Refining Annotation Methodology for Comparative Genomics in Conifers

Sumaira Zaman

January 13, 2018

@Sumaira_Zaman_

PineRef Seq Project

Project Objectives:

- Gain Fundamental Genetic Information
- Development of Genomic Technologies

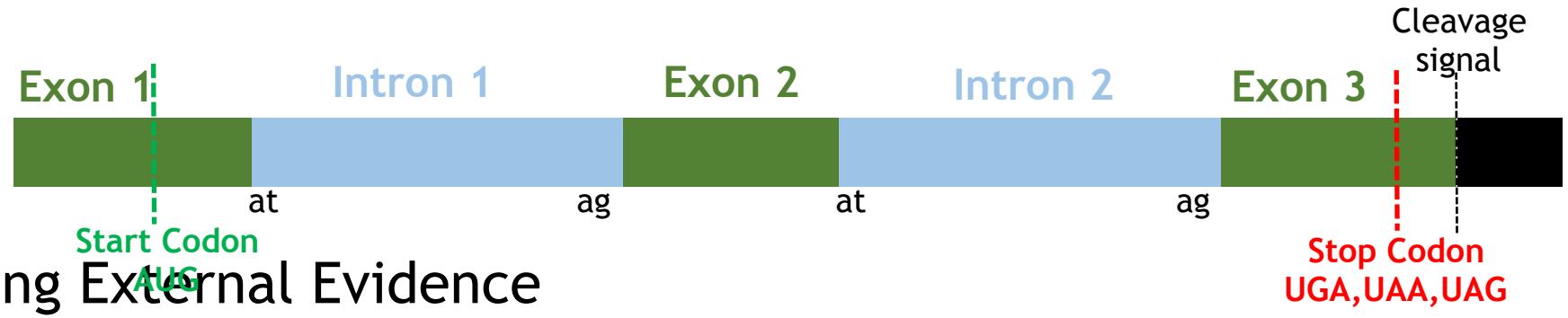
Pseudotsuga menziesii
(Douglas fir)

Pinus taeda
(loblolly pine)

Pinus lambertiana
(sugar pine)

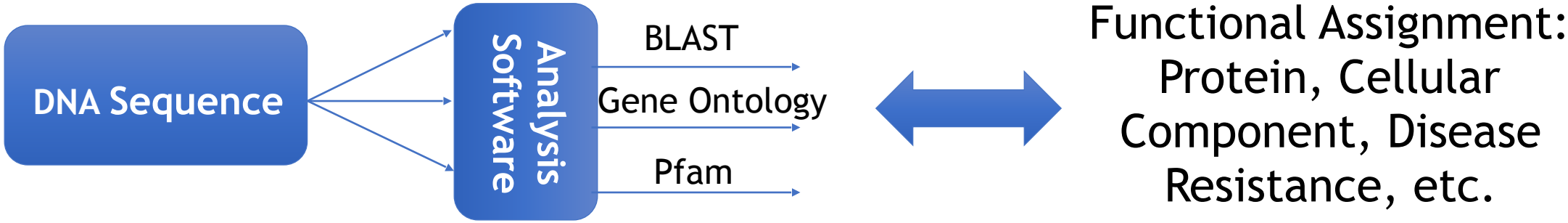
Genome Annotation

- Structural

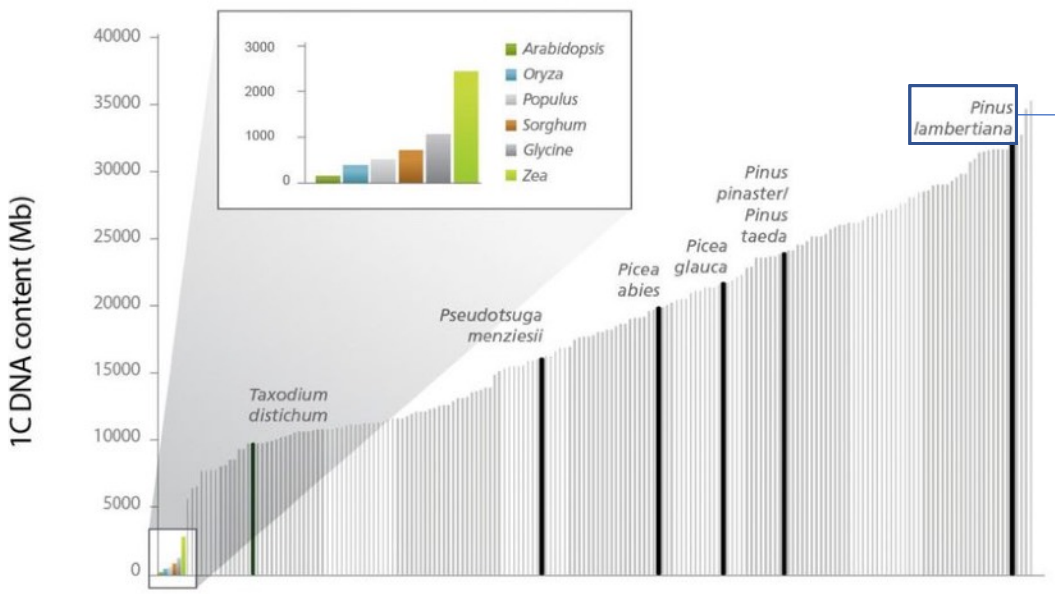


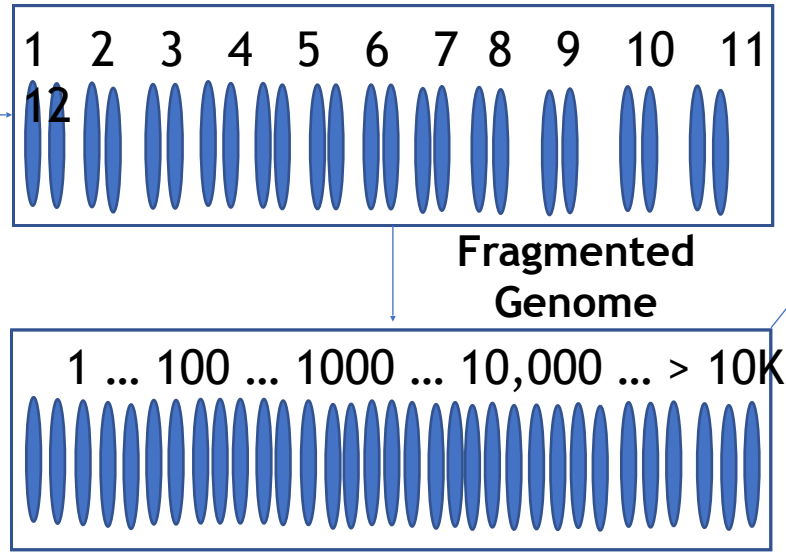
  - Aligning External Evidence
  - *Ab initio* Gene Prediction

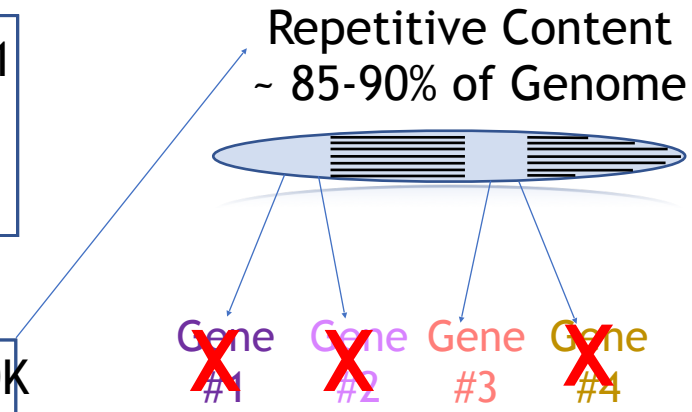
- Functional



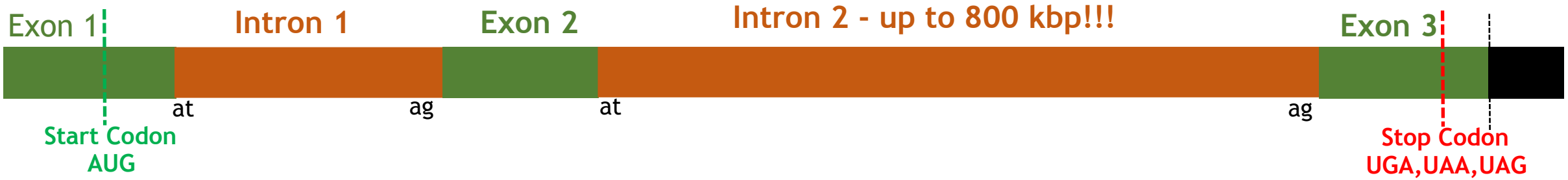
Challenges Facing the Genome







Uncharacteristic gene structure



Sequencing & Annotation Timeline

P. menziesii

v1.0 → N50 = 340 Kb
High Quality Genes ~20K

P. taeda

v1.01 → N50 = 67 Kb
High Quality Genes ~15K

v2.01 → N50 = 107 Kb
High Quality Gene ~46K

P. lambertiana

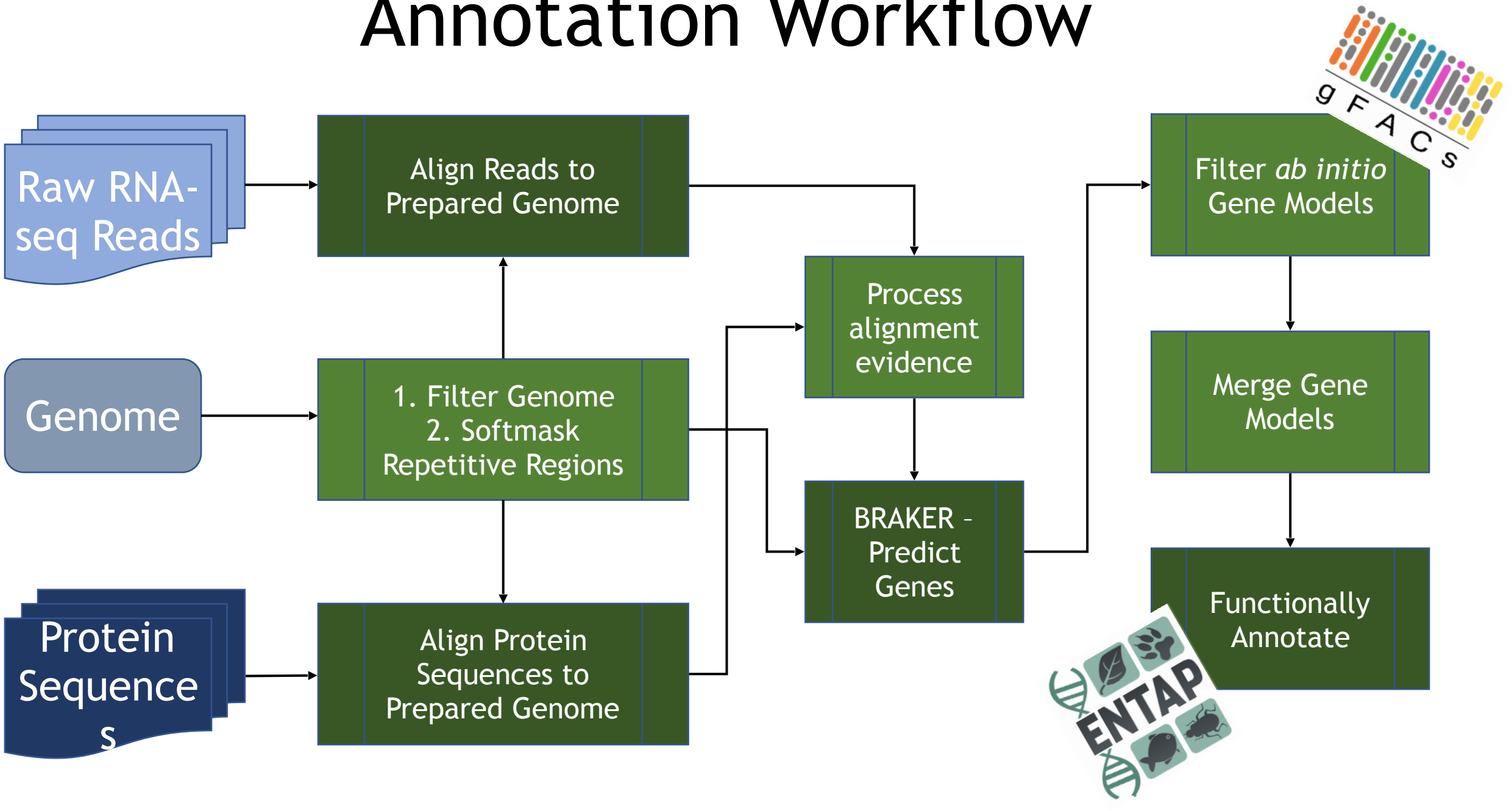
v1.0 →N50 = 247 Kb
High Quality Gene ~14K

v1.5 →N50 = 2500 Kb
High Quality Gen ~31K

Current challenges with prior methodology:

- Number of fragmented genes
- High number of partial genes
- Computational run time

Annotation Workflow



Annotation Results

Species	# of Genes Initially Predicted	High Quality Genes	High Quality Genes Annotated
P. menziesii (Douglas fir)	292,358	46,688	46,682
P. taeda (loblolly pine)	345,382	47,515	46,458
P. lambertiana (sugar pine)	224,590	31,253	31,250

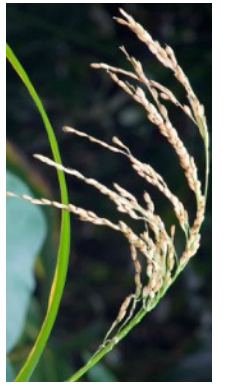
Ginkgo biloba 30,001

Picea abies 42,798

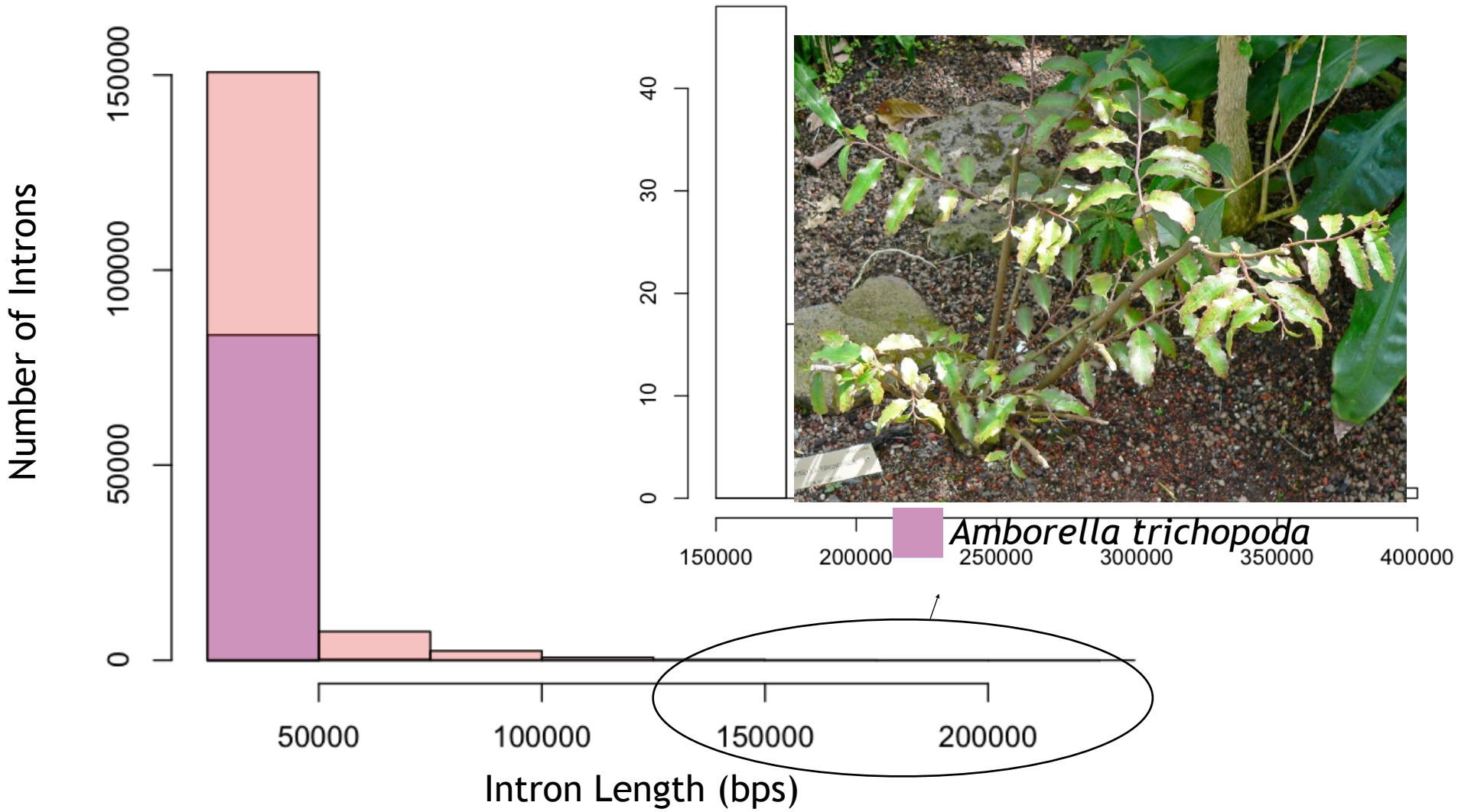
Sorghum bicolor 46,950

Vitis vinifera 54,204

Oryza sativa 49,061



Intron Size Comparison



Pseudotsuga menziesii

Sequencing *Pinus lambertiana* Genome

Sequence → Filtering → Assembly

Haploid

Diploid

Error Correct

Super Read Reduction

Mate pair filters

Assembly N50 = 35 Kbps

Assembly N50 = 196 Kbps

Gap Closing

Transcriptome Scaffolding N50 = 247 Kbps

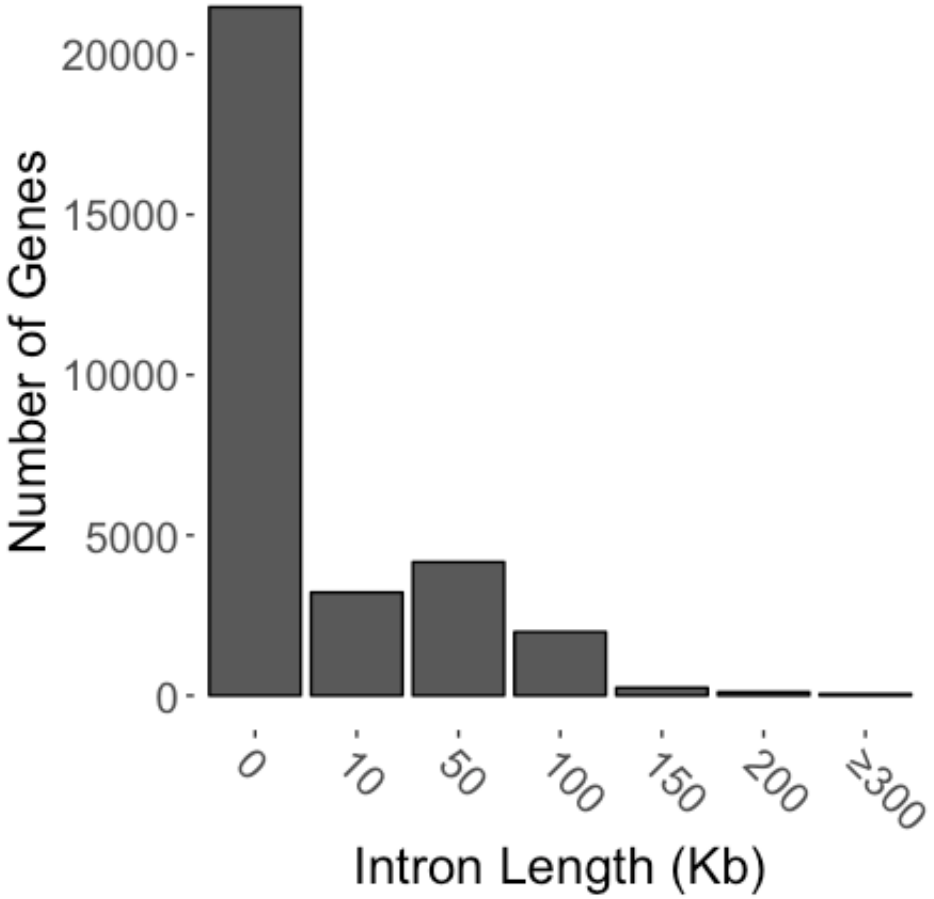
10X Genomics N50=2500 Kbps

Sugar Pine Structural Annotation

Structural Annotation Results

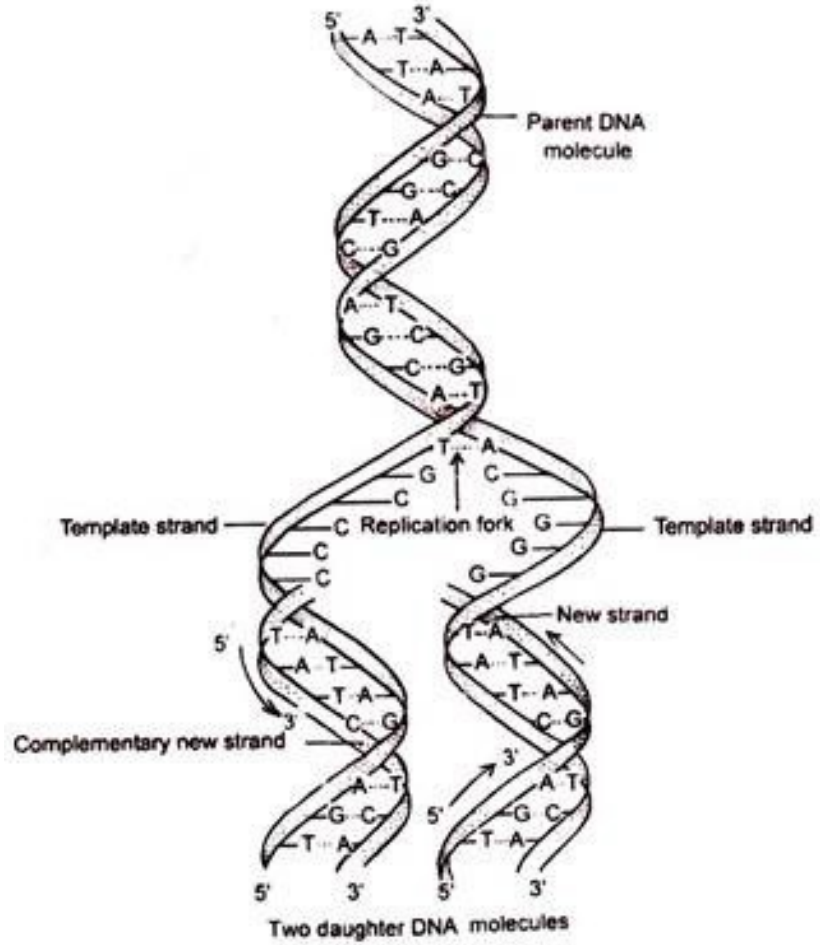
Genic Feature	Length (bp)
Average Gene Size	40,820
Average CDS Size	1,184
Average Exon Size	241
Average Intron Size	10,165
Maximum Intron Size	805,506

Intron Size Distribution

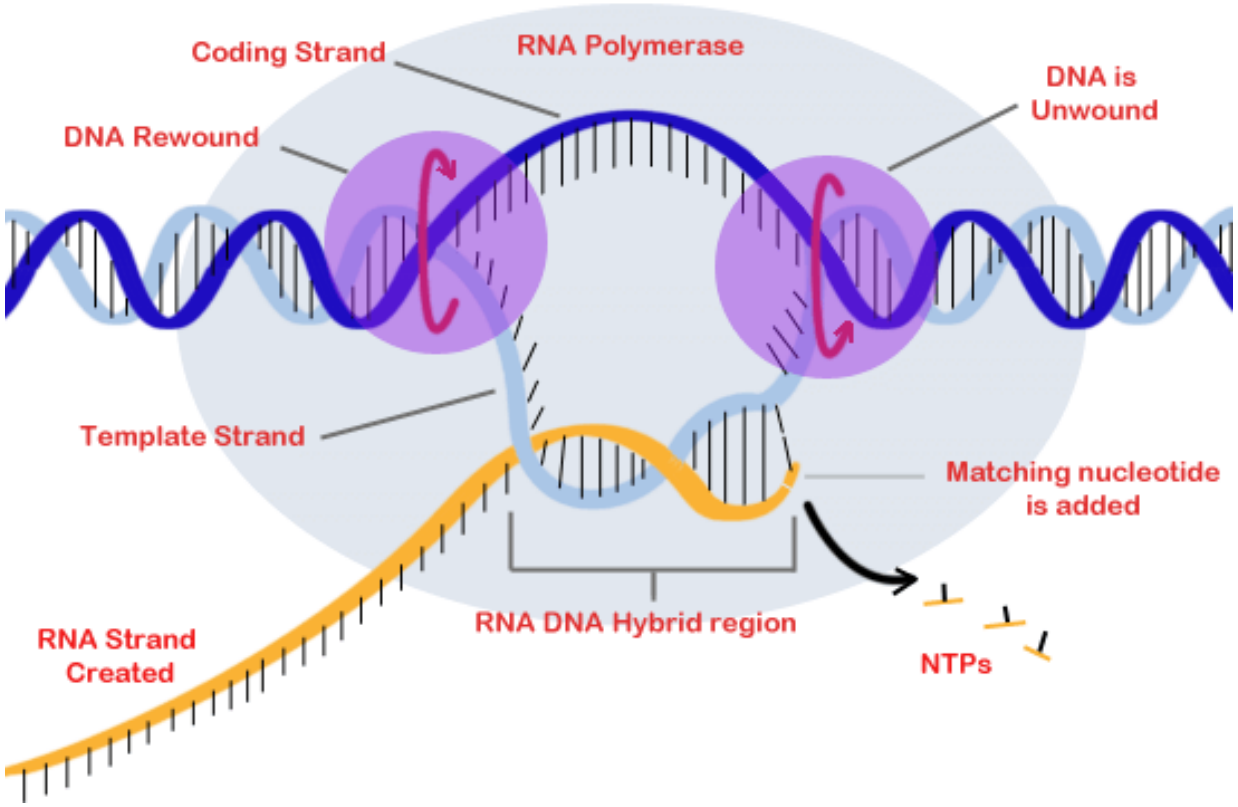


The Burden of Long Introns

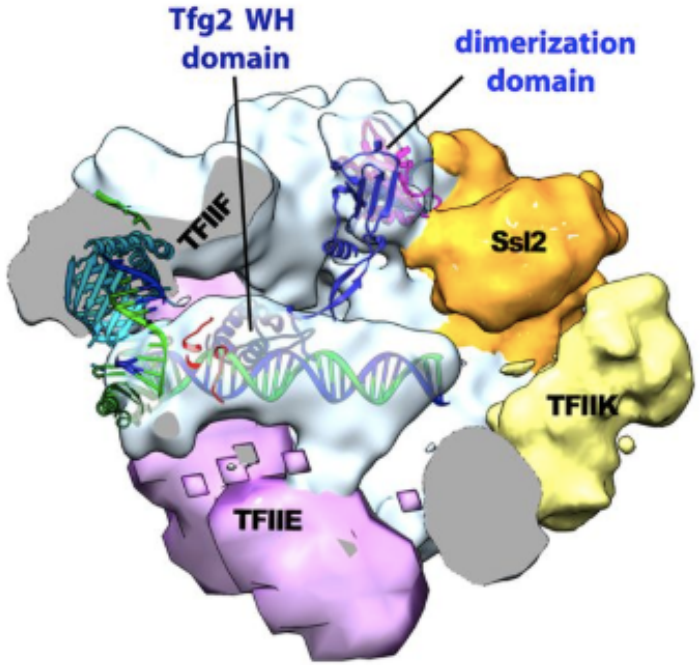
DNA Replication

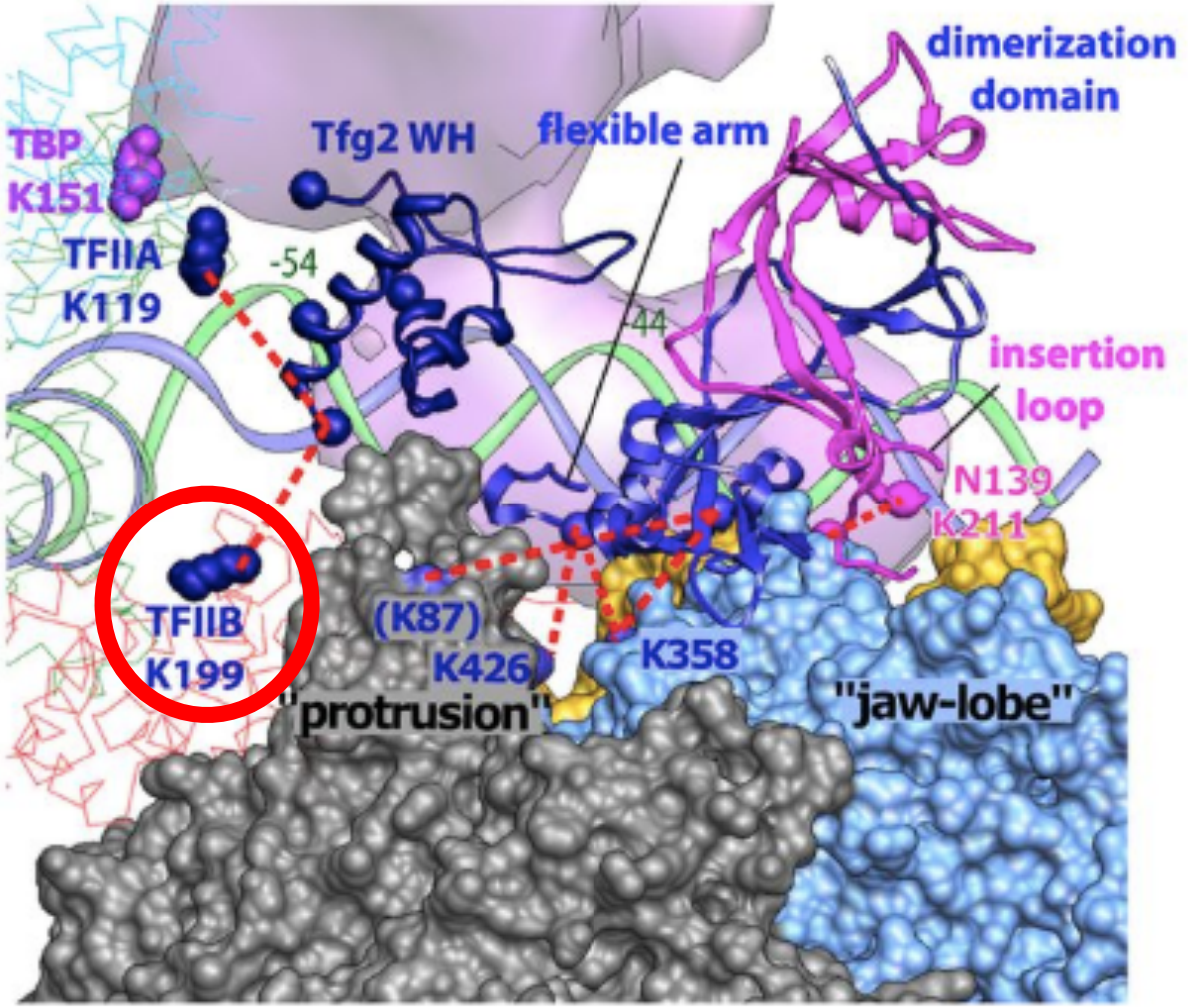


Transcription



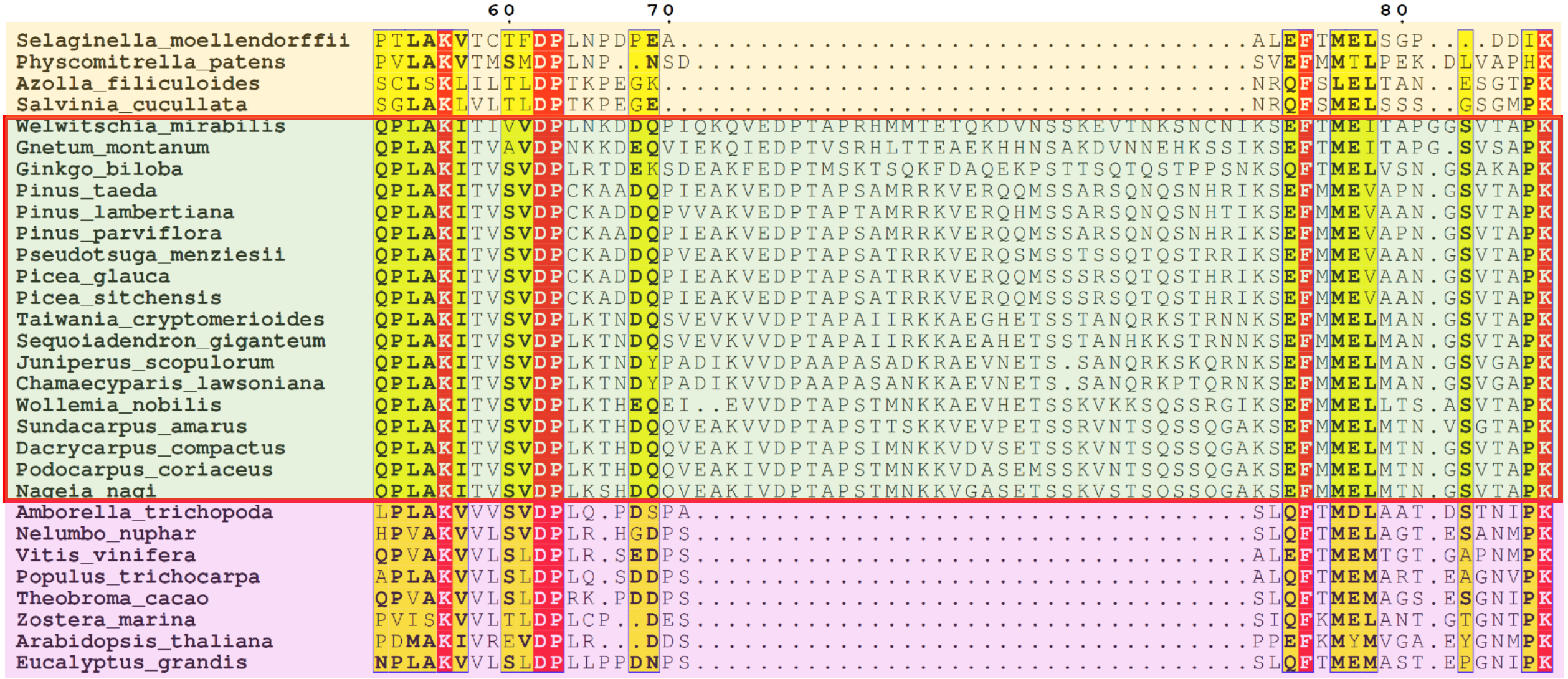
Transcription Machinery – RNA Polymerase II

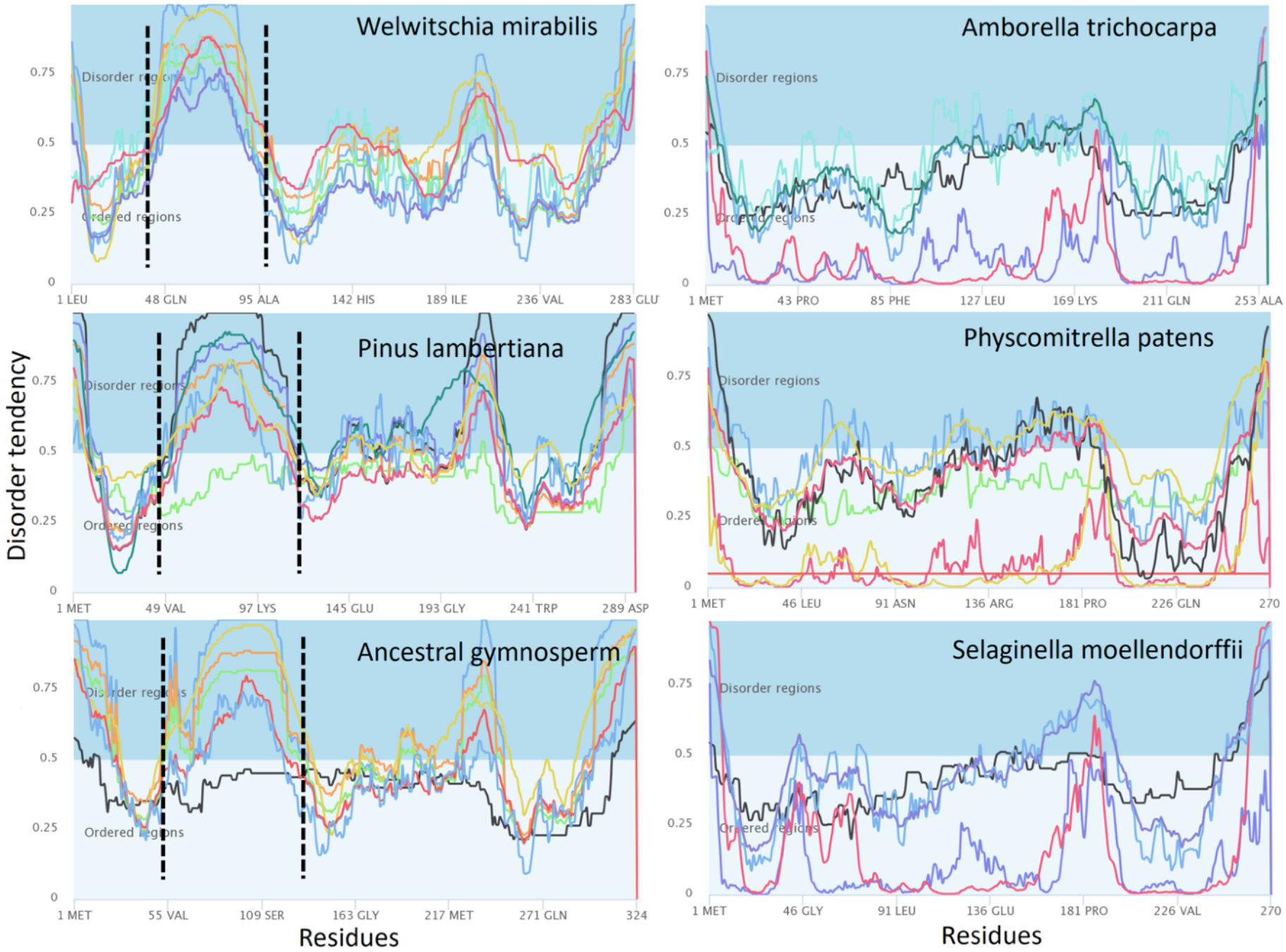




Reference: Murakami, Kenji, et al. “Architecture of an RNA Polymerase II Transcription Pre-Initiation Complex.” *Science* 342, no. 6159 (November 8, 2013): 1238724. doi: 10.1126/science.1238724.

Role of Transcription Factor TFIIFb

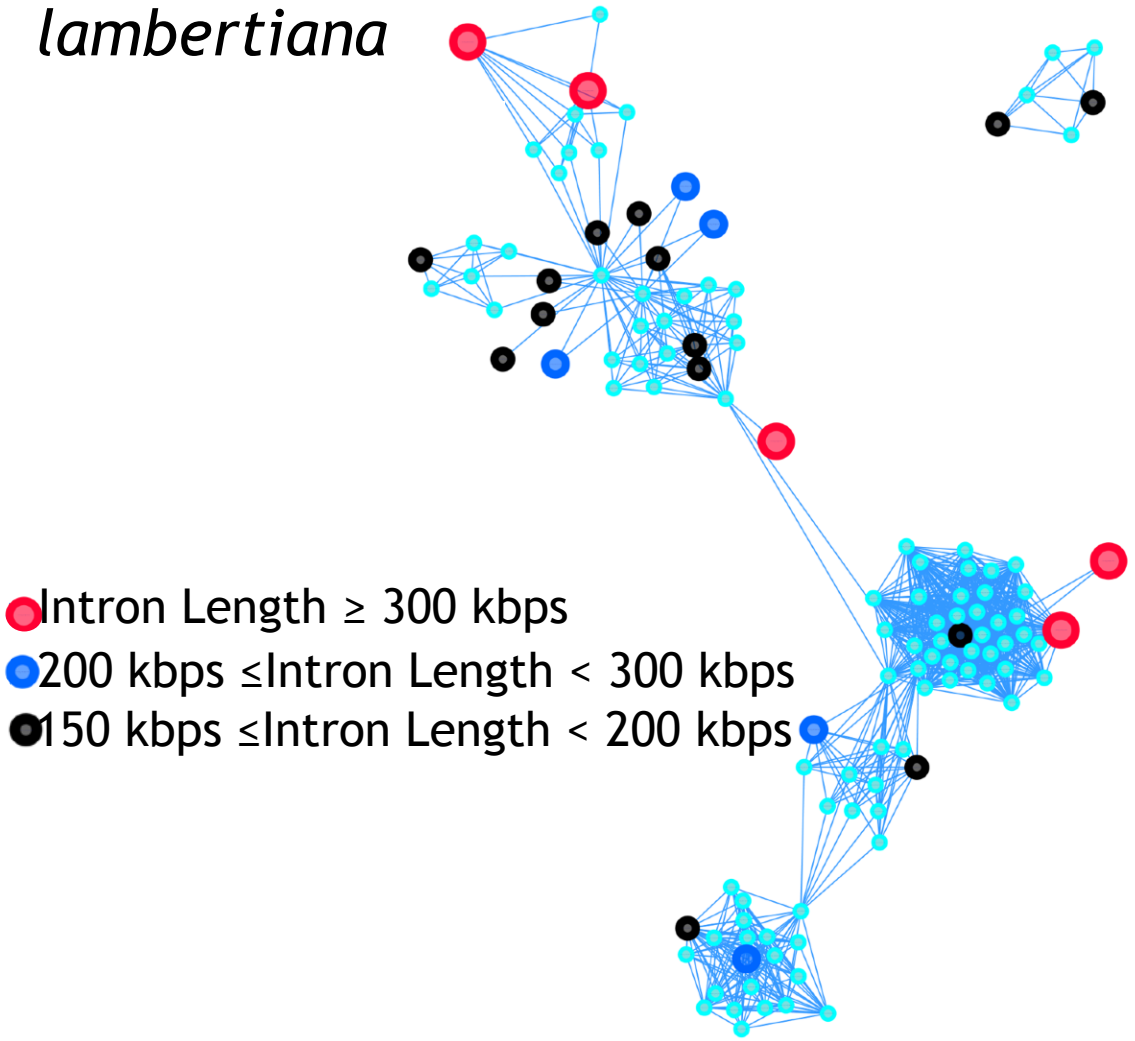




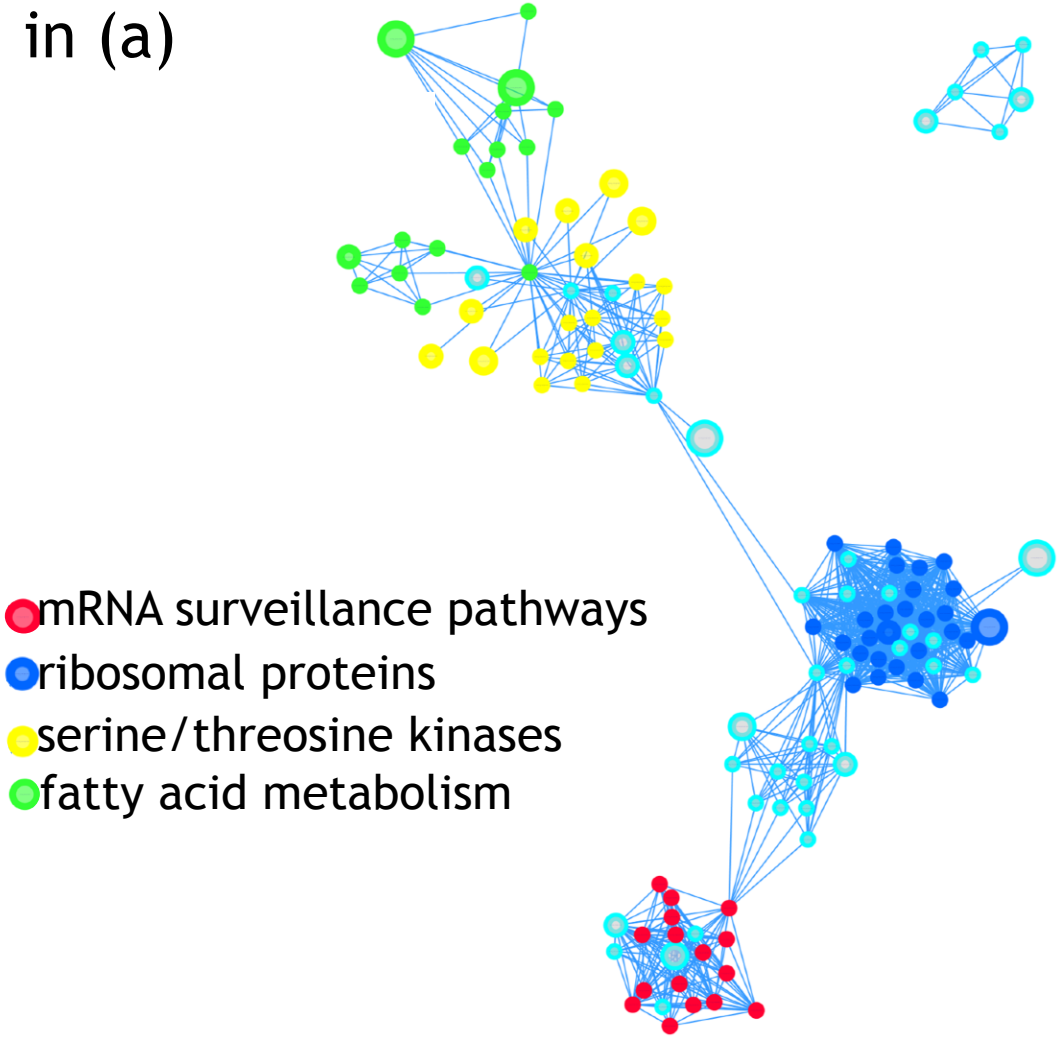
"Order" of this Protein Domain Insertion

Protein Interacting Network Using *V. vinifera*

(a) Interactome network of proteins coded by long intron genes from *P. lambertiana*

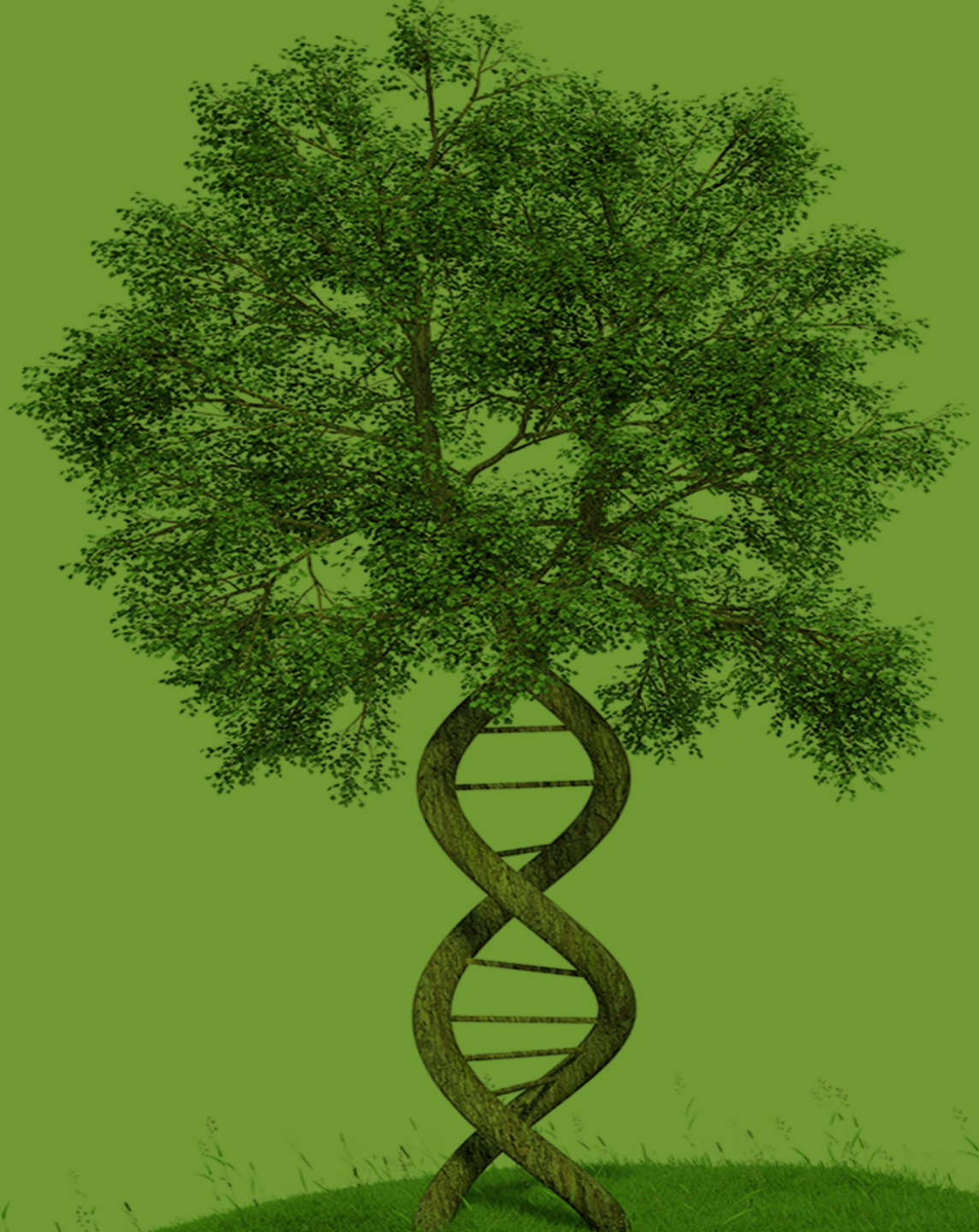


(b) Functional enrichment of the long intron genes present in interactome in (a)



Future Work

- Evaluating selection pressure of long intron genes
- Investigating conifer specific orthogroups
- Intron retention



Thank you!
Questions?

Uzay Sezen Smithsonian/UCONN
Madison Caballero UCONN
Jill Wegrzyn UCONN

USDA NSF UCONN