**n**T**ranslated **R**egion)

2. Perform Gene Modeling: Exons and Introns

>P.vulgaris v2.1 | Phvul.002G152700

```
AAGGCCAAGGAGAGGAATTGGGTGTTCAGGAAGTTAACGTTCAAGAGTTGAATAGTTGAAGTGAAGTAAA
AAGCGTTGAATTGGAGAGATGACGGGTATGGTGGCTCAAAGAGTTGAAAGCTTGGCTAGCAGTGGGATGA
AGCATATCCCGAAGGAGTACGTGAGGCCGCAAGAGGAGTTGGACAACATAGGGAACGTCTTCGAGGAGGA
GAAGAAGGAAGGGCCTCAGGTTCCAACCATTGACCTGGCAGAGATAGATTCCCCCTCCGAGGTTGTTCGA
GGGAAGTGTCGGGAGAATCTTAAGAAAGCGGCGGAGGAATGGGGCGTCATGAACTTGGTCAACCATGGCA
TCCCTGAGGACCTCTTGAATCGGCTGCGTAAAGCAGGGGAAACCTTCTTCTCTCTTCCCATTGAGGAGAA
GGAGAACTACGCCAACGACCAAGCCTCTGGGAAGATTCAGGGCTATGGGAGCAAGCTTGCTAACAATGCC
AGTGGCCAATTGGAGTGGGAAGATTACTTCTTCCACCTTGTTTATCCCGAGGAGAAGCGTGACCTCTCCA
TCTGGCCAACCAAACCTTCTGATTATACGTGAGCATTCCTCATCCCTTTTCTTTCTTCTTTCACTCTTTT
TTTATTTATAAAACTTTCTTAATTATCAACAAATTTGACATCTCAGTGAGGCTACAAGCGAATATGCAAG
GCGATTGAGGAAGCTTGCGACGAAGATACTAGAGGCACTTTCTGTTGGATTGGGGTTGGAAGGTGGAAGA
CTAGAGAAGGAAGTTGGTGGAATGGAGGAGCTTTTGCTTCAATTGAAGATAAACTACTACCCAATTTGTC
CCCAGCCAGAATTGGCTCTGGGAGTTGAAGCTCACACGGATATAAGTTCACTCACCTTCCTCCTCCACAA
CATGGTGCCAGGTCTGCAACTTTTCTACGAGGGCAAATGGATCACAGCAAAGTGTGTGCCTAATTCCATT
TTGATGCACATTGGGGACACCATTGAGATCCTGAGTAACGGCAAGTACAAGAGTATTCTCCACAGGGGAT
TGGTGAACAAAGAAAAGGTTCGAATATCATGGGCAGTGTTCTGTGAACCACCCAAGGAGAAGATAATCTT
GCAGCCACTTCCTGAACTTGTGACTGAGAAAGACCCAGCTCGTTTTCCTCCTCGCACTTTTGCTCAACAT
ATTCACCACAAACTTTTCAGGAAGGACGAGGAAAGTCTCCCAAAATGAGTCTGTGTCTCCTCCTTCAATG
CCTTCTCTTCTGCACTTCTTAGTTCTTATGGCTTGTACCAATAAAATGACCATTCATGTGGTCTCCTTCT
CATTCTCATGTTAA
```

Exon

Intron

Exon

3'-UTR
(**U**n**T**ranslated **R**egion)

3. Identify the mRNA Transcript

Start Codon

>P.vulgaris v2.1 | Phvul.002G152700

```
AAGGCCAAGGAGAGGAATTGGGTGTTCAGGAAGTTAACGTTCAAGAGTTGAATAGTTGAAGTGA
AGTAAAAAGCGTTGAATTGGAGAGATGACGGGTATGGTGGCTCAAAGAGTTGAAAGCTTGGCTA
GCAGTGGGATGAAGCATATCCCGAAGGAGTACGTGAGGCCGCAAGAGGAGTTGGACAACATAGG
GAACGTCTTCGAGGAGGAGAAGAAGGAAGGGCCTCAGGTTCCAACCATTGACCTGGCAGAGATA
GATTCCCCCTCCGAGGTTGTTCGAGGGAAGTGTCGGGAGAATCTTAAGAAAGCGGCGGAGGAAT
GGGGCGTCATGAACTTGGTCAACCATGGCATCCCTGAGGACCTCTTGAATCGGCTGCGTAAAGC
AGGGGAAACCTTCTTCTCTCTTCCCATTGAGGAGAAGGAGAACTACGCCAACGACCAAGCCTCT
GGGAAGATTCAGGGCTATGGGAGCAAGCTTGCTAACAATGCCAGTGGCCAATTGGAGTGGGAAG
ATTACTTCTTCCACCTTGTTTATCCCGAGGAGAAGCGTGACCTCTCCATCTGGCCAACCAAACC
TTCTGATTATACTGAGGCTACAAGCGAATATGCAAGGCGATTGAGGAAGCTTGCGACGAAGATA
CTAGAGGCACTTTCTGTTGGATTGGGGTTGGAAGGTGGAAGACTAGAGAAGGAAGTTGGTGGAA
TGGAGGAGCTTTTGCTTCAATTGAAGATAAACTACTACCCAATTTGTCCCCAGCCAGAATTGGC
TCTGGGAGTTGAAGCTCACACGGATATAAGTTCACTCACCTTCCTCCTCCACAACATGGTGCCA
GGTCTGCAACTTTTCTACGAGGGCAAATGGATCACAGCAAAGTGTGTGCCTAATTCCATTTTGA
TGCACATTGGGGACACCATTGAGATCCTGAGTAACGGCAAGTACAAGAGTATTCTCCACAGGGG
ATTGGTGAACAAAGAAAAGGTTCGAATATCATGGGCAGTGTTCTGTGAACCACCCAAGGAGAAG
ATAATCTTGCAGCCACTTCCTGAACTTGTGACTGAGAAAGACCCAGCTCGTTTTCCTCCTCGCA
CTTTTGCTCAACATATTCACCACAAACTTTTCAGGAAGGACGAGGAAAGTCTCCCAAAATGAGT
CTGTGTCTCCTCCTTCAATGCCTTCTCTTCTGCACTTCTTAGTTCTTATGGCTTGTACCAATAA
AATGACCATTCATGTGGTCTCCTTCTCATTCTCATGTTAA
```

StopCodon

4. Define the Protein Sequence

>P.vulgaris v2.1 | Phvul.002G152700

```
MTGMVAQRVESLASSGMKHIPKEYVRPQEELDNIGNVFEEEKKEGPQVPTIDLAEIDSPSEVVR
GKCRENLKKAAEEWGVMNLVNHGIPEDLLNRLRKAGETFFSLPIEEKENYANDQASGKIQGYGS
KLANNASGQLEWEDYFFHLVYPEEKRDLSIWPTKPSDYTEATSEYARRLRKLATKILEALSVGL
GLEGGRLEKEVGGMEELLLQLKINYYPICPQPELALGVEAHTDISSLTFLLHNMVPGLQLFYEG
KWITAKCVPNSILMHIGDTIEILSNGKYKSILHRGLVNKEKVRISWAVFCEPPKEKIILQPLPE
LVTEKDPARFPPRTFAQHIHHKLFRKDEESLPK
```

5. Annotate the Gene

- Anthocyanin Synthase (ANS)

Compare with another genes from other species

6. Describe the Gene Function

- Enzyme in Flavonoid Pathway
  - Leucoanthocyanidins ----------→ Anthocyanidins

ANS

Compare with function of the gene from other species

Genome Annotation

Genome Sequencing

- **Initially, the costliest aspect of sequencing the genome**
  - But
    - **Devoid of content**
- Genome must be **annotated**
  - Annotation definition
    - **Analyzing the raw sequence of a genome and describing relevant genetic and genomic features such as genes, mobile elements, repetitive elements, duplications, and polymorphisms**
  - Annotation costs may eventually exceed the sequencing cost
    - Why??
      - *Continued reanalysis is required to define all the genes and the phenotypes they control*

What Does Annotation Describe???

- Genome duplications
- Genes
- Mobile genetic elements
- Small repeats
- Genetic diversity

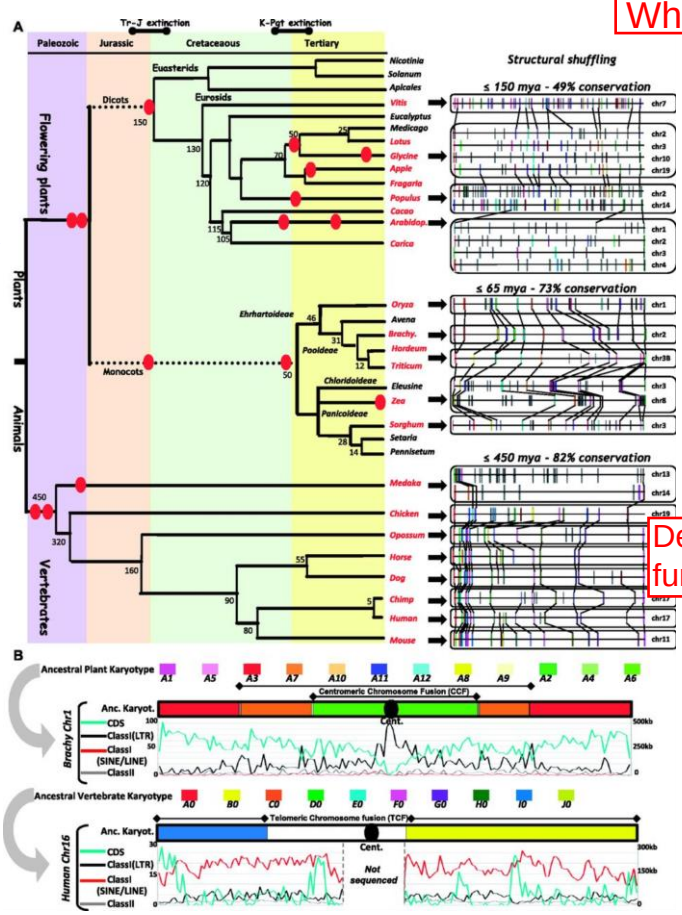
Whole Genome Duplications

- Whole genome duplications are a major event in the history of all eukaryotic groups of species
- **Duplications** can be of
  - *The full genome of one species*
    - **Autopolyploid**
  - *The mating and retention of both chromosome sets of two species*
    - **Allopolyploid**
- Duplications in the biological kingdoms
  - **Animals**
    - **Three ancient whole genome duplications**
  - **Plants**
    - **Multiple duplications along the two major lineages**
      - Dicots and monocots

Plants = shared ancestral duplicatons + lineage specific duplications

Red elipse = Whole Genome Duplication

Animals = shared ancestral duplicatons



Develolp new funciton

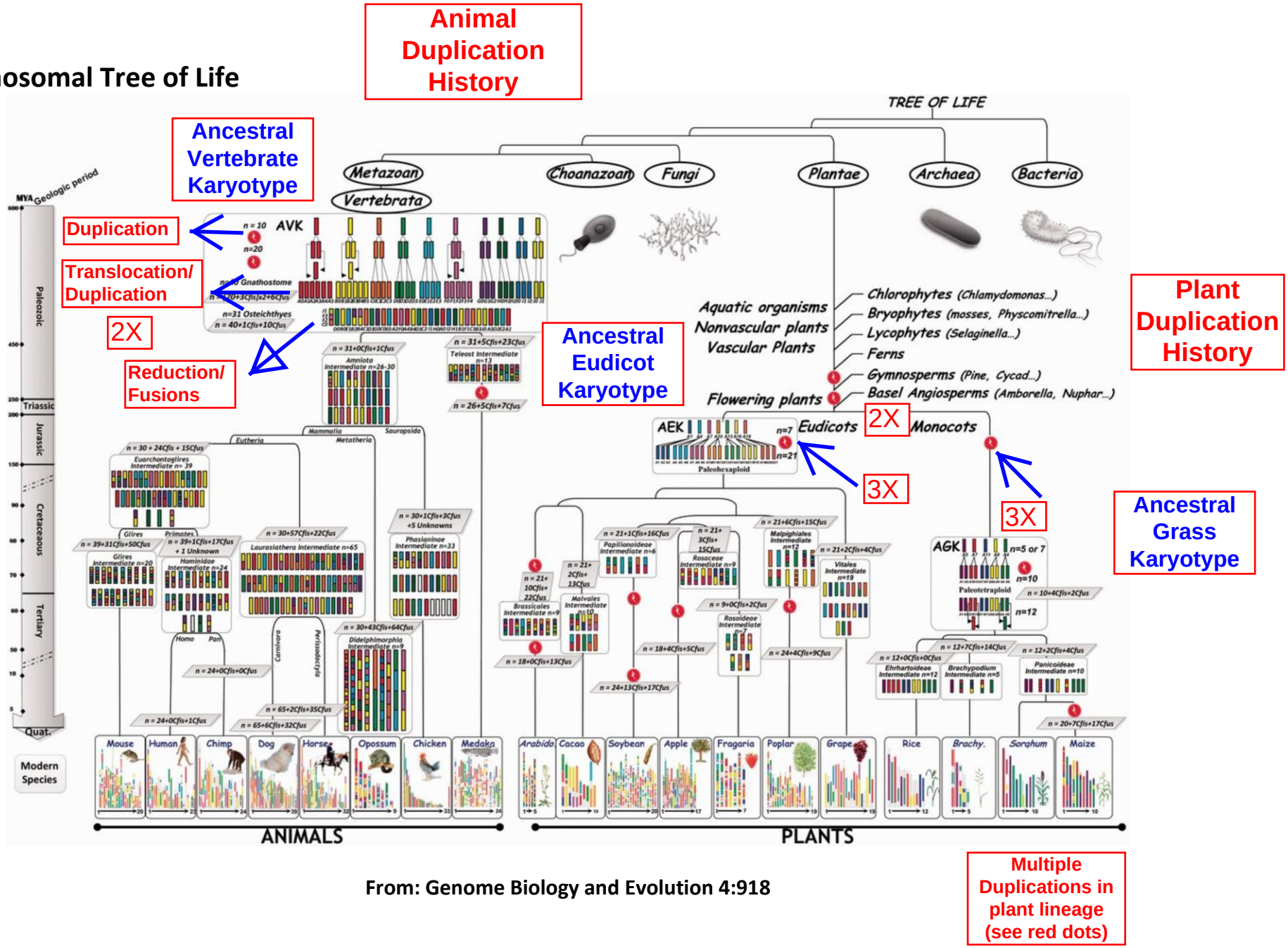
Maintain original funciton

From: Genome Biology and Evolution 4:918

- **Post whole-genome duplication events**
  - **Paleopolyploid**
    - **An ancient species that results from a genome duplication**

- **Duplications define**
  - **Modern relationships between species**
    - Ancestral shared segments exist between distant/close
      - **Derived from an ancient ancestral species that does not exist today**

Chromosomal Tree of Life



From: Genome Biology and Evolution 4:918

Multiple Duplications in plant lineage (see red dots)

- **Major result of duplications**
  - **Many of the genes in all genomes are related by descent**
  - **Protein sequences are similar**
  - **Conserved function can be inferred**
    - **BUT NOT PROVEN FROM SEQUENCE ANALYSIS**

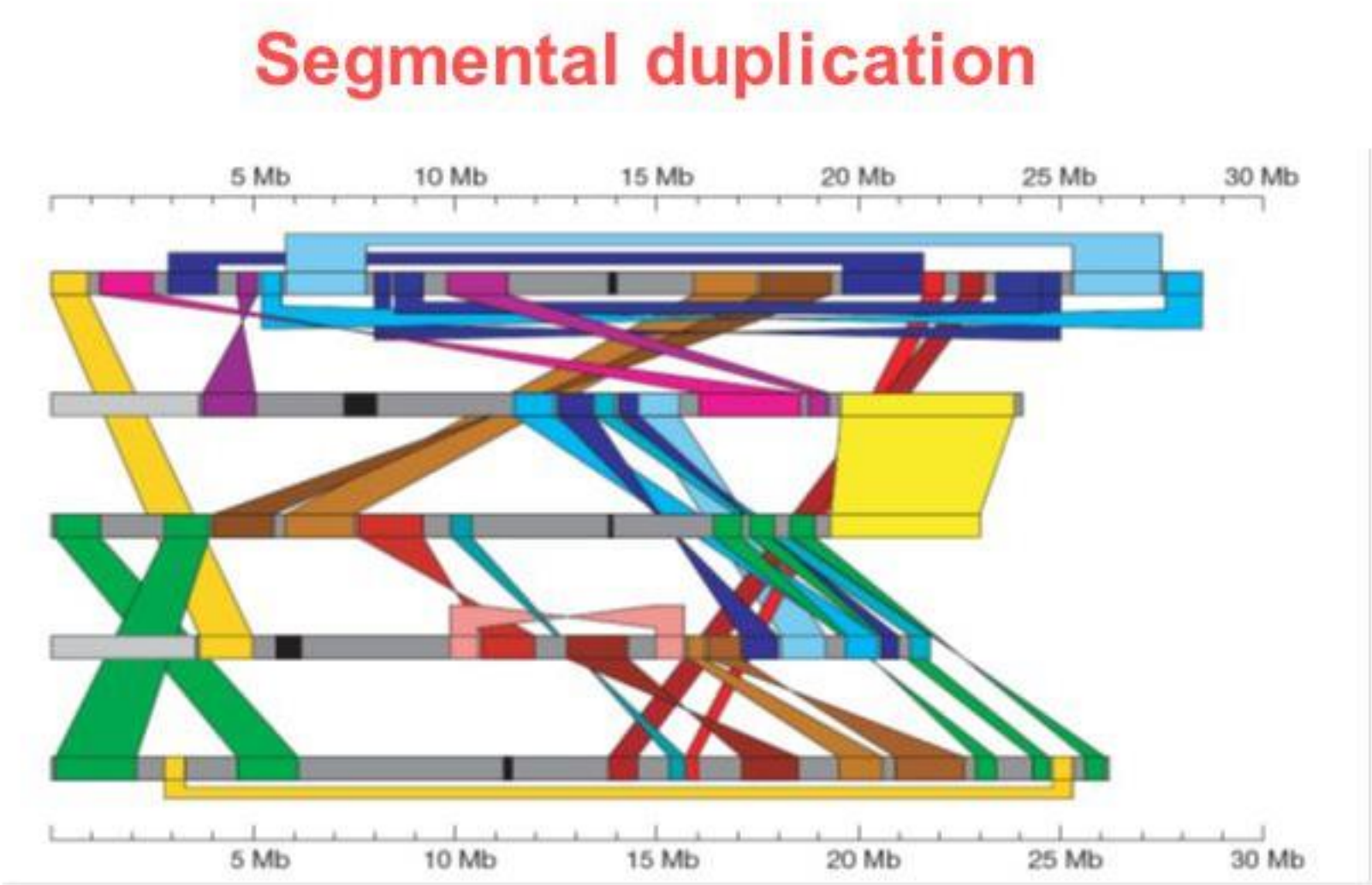
Segmental Gene Duplications

- **What are they???**
  - **Large gene blocks duplicated in the genome**
    - **Intrachromosomal duplication**
      - Duplicated region moved to ***same*** chromosome
    - **Interchromosomal duplication**
      - Duplicated region moved to ***another*** chromosome
      - Confirms/determines biological function

Are their large blocks of duplications in genomes? YES

  - ***Arabidopsis* genome**
    - Model plant organism
      - Originally considered devoid of gene duplications
    - **First Sequencing discovery**
      - Large blocks of segmental duplication, both
        - Local duplication
          - **Intrachromosomal**
        - Distal duplications
          - **Interchromosomal**
  - **Human genome**
    - Duplication pattern similar to *Arabidopsis*
  - **Mouse genome**
    - Lesser degree of duplication

Current Arabidopsis Chromosomes Showing Shared Genomic History



- Fragment between two chromosomes that share a color are similar because of the duplication history of the species lineage.

Genes

- **Collection of Genes define the phenotypic life-cycle of a species**
  - **Development**
  - **Reproduction**
  - **Response to environment**
    - **Biotic**
    - **Abiotic**

Major functions of all biological organisms

- **What defines the gene???**
  - **Coding region**
    - *Contains the information that defines the nature of an expressed protein or a functional RNA molecule*
  - **Controlling regions**
    - *Sequences that define where, when, and how much the gene will be expressed*
- **Major goal of gene annotation**
  - *Defining genes and their controlling regions*

Repetitive Elements

- **Repetitive elements**
  - **Often the major component of genome**
    - Generally conserved
  - **Fairly ease to discover**
    - Example
      - **Retrotransposons**
        - Reverse transcriptase protein is conserved
- **Cataloging the repetitive elements**
  - **The first step of annotation**
    - *Greatly reduces the amount of sequence that must be searched for genes*
    - **Repeat masking the next step**
      - *Procedure that removes the repetitive elements from the gene discovery process*

Mobile Genetic Elements

- Also called transposable elements
  - A major component of some genomes
- **Classes**
  - **Class I elements: Retrotransposons**
    - **Most abundant class of repeats**
    - Abundance
      - **Human** 3 Gb
        - 50% of genome is mobile elements
      - **Arabidopsis** 0.15 Gb
        - 10 % of total DNA
        - 20 % of gene rich-region
  - **Class II elements: DNA elements**
    - McClintock elements

Major Classes of Transposable Elements

Super families of TEs	Number of TEs ($x10^3$)	Coverage of TEs (bp)	Fraction of genome (%)
Class 1 (RNA-based)	**283.1**	**195,948,599**	**41.47**
LTR retrotransposon	**244.3**	**181,963,056**	**38.51**
Ty3-gypsy	**146.7**	**125,312,211**	**26.52**
Ty1-copia	62.2	47,126,880	9.97
Others	35.3	9,523,965	2.02
LINEs	37.8	13,825,275	2.93
SINEs	1.0	160,268	0.03
Class 2 (DNA-based)	**87.4**	**26,832,637**	**5.68**
CACTA	44.0	13,295,207	2.81
Harbinger/PIF	0.5	263,181	0.06
hAT	4.0	1,062,438	0.22
Helitron	18.3	5,095,472	1.08
MULE	20.7	7,116,339	1.51
Unclassified TEs	14.7	2,728,570	0.58
Total	**385.2**	**225,509,806**	**47.73**

Table S8. Summary of transposable elements (TEs) in *Phaseolus vulgaris*. Schmutz et al. (2014).

Other Repeat Elements

- **Simple Sequence Repeats**
  - SSRs
    - Defined as
      - **Localized repetitions of di- or tri-nucleotides**
    - Major repetitive class found in genomes
      - >100,000 in many eukaryotic genomes
      - Widely used as molecular markers

Example: (ATA)n; n=10
ATA repeated ten times

Genetic Diversity

- **How can gene sequences among individuals of a species vary?**
  - **Large deletions**
    - *Eliminate gene function*
  - **Small deletions**
    - *Gene often expressed but phenotype is changed*
      - Example
        - **Cystic fibrosis gene**
          - *Three nucleotides (triplet) lost = phenylalanine deleted in protein*
            - **Mutant CF phenotype expressed**

Single nucleotide polymorphisms or variants (SNPs or SNVs)

- **A difference in a single nucleotide between two alleles**
- Resequencing discovers differences
  - Uncovers SNP diversity related to function
    - Examples of functional SNPs
      - **Sickle cell anemia**
        - *Adenine → thymine* in sixth amino acid codon of **β-globin gene**
        - **Change leads to sickle cell phenotype**
      - **Mendel's plant height gene (Le)**
        - *Guanine → adenine* change at nucleotide 685 of the mRNA
        - First nucleotide of 229th codon
          - Amino acid: Alanine → Threonine
        - Changes function of **gibberellin 3 beta-hydroxylase**
          - **Plant is short rather than tall**

Exon Definition Model

- **Key principle in gene modeling**
  - **Genes consist of exons and introns**
  - *Based on key sequences that define exons, introns, and 5' and 3' region of genes*
    - Defines the sequences that define exons
      - **These sequences are the keys to *computationally* discovering gene models**

5' End of Intron
Donor Site
AG|GURAGU

Y=C or T
R=A or G

3' End of Intron
Acceptor Site
$Y_{(10)}$NCAG|G

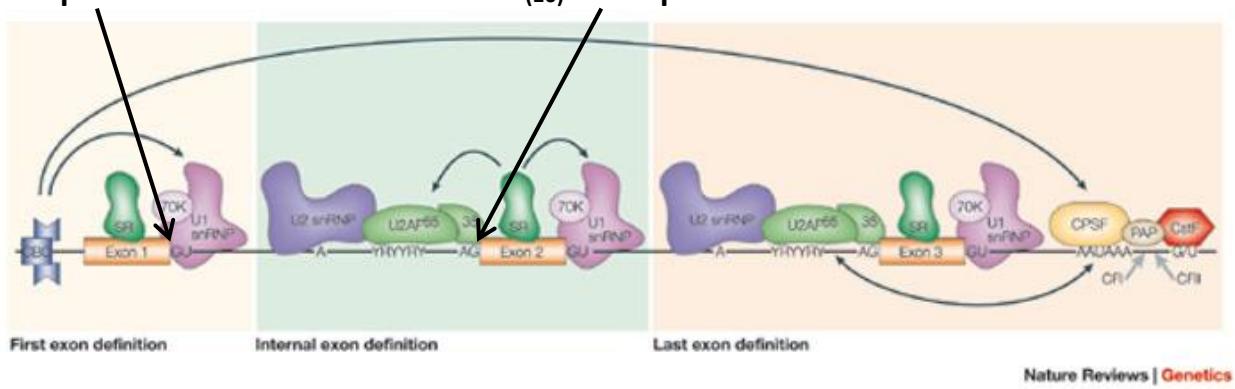


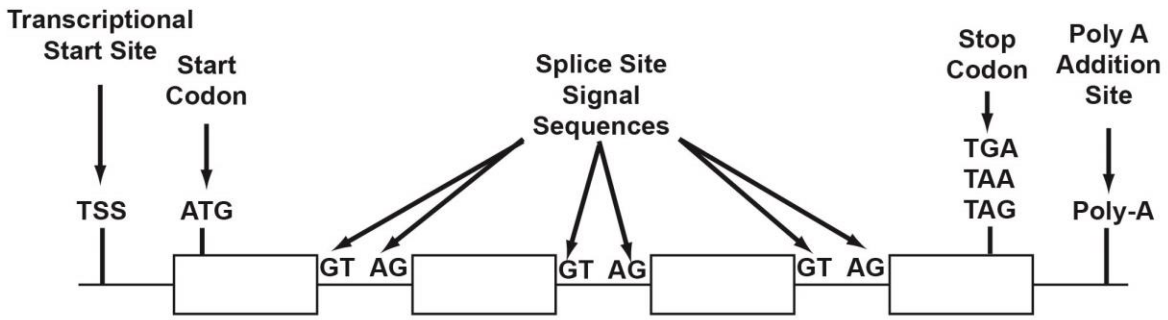
Figure 3 | Exon-definition model. Typically, in vertebrates, exons are much shorter than introns. According to the exon-definition model, before introns are recognized and spliced out, each exon is initially recognized by the protein factors that form a bridge across it. In this way, each exon, together with its flanking sequences, forms a molecular, as well as a computational, recognition module (arrows indicate molecular interactions). Modified with permission from Ref. 26 © (2002) Macmillan Magazines Ltd. CBC, cap-binding complex; CFI/II, cleavage factor I/II; CPSF, cleavage and polyadenylation specificity factor; CstF, the cleavage stimulation factor; PAP, poly(A) polymerase; snRNP, small nuclear RNP; SR, SR protein; U2AF, U2 small nuclear ribonucleoprotein particle (snRNP) auxiliary factor.

From: Nature Review Genetics (2002) 3:698-709

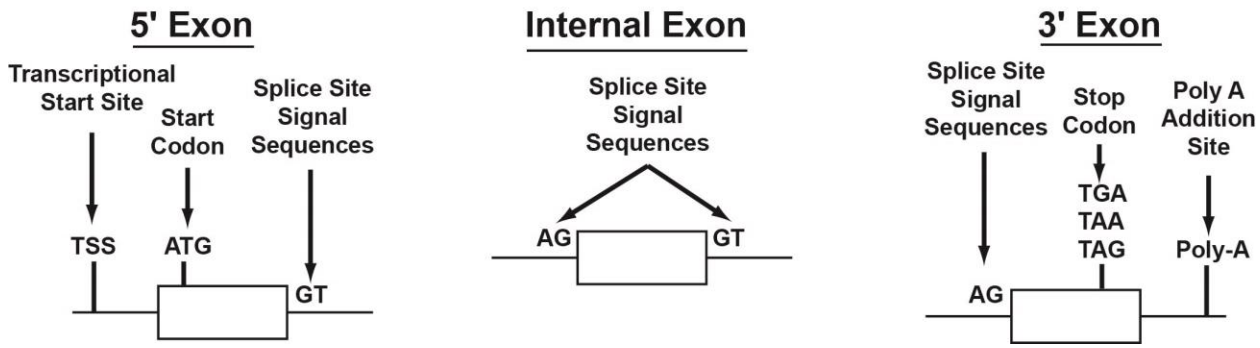
Intrinisic gene structure data

**data defines some structure found in all gene models

General Features of a Eukaryotic Gene



Specialized Gene Regions That Need Unique Gene Prediction Models



All of this information is called

INTRINSIC DATA!!!

Finding Genes In a Sea of Nucleotides

Extrinsic gene structure data

**data related to some function of a gene model

**function = expressed as a RNA molecule

Extrinsic Content Detection

- **Uses data from databases to discover genes**
  - Search genomic sequence as a query against
    - *Protein databases*
    - *Nucleotide databases*

Historical data sources used to discover gene models

- RNA-seq (and historically EST) data
  - **Best information for gene structures**
    - Known to represent genes
    - Contain 3' sequences that are normally gene specific

Best current data sources to discover gene models

- **Problems**
  - Exon-intron borders not always easy to predict
  - 5'-UTR sequences cannot be predicted

RNA-seq data is EXTRINSIC DATA!!!

In plants today, 90% of gene models are being predicted using RNA-seq type data!!!

Hidden Markov Models

***Ab initio* gene modeling**
**Computer aided gene modeling
****Uses intrinsic data**

Definition

- A statistical approach that builds on ***prior knowledge*** and assigns a ***probability*** that a certain ***sequence set*** is a ***member of specific state***.

State

- 5'UTR
- Exon
- Intron
- 3'UTR
- Polyadenylation site

Example: Predicting a "GT" splice site

- **Prior knowledge**
  - Sequence of splice site begins with GT
- **Given a specific nucleotide sequence**
  - GTGAG
  - ***What is the probability that the 5' G is start of an intron splice site if the next nucleotide is a T?***
- **If the next nucleotide is indeed a T, a high probability exists that it is part of the GT splice site IF other properties of the state associate it with the GT splice site state**
  - ***Is the GT preceded by an [A/C]AG sequence?***
  - ***Is the GT followed by [A/G]AGT?***
    - **If yes, then all criteria met, so**
      - **[A/C]AGGT[A/G]AGT is a splice site and defines an intron**

Intrinsic Content Detection

- ***Finding internal exons***
  - Based on splice site features
    - Acceptor site
      - ***Exon|Intron***
      - AG|**<u>GT</u>**RAGT (R = A or G)
    - Donor site
      - ***Intron|Exon***
      - YYYYYYYYYYNC**<u>AG</u>**|G
        (Y = C or T; N = any nucleotide)
- ***Finding 5' exons***
  - Difficult process
    - 5' signal not fully defined
      - Transcriptional start site (TSS)
        - Few known
      - Promoter
        - Variation among promoters known
  - Algorithms search for:
    - CpG island
      - Normally gene rich
      - Some algorithms find TSS within the island
      - Some algorithms find TSS associated with TATA box in island
      - Identify ATG start site in island
- ***Finding 3' exons***
  - Identify polyadenylation addition site signal
    - AATAAA
  - Use stop codon as a 3' prediction signal
    - Essential for determining where one gene ends and another begins

- Finding intronless exons
    - Difficult task
    - Must distinguish these from
        - Long internal exon
        - Pseudogenes that occurred by lost of intron in normal gene

What does a gene prediction program do?

- Calculates best scores for all gene features
  - Defines likelihood that neighboring coding features are really part of a gene
  - Likelihood is calculated as a
    - Weight
    - Probability
      - Hidden Markov Model (HMM) approach is currently preferred approach
        - DNA fragments (a few nucleotides at a time) are defined as a state
        - Probability that a neighboring state can be coupled with the first state to form a gene feature is calculated
        - This allows interdependencies between exons to be explored
        - Calculation based on a training set of genes
          - Training set are genes from a similar taxonomic group with "putative" similar gene features
        - Probability that multiple states can define a gene is calculated

Predicting Multiple Genes

- HMM approaches easily extended to study both strands of a DNA sequence simultaneously
  - Value of modeling both strands
    - Prevents predicting two genes that overlap on the two strands
      - A rare eukaryotic event
- Need to understand features common across chromosomes
  - Insulator elements
  - Boundary elements
  - Matrix attachment regions
- Scaffold attachment regions

Comparative analysis

- One gene set can aid discovery in a related species
  - Gene order is conserved
  - Gene structure is conserved
  - Provides additional training set data for gene prediction
  - Example: Human gene models supporting mouse gene discovery

Synteny: important concept

****Shared gene order between two species**

****Result of related duplication histories**

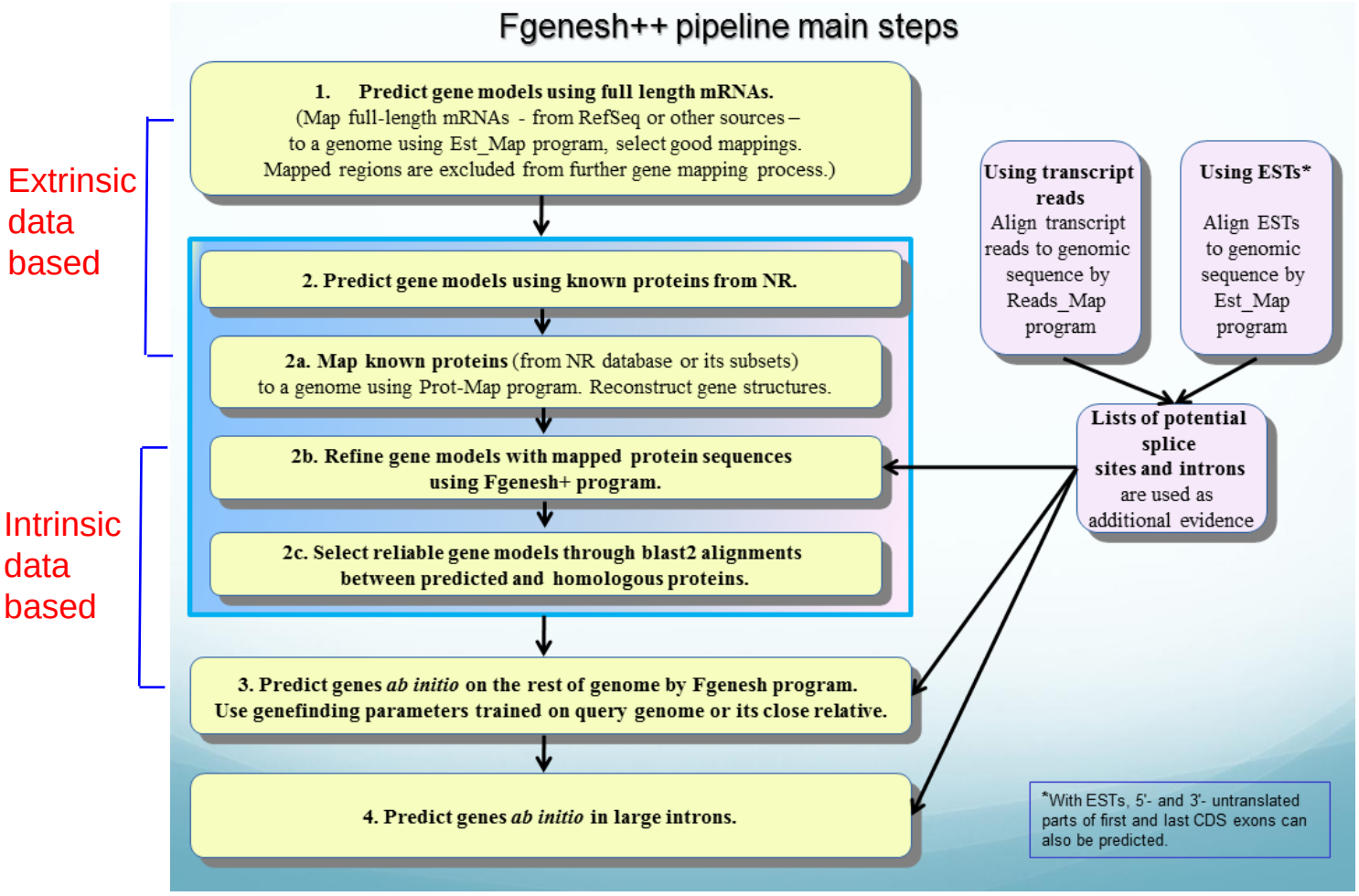
If order conserved,

****The newer genome can inherit gene names of older genomes**

****If the genes are similar**

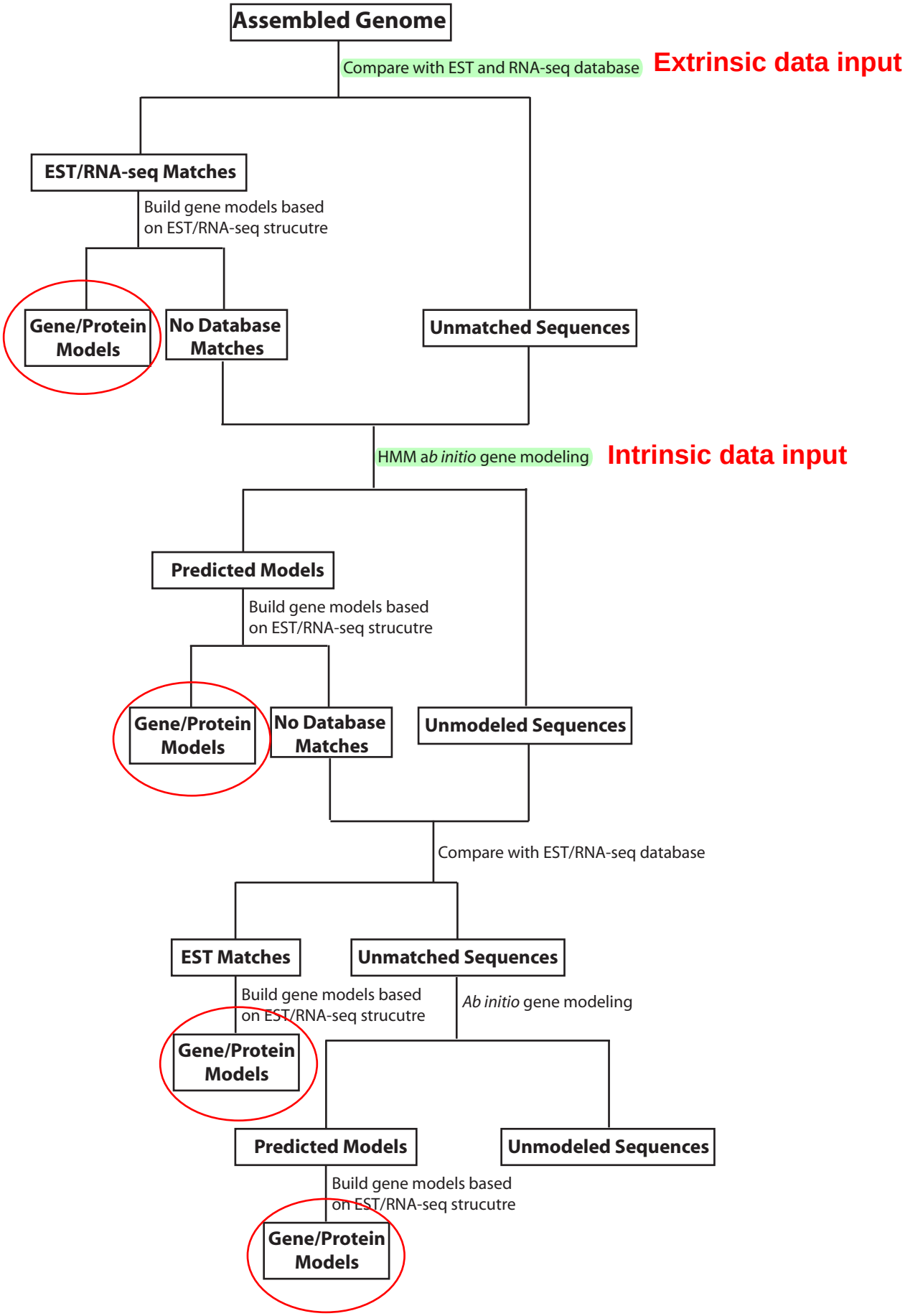
How does a popular *ab initio* software package do it??

FGENESH++C Pipeline (Softberry; quoted directly from company brochure)



FGENESH Gene Modeling Procedure

(www.softberry.com)



Annotation

Naming The Genes

- **Gene naming follows the discovery of potential genes**
- **Relies upon the significant amount of research already available from other genome projects**
    - Historically done on a gene-by-gene approach
        - Goals of gene-by-gene research goal is to clone and characterize an individual gene
    - Each gene is of interest to a specific research group
        - Housekeeping genes
            - Necessary for basic cellular biochemical processes of a cell
            - Nearly all are characterized at the nucleotide and protein levels
            - Sequence information is stored in large databases

The Naming Process

- BLAST ← Starting Tool for Annotation
  - Software tool most often used to annotate (or name) a gene
  - **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Series of computer programs
    - Looks for sequence similarities between two sequences
  - Analysis consists of
    - Query ← All predicted gene models
      - Sequence to which you are looking for a match
      - Nucleotide or amino acid sequences
    - Database ← e.g. Model genome
      - Set of sequences that may be like the query
        - Nucleotide sequences
          - GenBank
        - Protein sequences
          - Swiss-Prot is used to uncover sequences that are similar to the query
    - Translations possible
      - Nucleotide query sequence can be translated
      - Amino acid database sequences can be reverse translated
    - Recent BLAST innovations
      - Gaps can be incorporated to discover matches

Naming The Non-Genes Sequences

- **Other RNA molecules**
  - Important components of the genome
    - **Ribosomal RNAs and tRNAs**
      - Both essential for protein translation
    - **Small nuclear RNAs**
      - Important for RNA splicing
      - Necessary component of the genome
      - Highly conserved
        - Easily recognizable
    - **MicroRNAs**
      - Short RNA sequences
      - 21-25 nt long
      - Negative regulators of gene expression
        - Bind target gene mRNAs and prevent their expression
- **Programs that search specifically for these genes are available molecules**

Regulatory Sequences

- Gene regulation a major area of research
  - Key to understanding gene expression
- **Regulatory motifs discovered**
  - **Motifs**
    - ***Short sequences that define a function***
      - **Nucleotide sequences**
        - Sites where regulatory bind
          - Orthology searches
            - Scan promoter sequences
            - Search for conserved regulatory motifs
      - **Amino acid sequences**
        - Key sequences that bind DNA molecules
          - Orthology searches
            - Scan protein sequences
            - Search for DNA binding motifs
    - Discovered motifs must be tested experimentally
      - Testing motifs is a functional genomics concern

Transcription Factor Binding Motifs

- ***Sequences upstream to which transcription factors bind***
  - Multiple sites possible per transcription factor
  - Motif location can vary because the 5'-UTR length is not consistent

Motif/TF Family Table from: BMC Genomics (2014) 15:317

Motif	Transcription Factor Family
GCTGCCGGAGA	NAC
GCACGTGGAG ATGTGATGC	bHLH bHLH
GGTTGTGGT ACCAAACAT CACCTAAC	R2R3-MYB R2R3-MYB R2R3-MYB