

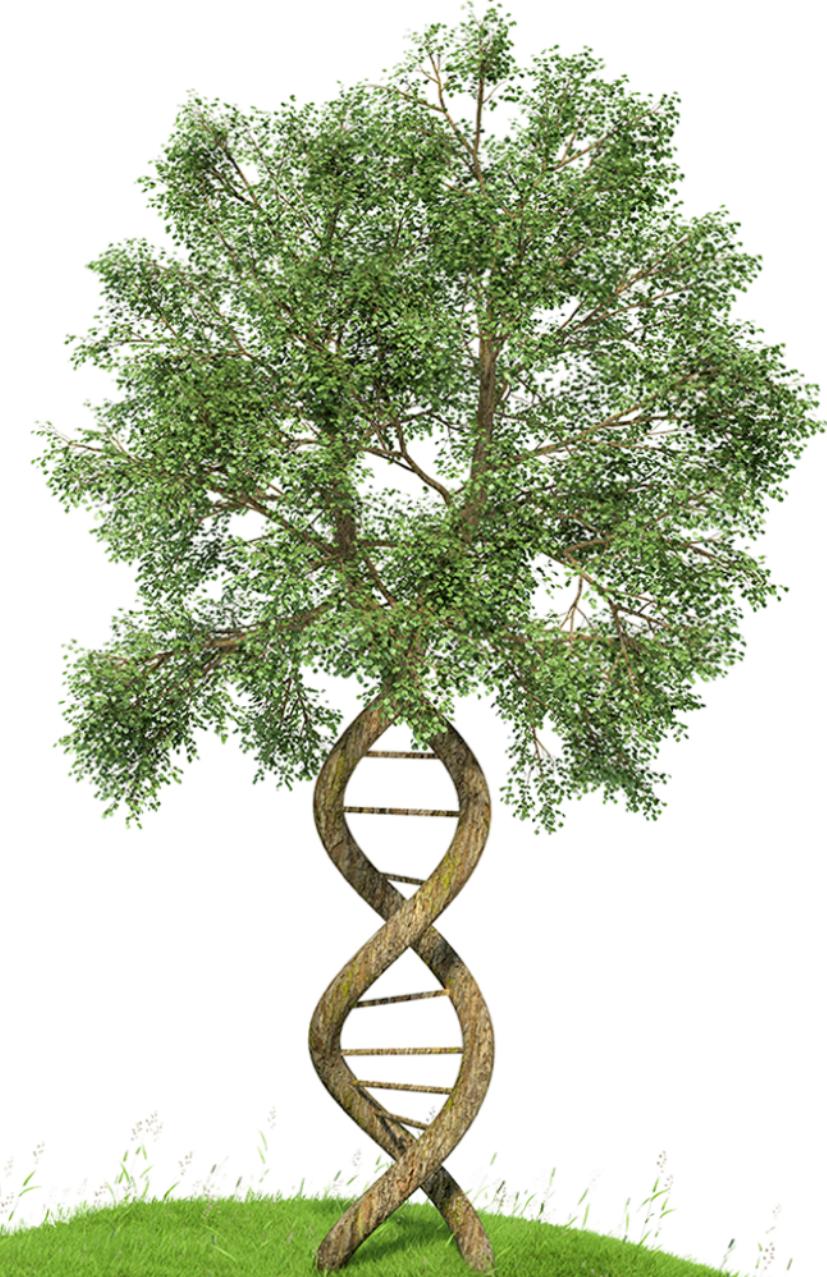
# Refining Annotation Methodology for Comparative Genomics in Conifers

Sumaira Zaman

January 13, 2018



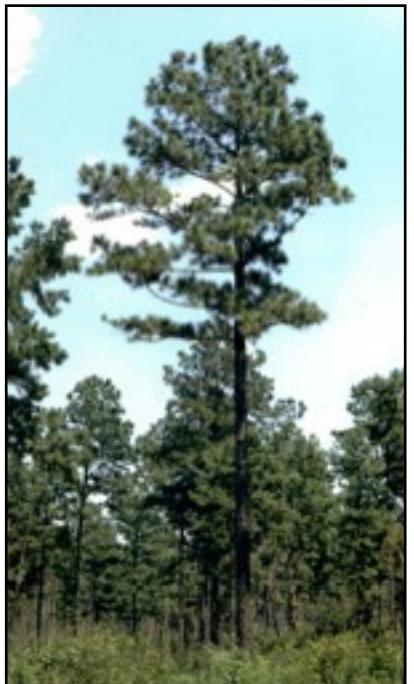
@Sumaira\_Zaman\_



# PineRef Seq Project

## Project Objectives:

- Gain Fundamental Genetic Information
- Development of Genomic Technologies



*Pseudotsuga menziesii*  
(Douglas fir)

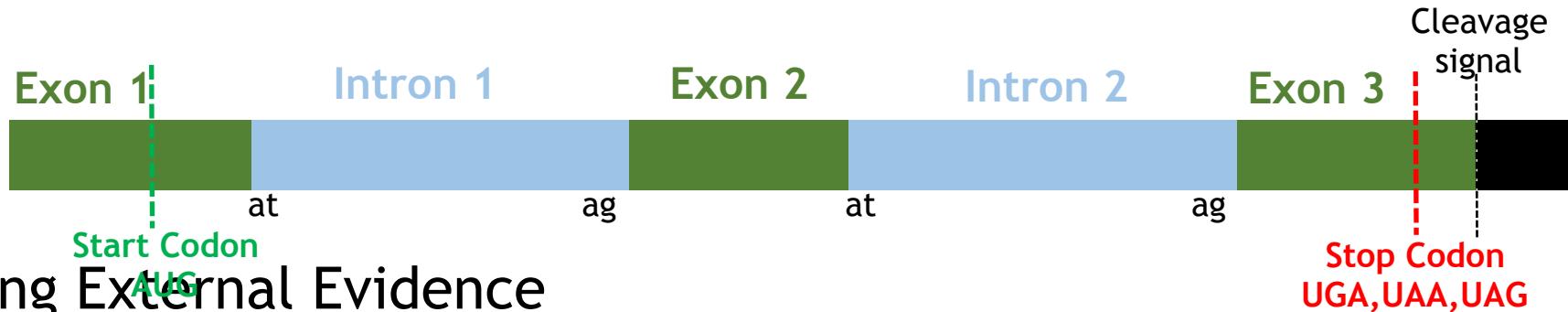
*Pinus taeda*  
(loblolly pine)

*Pinus lambertiana*  
(sugar pine)



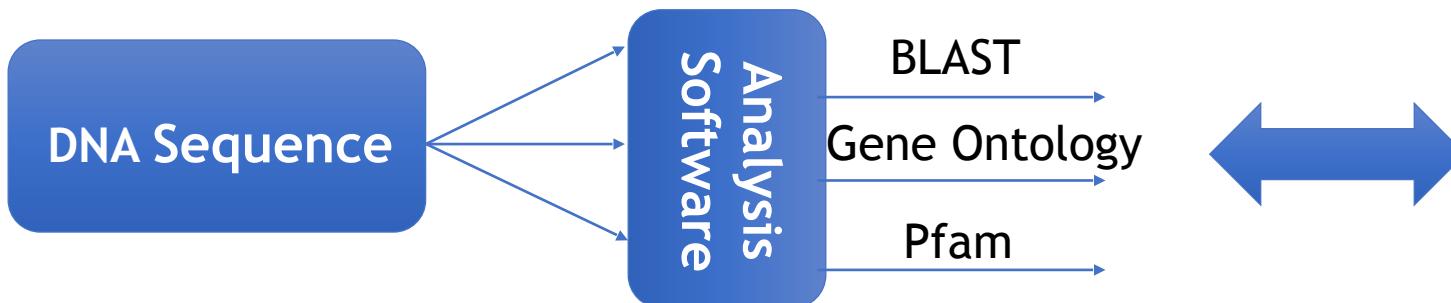
# Genome Annotation

- Structural



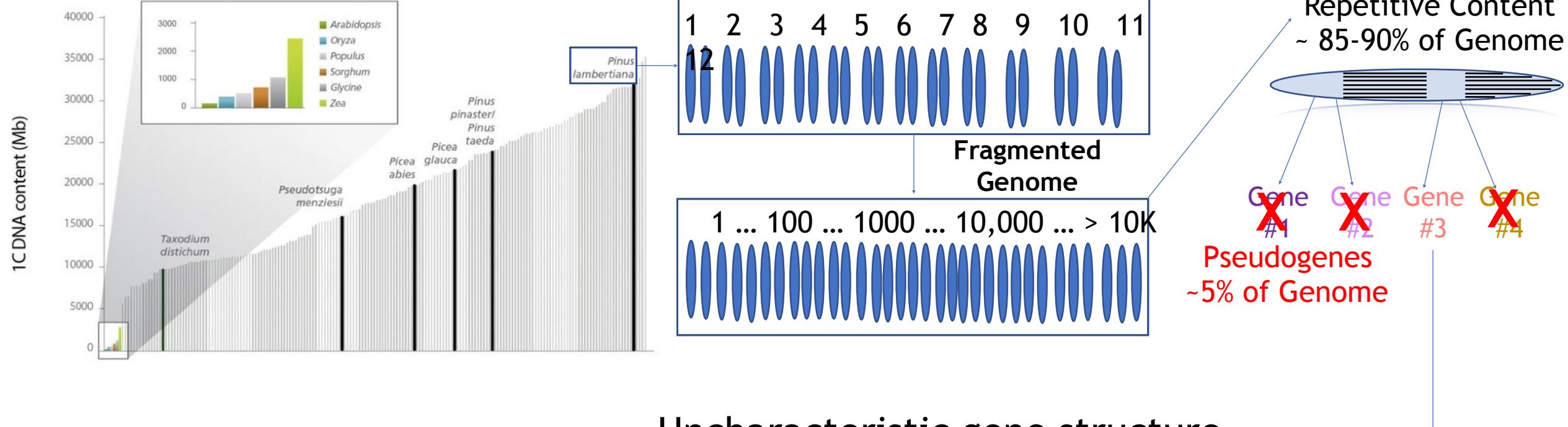
- Aligning External Evidence
- *Ab initio* Gene Prediction

- Functional



Functional Assignment:  
Protein, Cellular  
Component, Disease  
Resistance, etc.

# Challenges Facing the Genome



Uncharacteristic gene structure



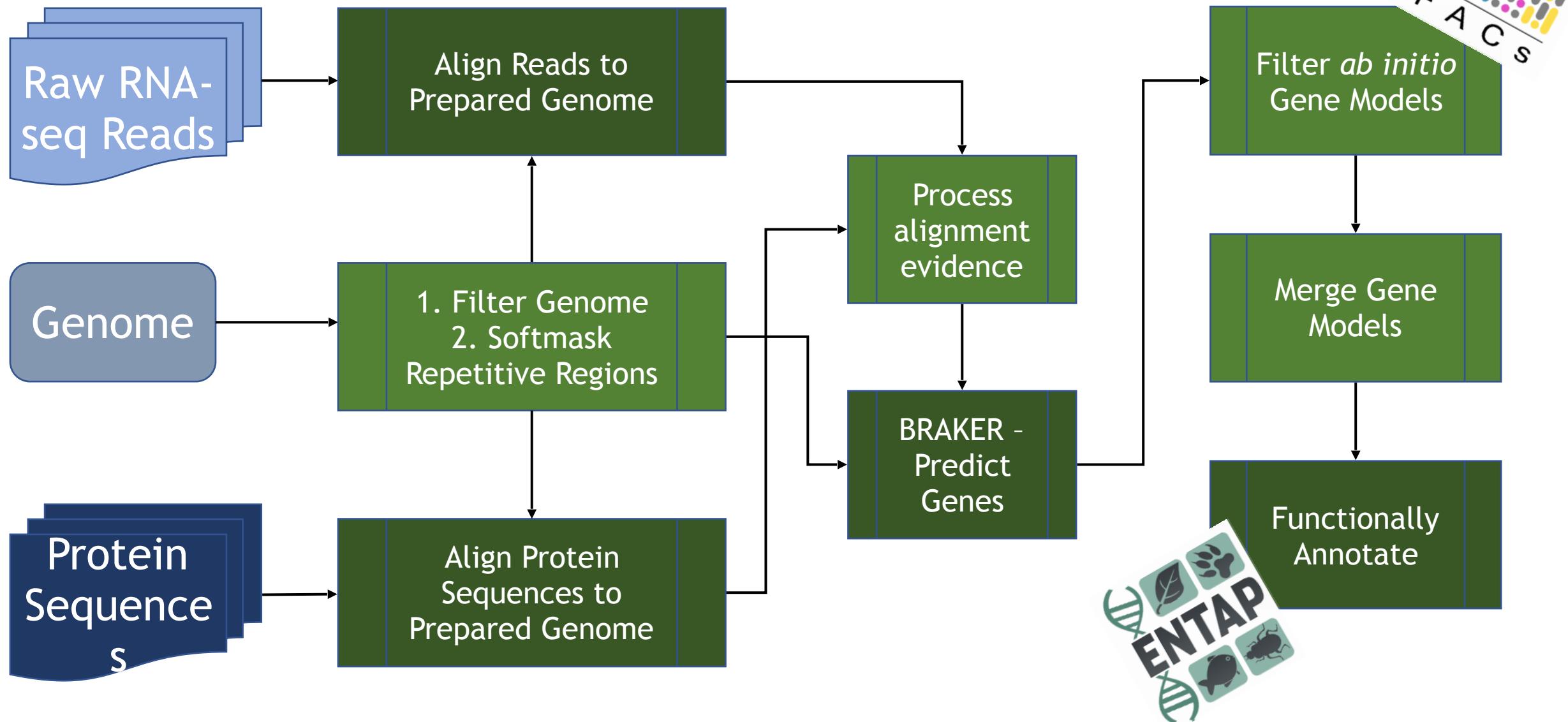
# Sequencing & Annotation Timeline

	2014	2016	2017
<i>P. menziesii</i>		v1.0 → N50 = 340 Kb High Quality Genes ~20K	
<i>P. taeda</i>	v1.01 → N50 = 67 Kb High Quality Genes ~15K		v2.01 → N50 = 107 Kb High Quality Gene ~46K
<i>P. lambertiana</i>		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

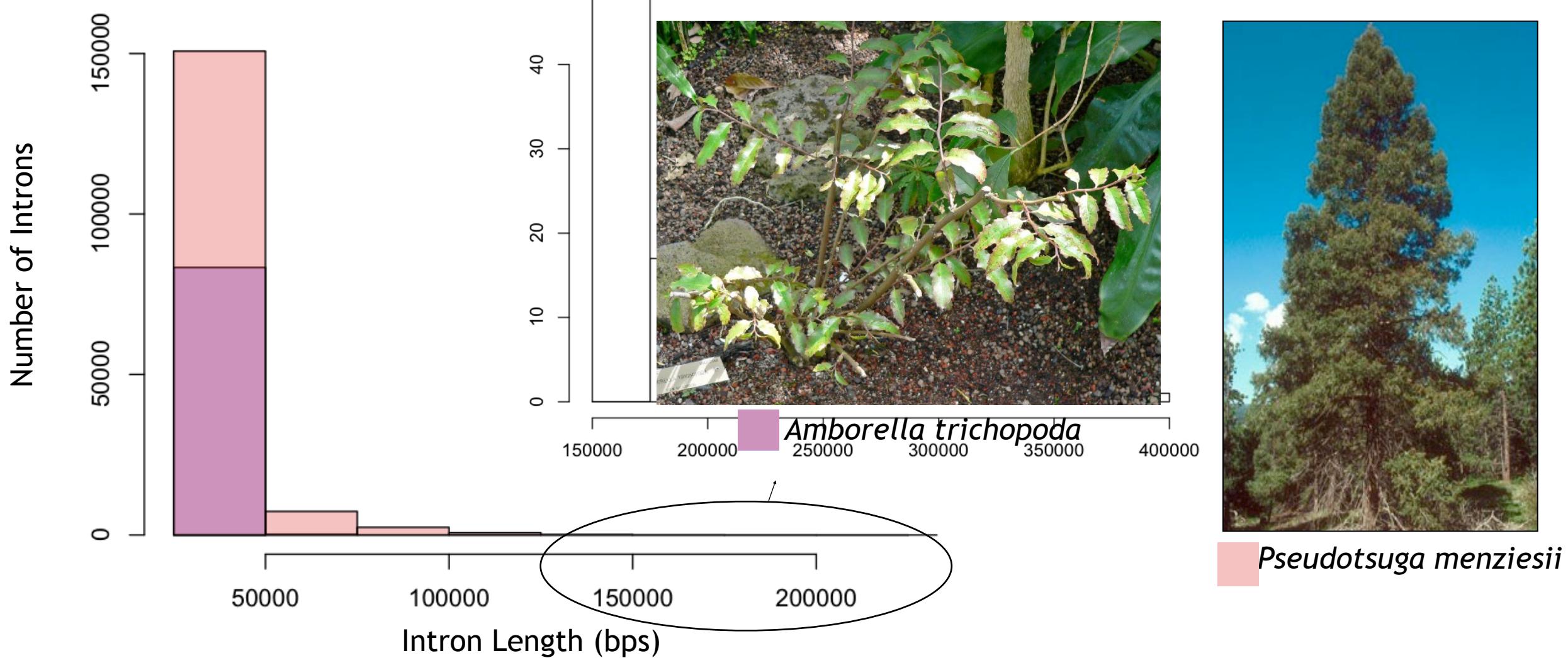
# Annotation Workflow



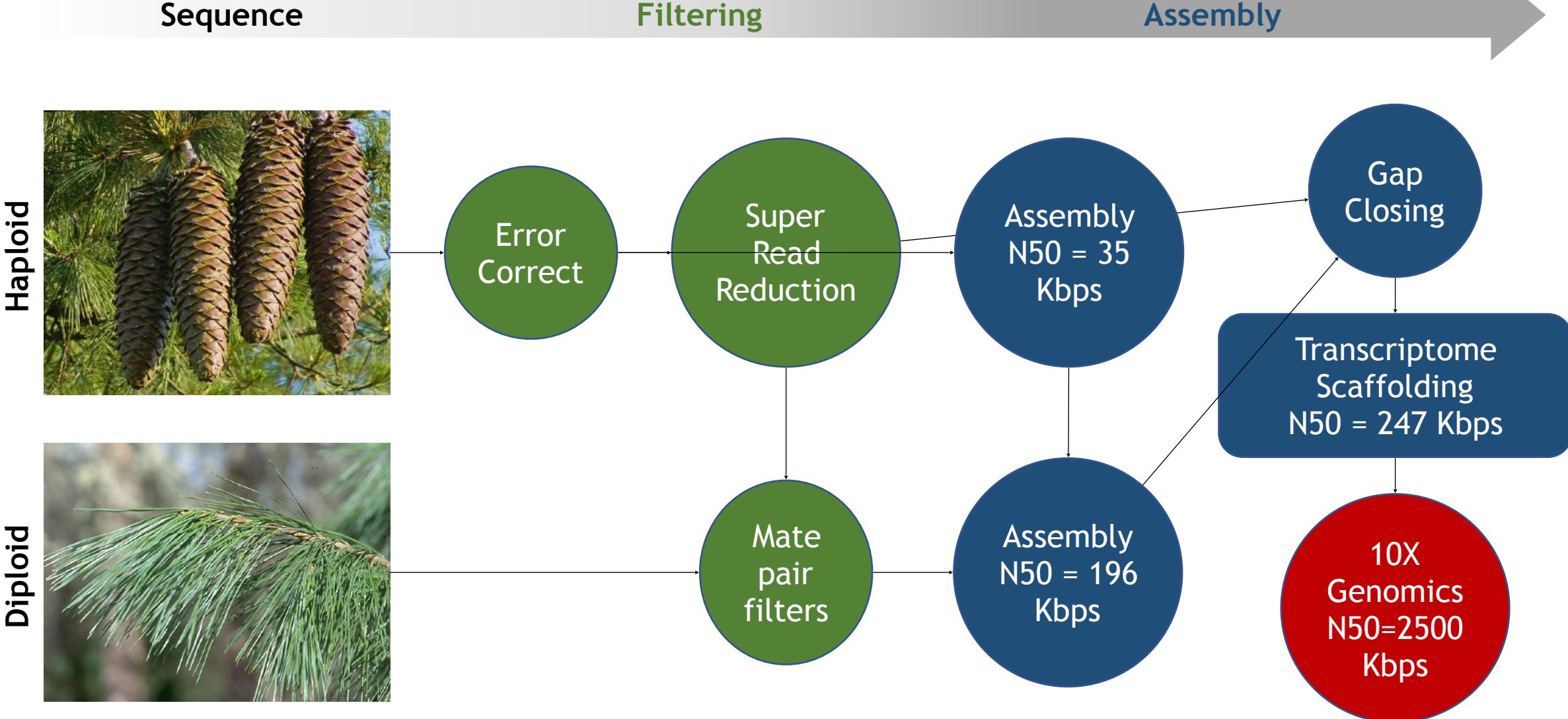
# Annotation Results

Species	# of Genes Initially Predicted	High Quality Genes	High Quality Genes Annotated	
<i>P. menziesii</i> (Douglas fir)	292,358	46,688	46,682	
<i>P. taeda</i> (loblolly pine)	345,382	47,515	46,458	
<i>P. lambertiana</i> (sugar pine)	224,590	31,253	31,250	
<i>Ginkgo biloba</i>	<i>Sorghum bicolor</i>	<i>Vitis vinifera</i>	<i>Oryza sativa</i>	
30,001	42,798	46,950	54,204	
				

# Intron Size Comparison



# Sequencing *Pinus lambertiana* Genome

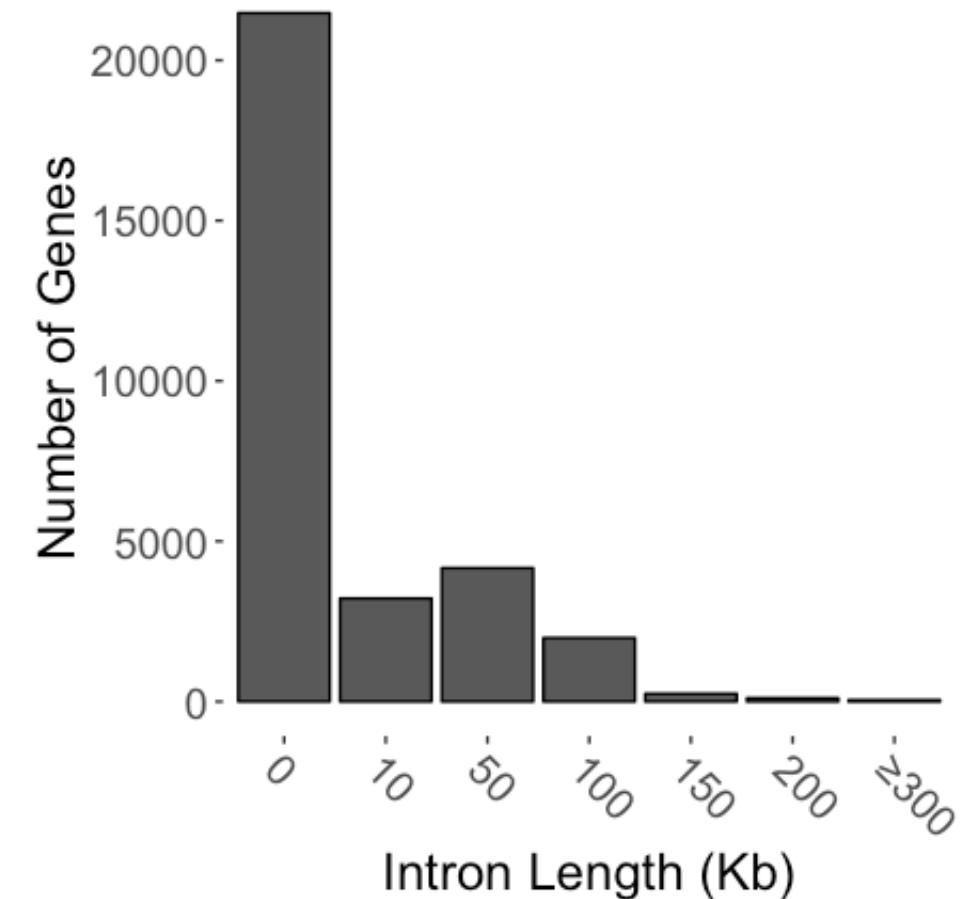


# Sugar Pine Structural Annotation

Structural Annotation

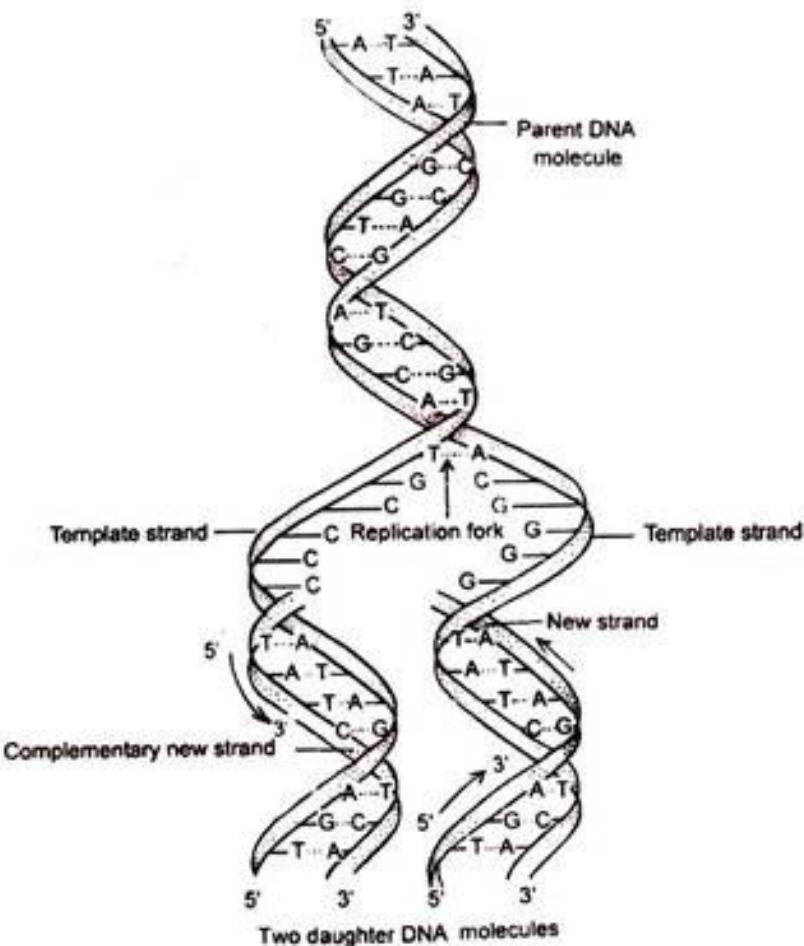
Genic Feature	Results	Length (bp)
Average Gene Size		40,820
Average CDS Size		1,184
Average Exon Size		241
Average Intron Size		10,165
Maximum Intron Size		805,506

Intron Size Distribution

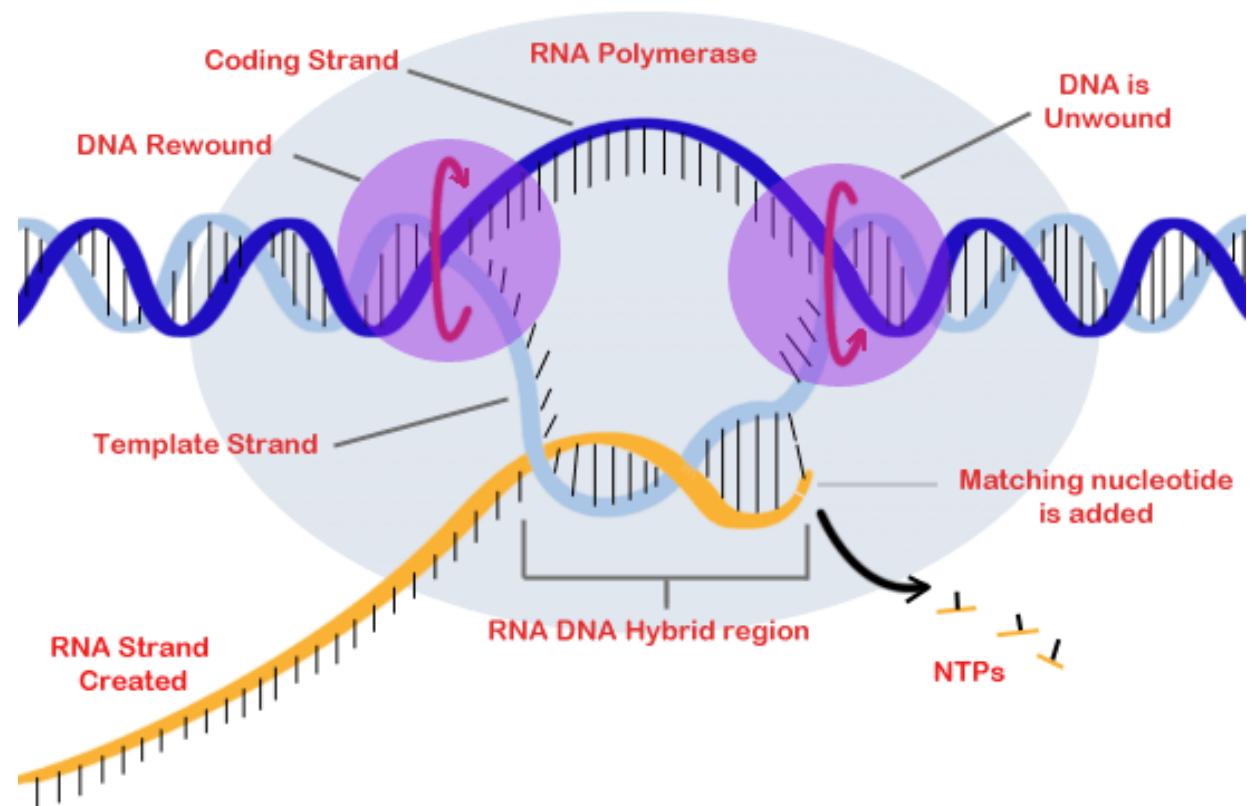


# The Burden of Long Introns

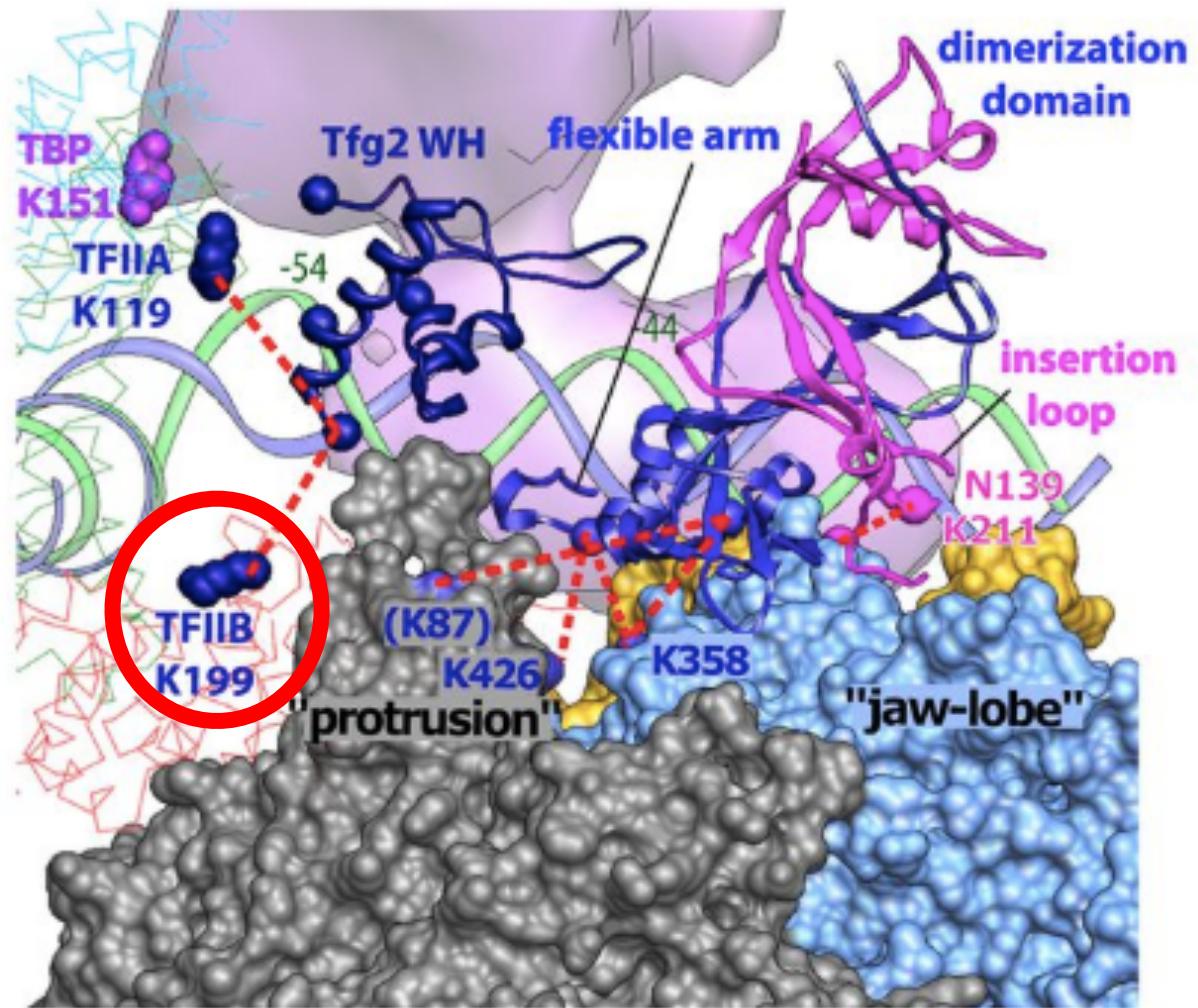
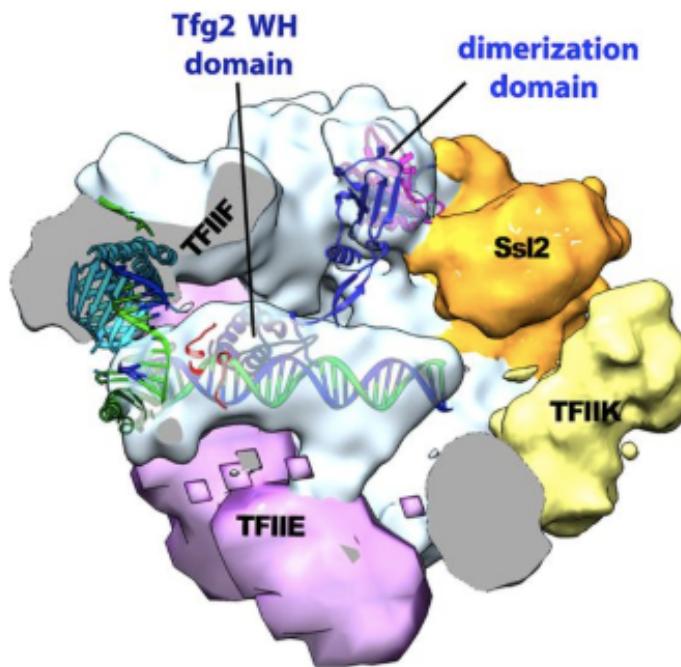
## DNA Replication



## Transcription



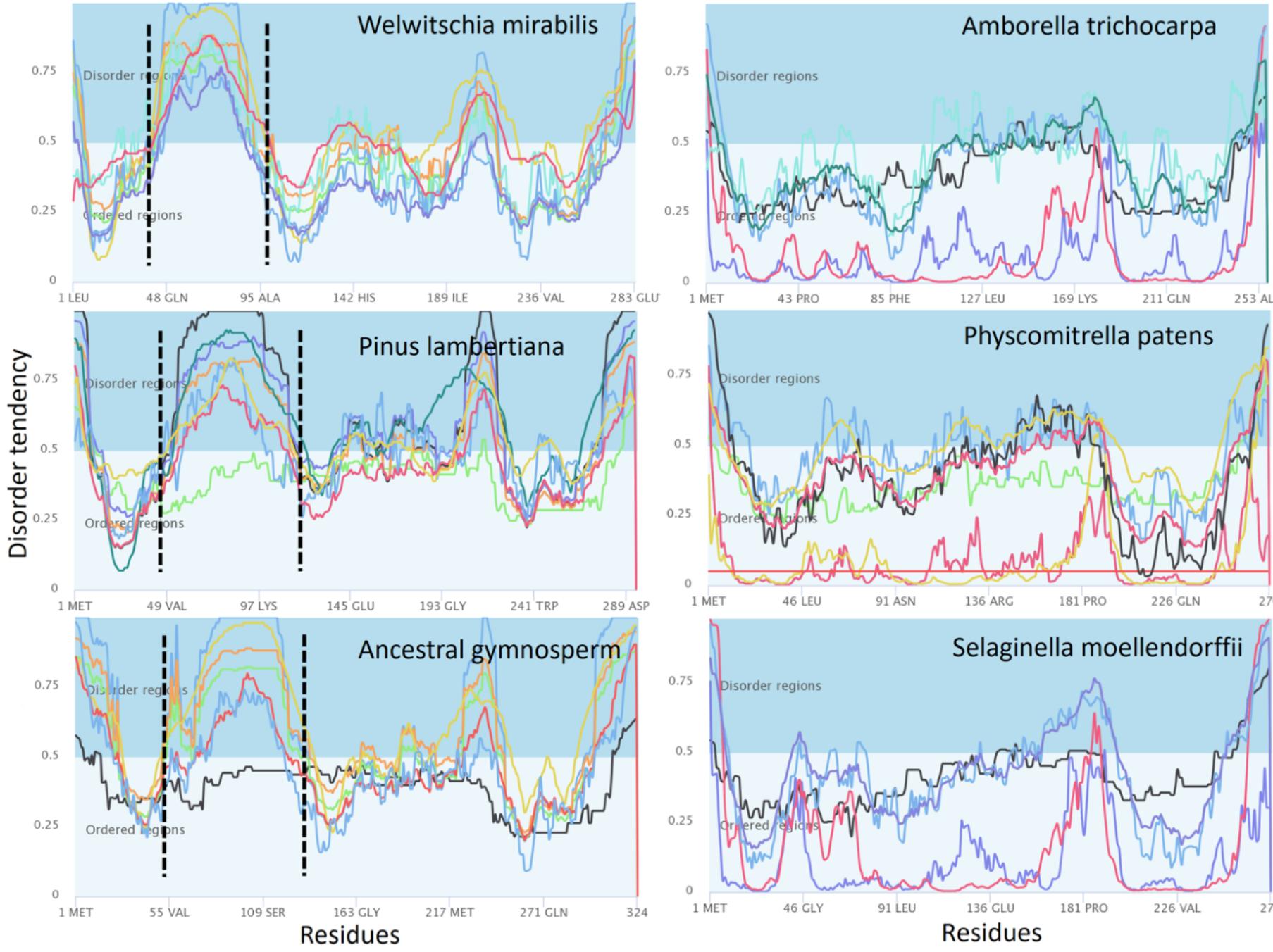
# Transcription Machinery - RNA Polymerase II



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 TFIIIfb

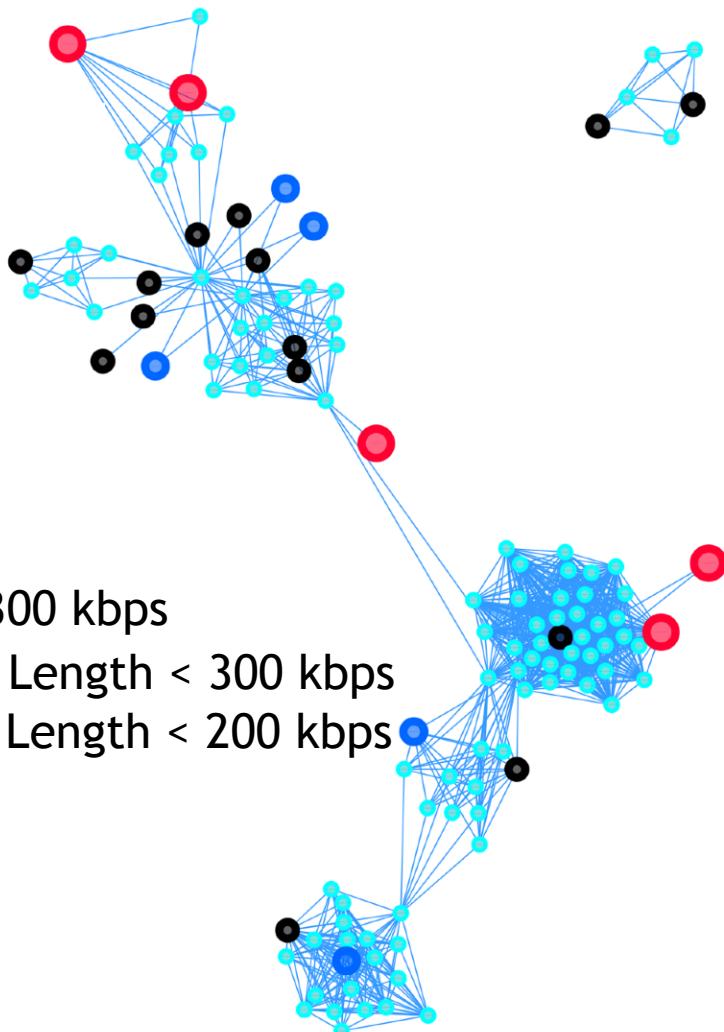
		60	70	80
<i>Selaginella_moellendorffii</i>	PTLAKV	TCTFDP	LNPDPPEA.	AL
<i>Physcomitrella_patens</i>	PVLAKV	TM <del>SM</del> DP	LNP..NSD	SVE
<i>Azolla_filiculoides</i>	SCLS <del>KL</del>	ILTLDP	TKPEGK..	NRQFS
<i>Salvinia_cucullata</i>	SGLAKL	VLTLDP	TKPEGE..	NRQFS
<i>Welwitschia_mirabilis</i>	QPLAKI	TIVVDP	LNKDDQPIQKQVEDPTAPRHMMTETQKDVNSSKEVTNKSNCNIKS	EFT
<i>Gnetum_montanum</i>	QPLAKI	TVAVDP	NKKDEQVIEKQIEDPTVSRHLTTAEAKHHNSAKDVNNEHKSSIKS	EFT
<i>Ginkgo_biloba</i>	QPLAKI	TVSVDP	LRTDEKSDEAKFEDPTMSKTSQKFDAQEKPSTTSQTSQTPPSNKS	QFT
<i>Pinus_taeda</i>	QPLAKI	TVSVDP	CKAADQPIEAKVEDPTAPSAMRRKVERQQMSSARSQNQSNHRIKS	EFT
<i>Pinus_lambertiana</i>	QPLAKI	TVSVDP	CKADDQPVVAKVEDPTAPTAMRRKVERQHMSSARSQNQSNHTIKS	EFT
<i>Pinus_parviflora</i>	QPLAKI	TVSVDP	CKAADQPIEAKVEDPTAPSAMRRKVERQQMSSARSQNQSNHRIKS	EFT
<i>Pseudotsuga_menziesii</i>	QPLAKI	TVSVDP	CKADDQPVVEAKVEDPTAPSATRRKVERQSMSSTSQTQSTRRIKS	EFT
<i>Picea_glauca</i>	QPLAKI	TVSVDP	CKADDQPIEAKVEDPTAPSATRRKVERQQMSSRSQTQSTRRIKS	EFT
<i>Picea_sitchensis</i>	QPLAKI	TVSVDP	CKADDQPIEAKVEDPTAPSATRRKVERQQMSSRSQTQSTRRIKS	EFT
<i>Taiwania_cryptomerioides</i>	QPLAKI	TVSVDP	LKTNDQSVEVKVVDPTAPAIIRKKAEGHETSSTANQRKSTRNNKS	EFT
<i>Sequoiadendron_giganteum</i>	QPLAKI	TVSVDP	LKTNDQSVEVKVVDPTAPAIIRKKAEAHETSSTANHKKSTRNNKS	EFT
<i>Juniperus_scopulorum</i>	QPLAKI	TVSVDP	LKTNDYPADIKVVDPAAPASADKRAEVNETS.SANQRKSKQRNKSE	EFT
<i>Chamaecyparis_lawsoniana</i>	QPLAKI	TVSVDP	LKTNDYPADIKVVDPAAPASANKKAEVNETS.SANQRKPQQRNKS	EFT
<i>Wollemia_nobilis</i>	QPLAKI	TVSVDP	LKTHEEQEI..EVVDPTAPSTMNKKAEVHETSSKVKSQSSRGIKS	EFT
<i>Sundacarpus_amarus</i>	QPLAKI	TVSVDP	LKTHDQOQVEAKVVDPTAPSTTSKKVEVPETSSRVNTSQSSQGAKS	EFT
<i>Dacrycarpus_compactus</i>	QPLAKI	TVSVDP	LKTHDQOQVEAKIVDPTAPSIMNKKVDVSETSSKVNTSQSSQGAKS	EFT
<i>Podocarpus_coriaceus</i>	QPLAKI	TVSVDP	LKTHDQOQVEAKIVDPTAPSTMNKKVDASEMSSKVNTSQSSQGAKS	EFT
<i>Nageia_nagi</i>	QPLAKI	TVSVDP	LKSHDOOQVEAKIVDPTAPSTMNKKVGASETSSKVSTSOSOGAKS	EFT
<i>Amborella_trichopoda</i>	LPLAKV	VVSVDP	LQ.PDSPA..	SLQFT
<i>Nelumbo_nuphar</i>	H <del>P</del> VAKV	VL <del>SV</del> DP	LR.HGDP..	SLQFT
<i>Vitis_vinifera</i>	QPVAKV	VL <del>SL</del> DP	LR.SEDPS..	ALE
<i>Populus_trichocarpa</i>	A <del>P</del> LAKV	VL <del>SL</del> DP	LQ.SDDPS..	ALQFT
<i>Theobroma_cacao</i>	QPVAKV	VL <del>SL</del> DP	RK.PDDPS..	ALQFT
<i>Zostera_marina</i>	PVIS <del>KV</del>	VL <del>TL</del> DP	LCP..DES..	SLQFT
<i>Arabidopsis_thaliana</i>	PDMAK <del>I</del>	VRE <del>EV</del> DP	LR...DDS..	PPE
<i>Eucalyptus_grandis</i>	NPLAKV	VL <del>SL</del> DP	LLPPDNPS..	SLQFT



# “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*

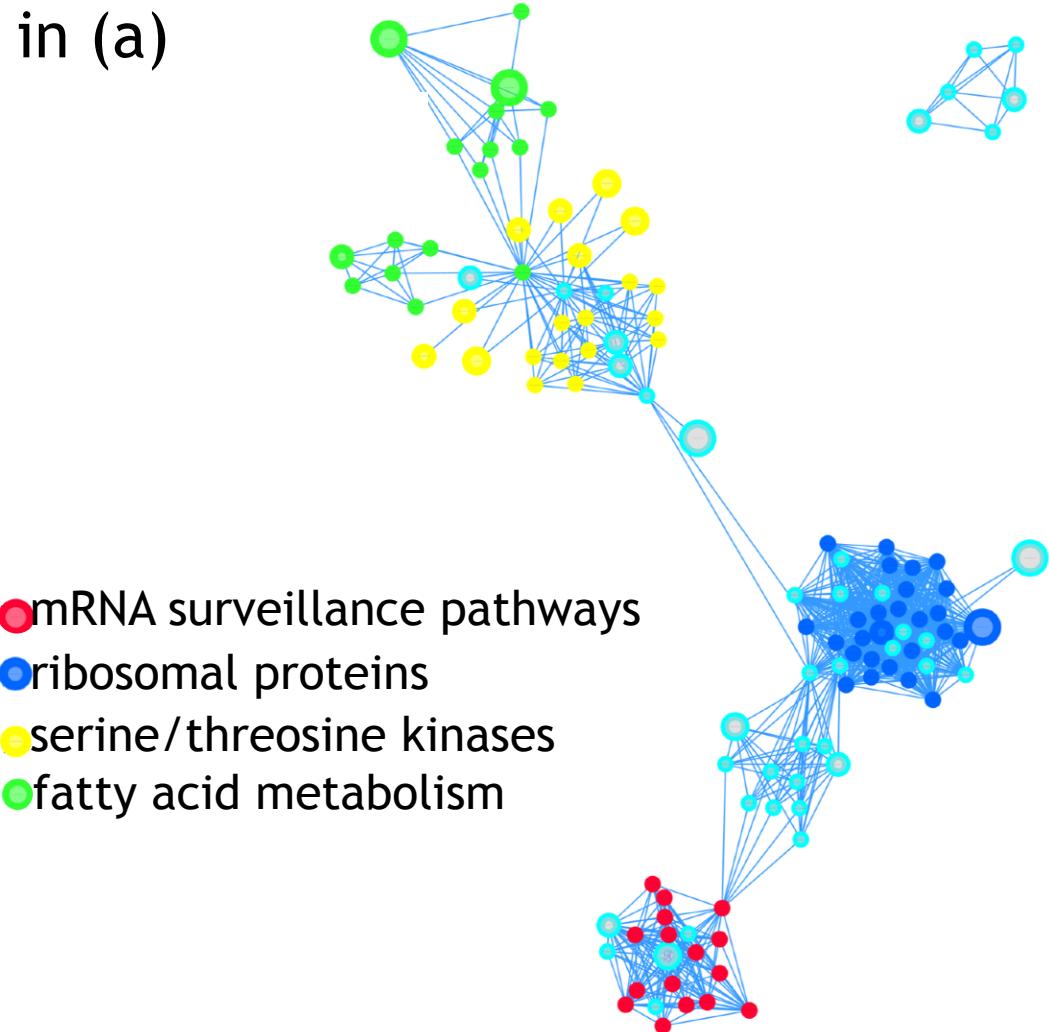


● Intron Length  $\geq$  300 kbps

● 200 kbps  $\leq$  Intron Length < 300 kbps

● 150 kbps  $\leq$  Intron Length < 200 kbps

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



● mRNA surveillance pathways

● ribosomal proteins

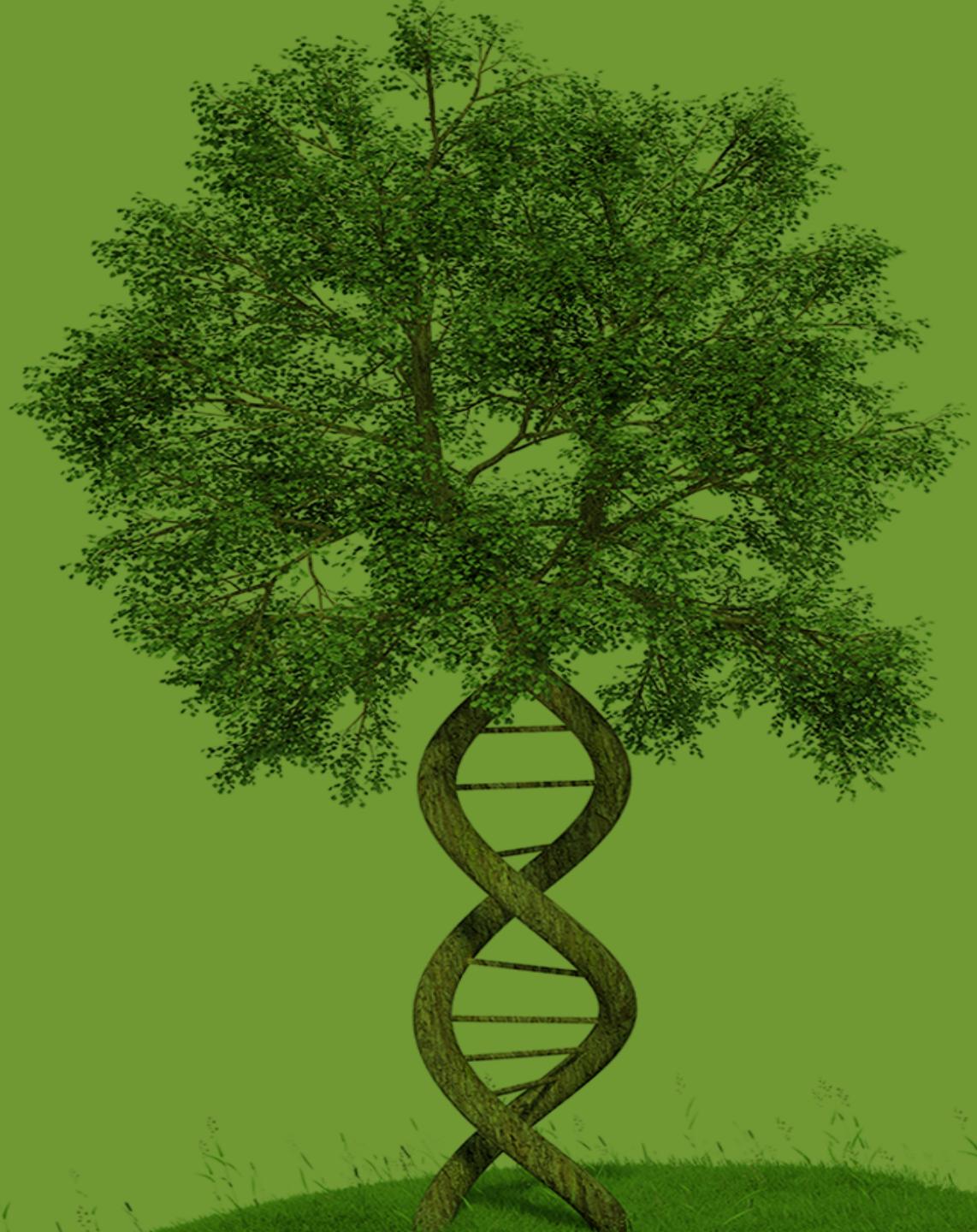
● serine/threonine kinases

● fatty acid metabolism

# 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

