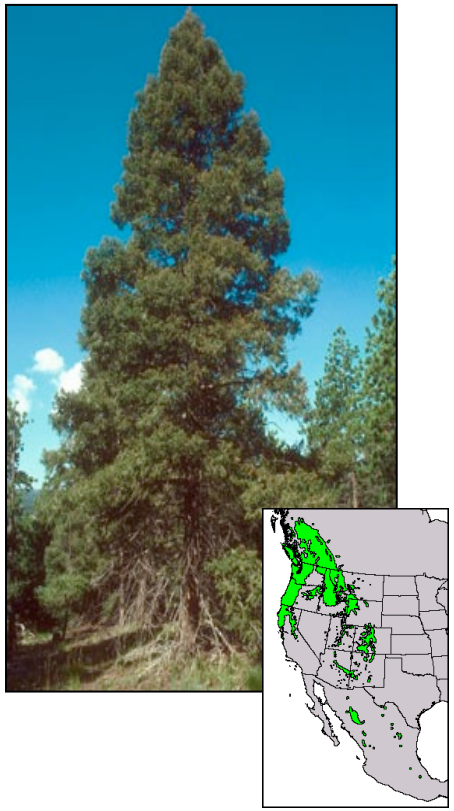
Comparative Evaluation of Gene Annotation Methods in Conifer Megagenomes

Sumaira Zaman

January 13, 2018

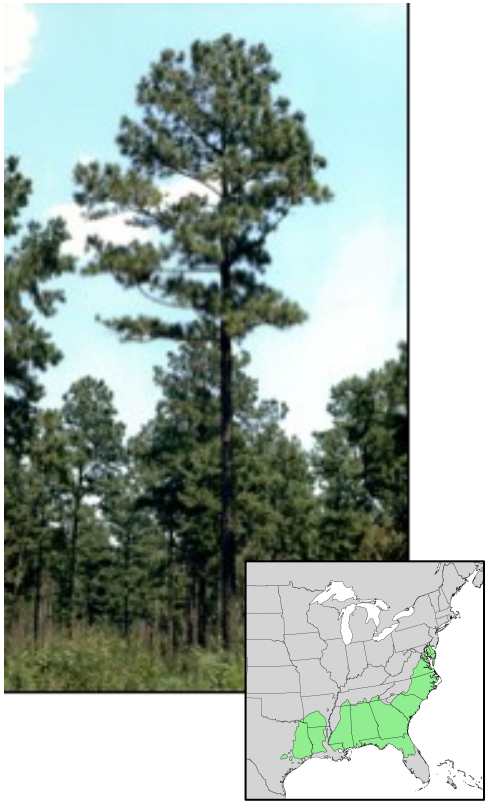
Pine RefSEQ Project

Douglas fir
(*Pseudotsuga menziesii*)



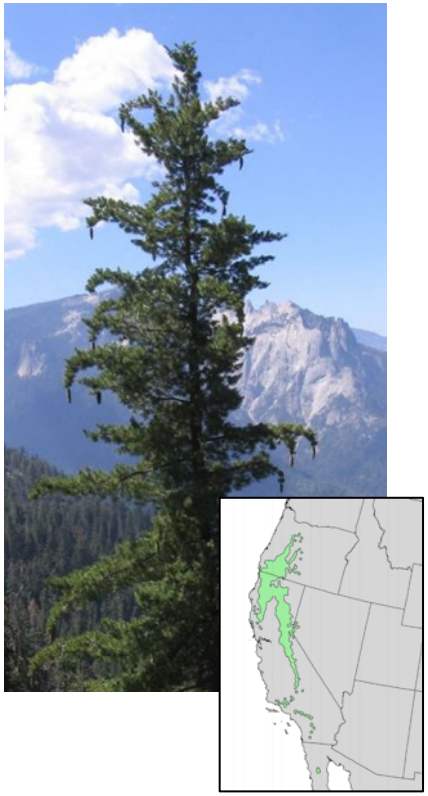
- n=13
- Genome = 18.7 Gb

Loblolly pine
(*Pinus taeda*)

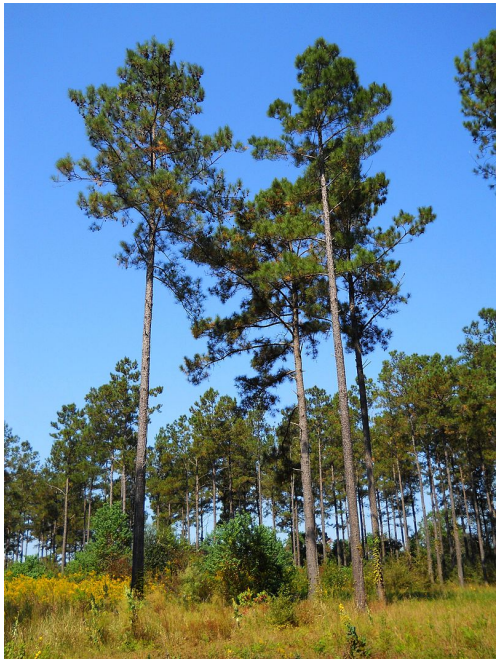


- n=12
- Genome = 22 Gb

Sugar pine
(*Pinus lambertiana*)



- n=12
- Genome = 33.5 Gb

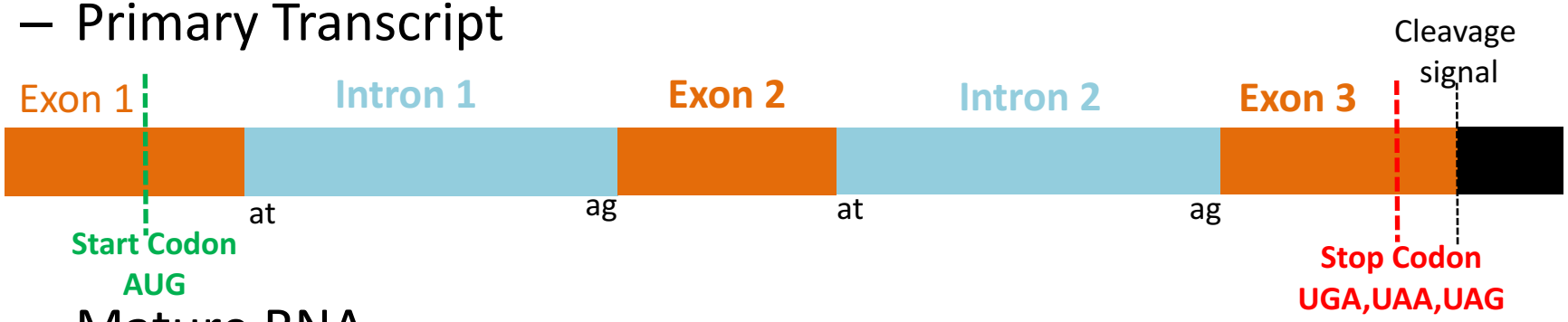


- 18% of world's industrial round wood
- Ecological model for carbon sequestration
- Potential source for renewable energy
- Breeding time: 20 years/generation

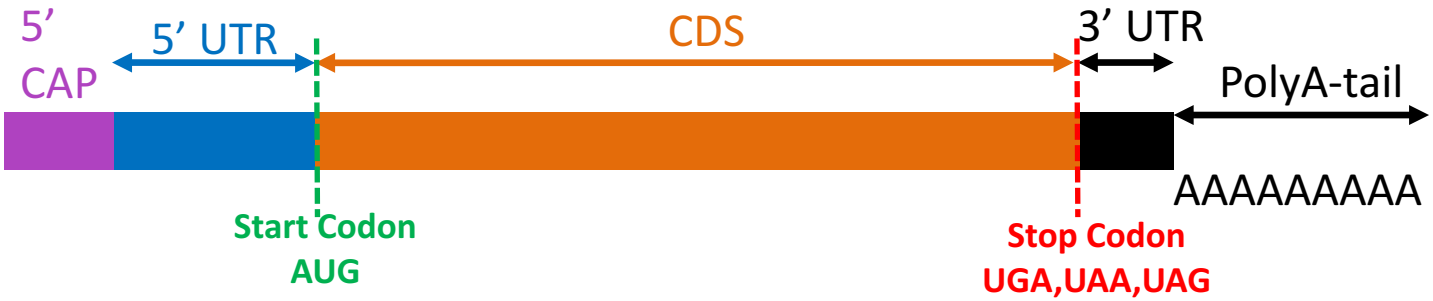
- Fusiform rust
- Consequentially affects loblolly and slash pine.
- Results in losses of millions anually.
- Solution: Discover genes responsible for resistance

Genome Annotation

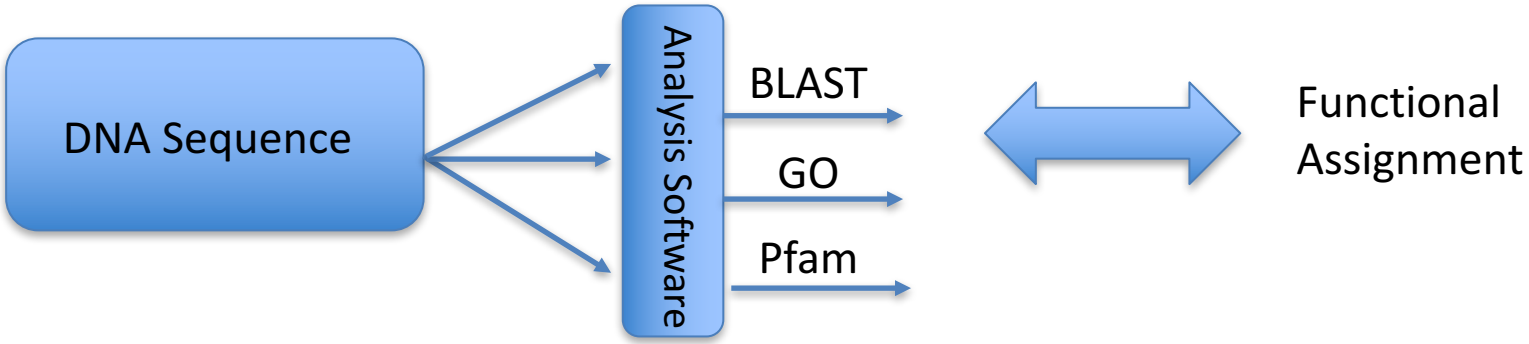
- Structural
  - Primary Transcript



  - Mature RNA



- Functional



CHALLENGES FACING THE GENOME

Size (22 Gbp)

Fragmentation (N50 = 107 kbps, number of scaffolds =)

Repetitive Content (~85-90% of the genome)

Pseudogenes (3-5% of the genome)

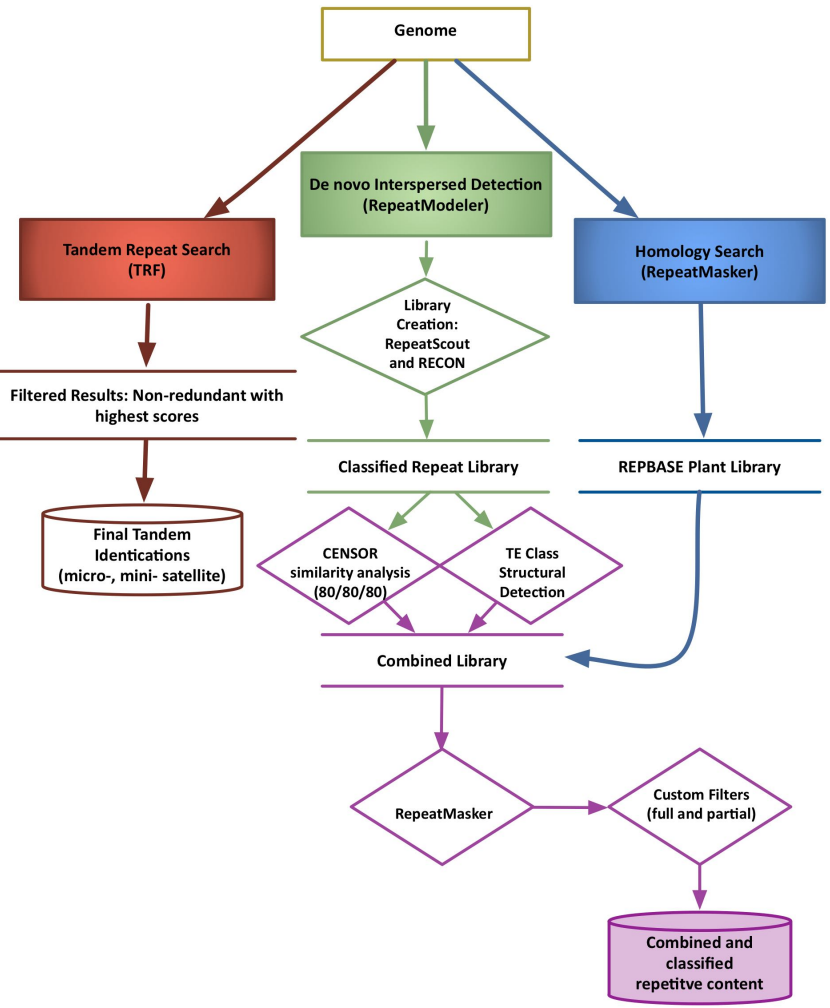
BIOINFORMATICS CHALLENGES

Time consuming (sequence similarity)

Quality of external evidence

Gene prediction (detection of true start sites)

Resolving Repetitive Regions



Creating Repeat Libraries

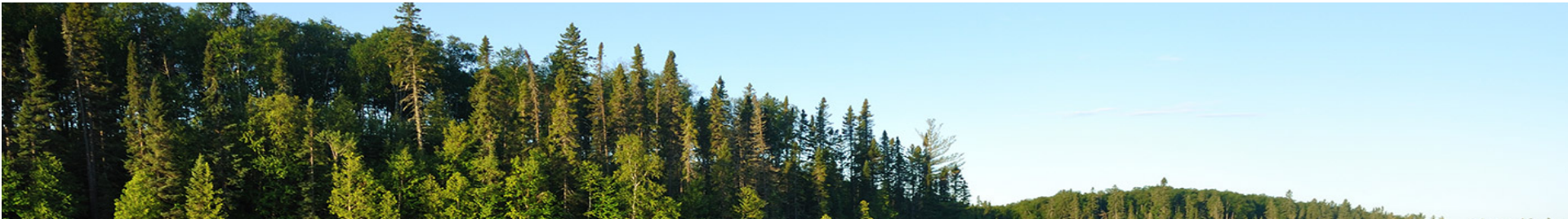
1. Tandem Repeat Identification
   - 2.86% of the genome
2. Homology Based Repeat Identification
   - 62% of the genome dominated by reterotransposons
3. De novo Repeat Identification
   - 1% of the genome yielded 8,155 repeats

Soft Masking the Genome

1. RepeatMasker
   - Generate masked genome, Estimate repeat content

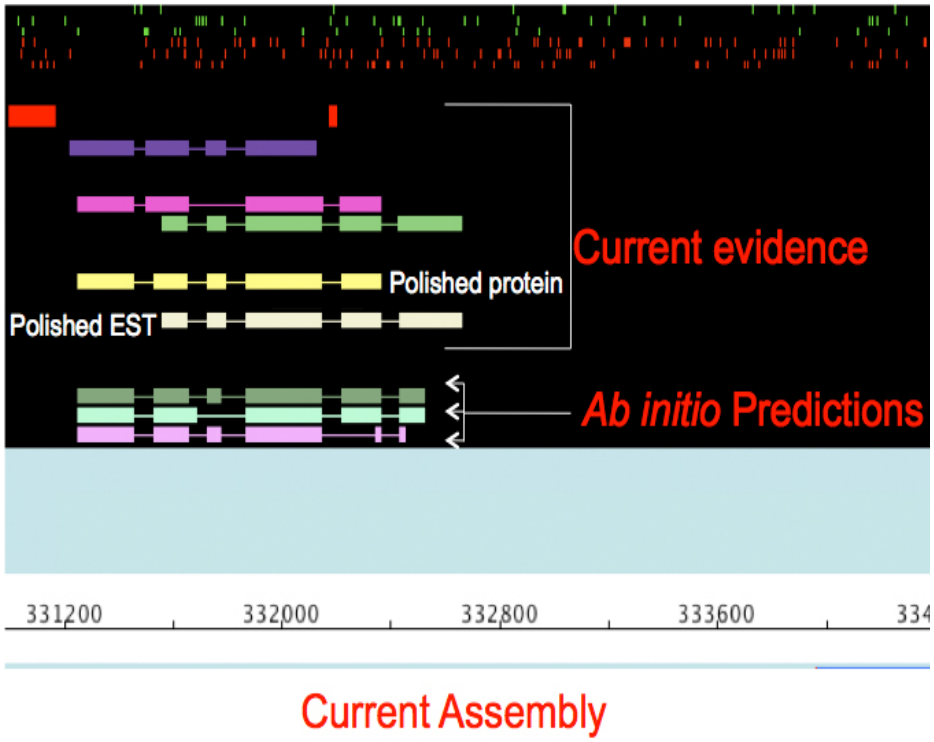
Pinus taeda Sequencing Timeline

Genome Version	*Pinus taeda 1.0*	*Pinus taeda 1.01*	*Pinus taeda 2.0*	*Pinus taeda 2.01*
Tissues and/or Materials Used	1. Haploid Megagametophyte 2. Diploid Needle Tissue	Transcriptome scaffolding	1. Short read data from *P. taeda* 1.0 2. Needle Tissue	1. Haploid Megagametophyte 2. Transcriptome scaffolding
Read Type	Illumina	Illumina (454)	Illumina,Pacbio	Pacbio Isoseq
Assembler	MaSuRCA	-	MaSuRCA, SOAPdenovo	-
Group	UC Davis	Texas A&M, IU	UC Davis	NCSU
N50	30.7 Kbp	66.9 Kbp	107 Kbp	~ >107 Kbp



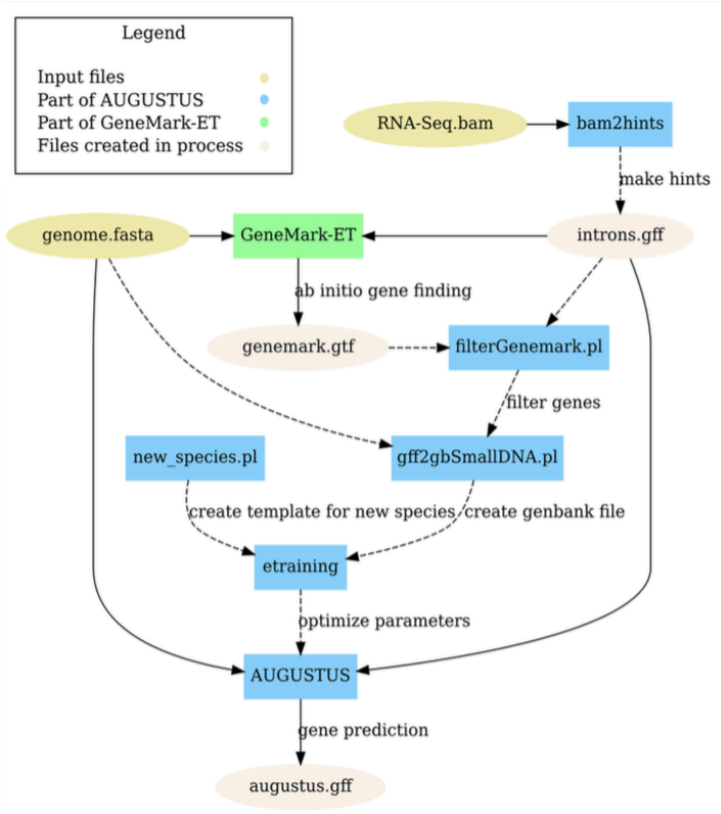
Genome Annotation Pipelines

MAKER



Reference: MAKER GMOD Wiki

BRAKER

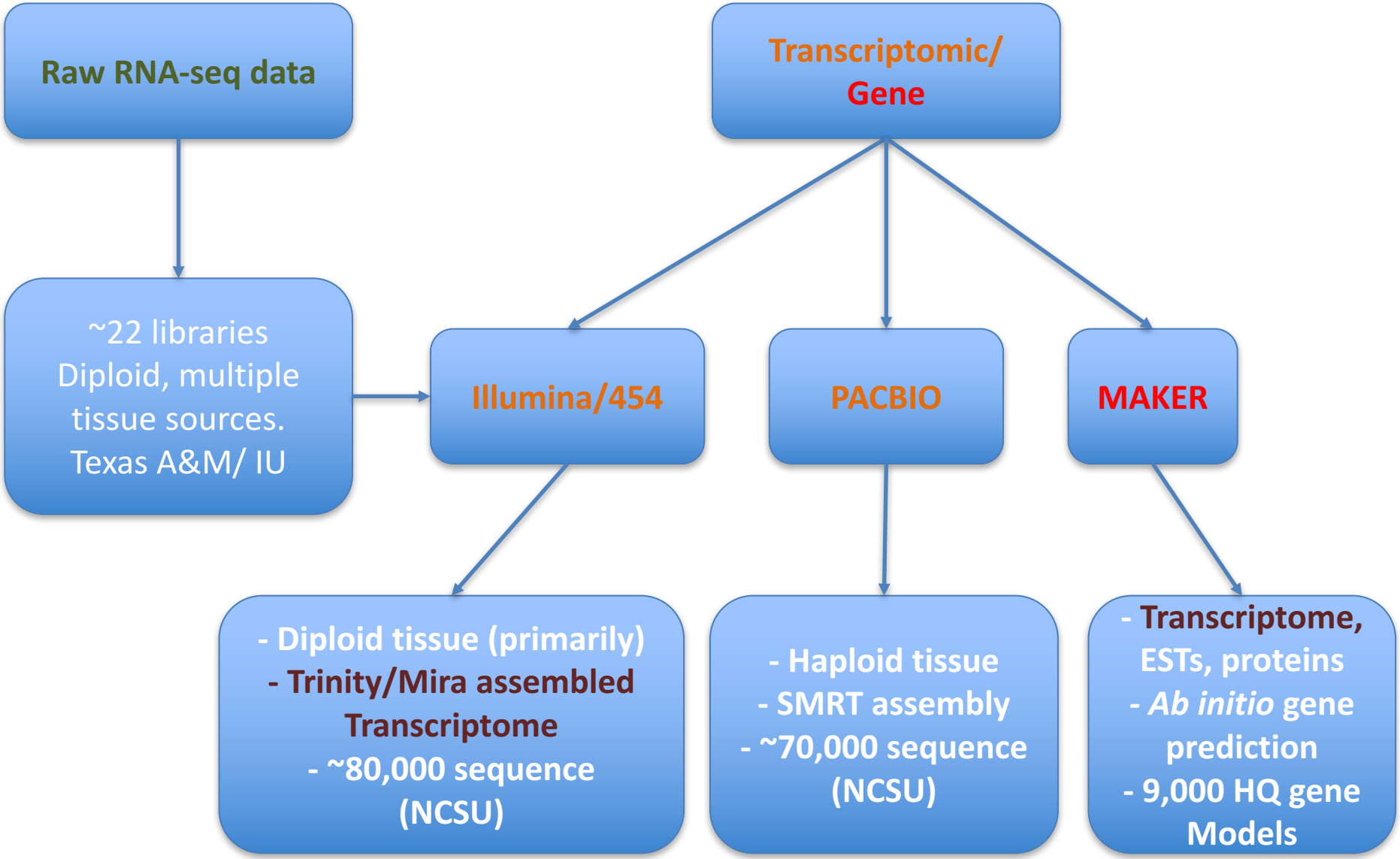


Reference: Katherina Hoff – Universitat Greifswald

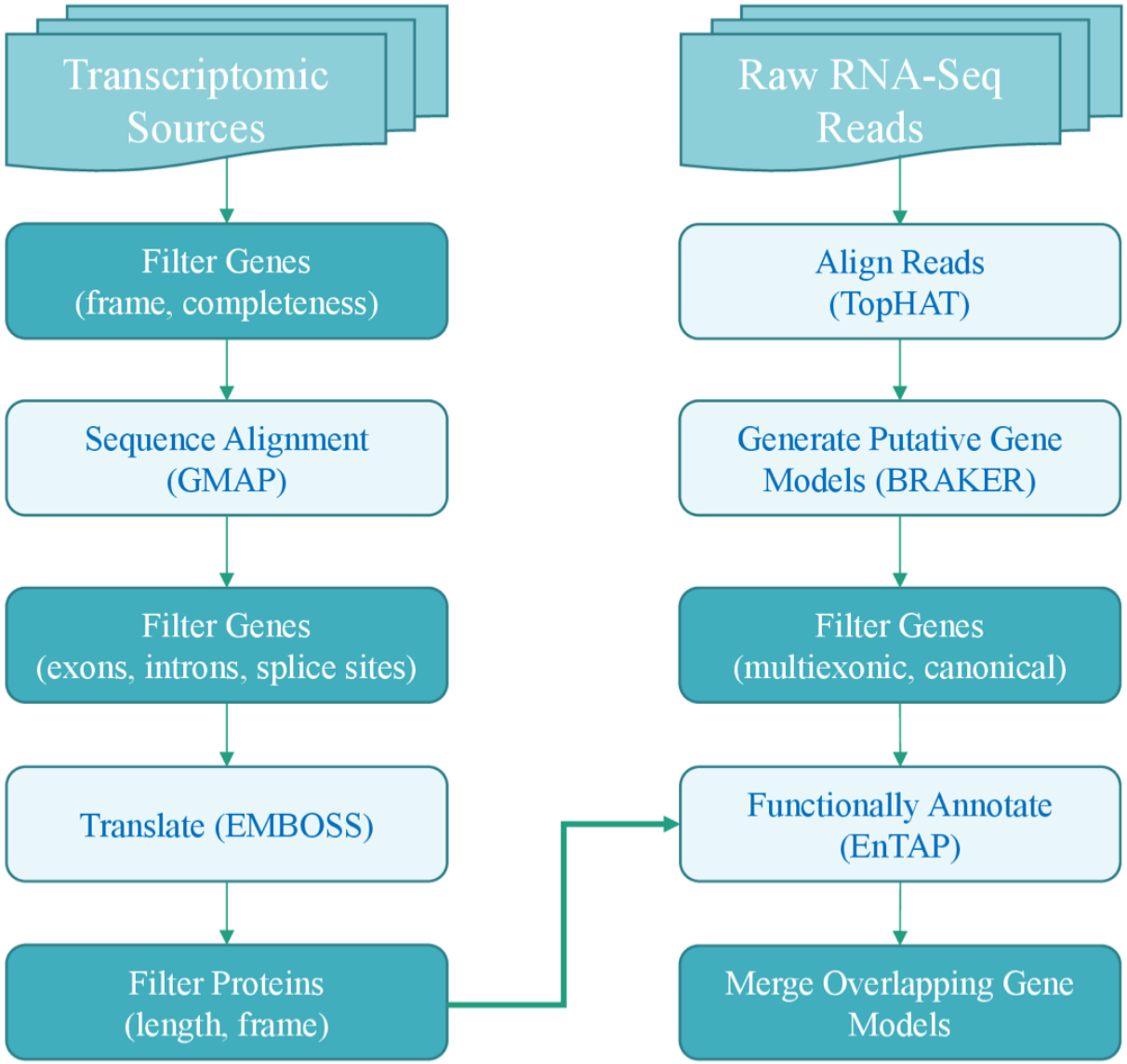
Pinus taeda Sequencing Timeline

Genome Version	*Pinus taeda 1.0*	*Pinus taeda 1.01*	*Pinus taeda 2.0*	*Pinus taeda 2.01*
Tissues and/or Materials Used	1. Haploid Megagametophyte 2. Diploid Needle Tissue	Transcriptome scaffolding	1. Short read data from *P. taeda* 1.0 2. Needle Tissue	1. Haploid Megagametophyte 2. Transcriptome scaffolding
Read Type	Illumina	Illumina, 454	Illumina, Pacbio	Pacbio Isoseq
Assembler	MaSuRCA	-	MaSuRCA, SOAPdenovo	-
Group	UC Davis	Texas A&M, IU	UC Davis	NCSU
N50	30.7 Kbp	66.9 Kbp	107 Kbp	~ >107 Kbp
Gene Annotation	-	**~32,000 MAKER models genes**	-	**?**

Inputs for Annotation of *P. taeda* 2.01



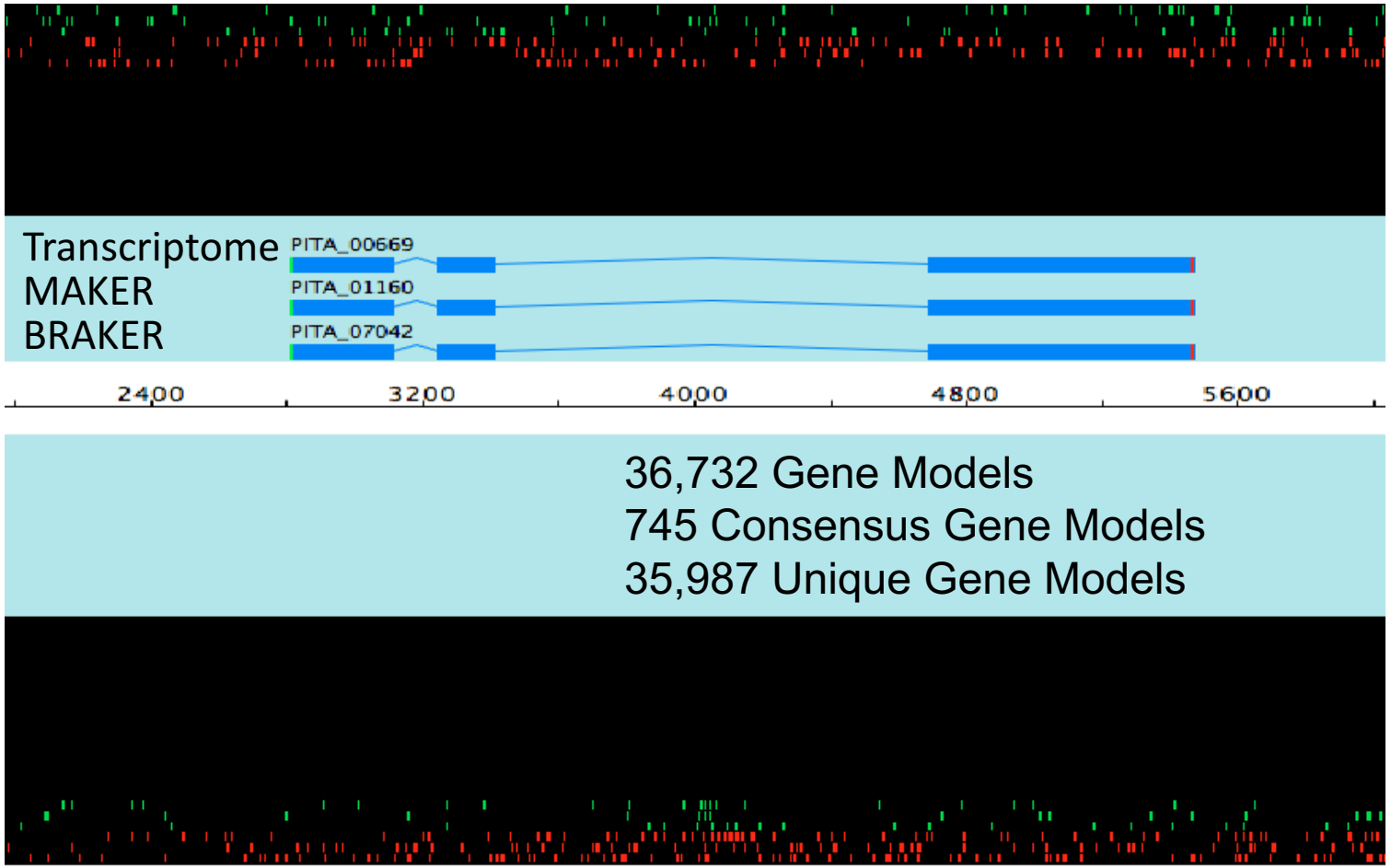
Annotation of *P. taeda* 2.01



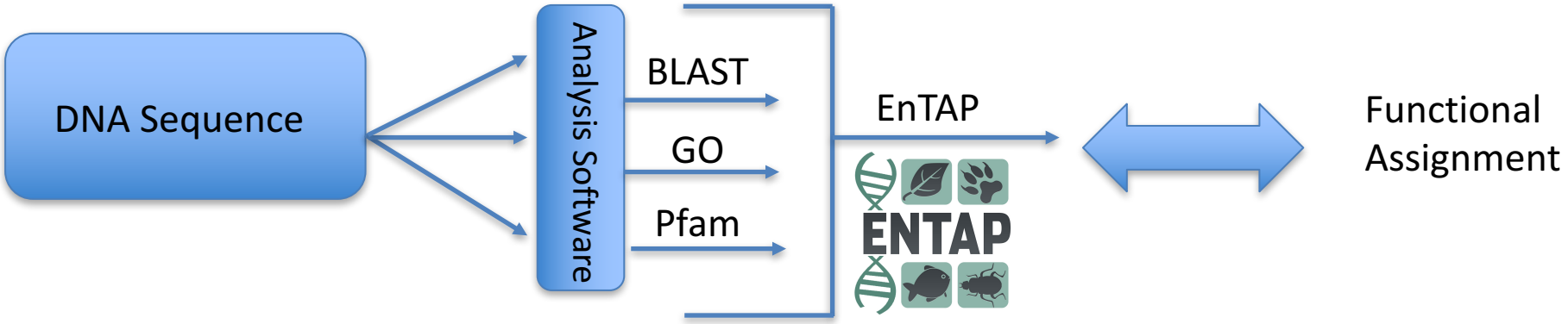
Gene Structure Across Methods

	Number of Genes Discovered	Average Gene Size (bps)	Average Intron Size (bps)	Average Number of Exons	Maximum Intron Size (bps)
BRAKER	35,326	1,790	365	3	25,759
MAKER	388	27,500	312	7	459,077
Illumina sourced transcriptome	430	21,263	4,299	8	461,597
Pacbio reads	586	62,528	3,504	8	568,970

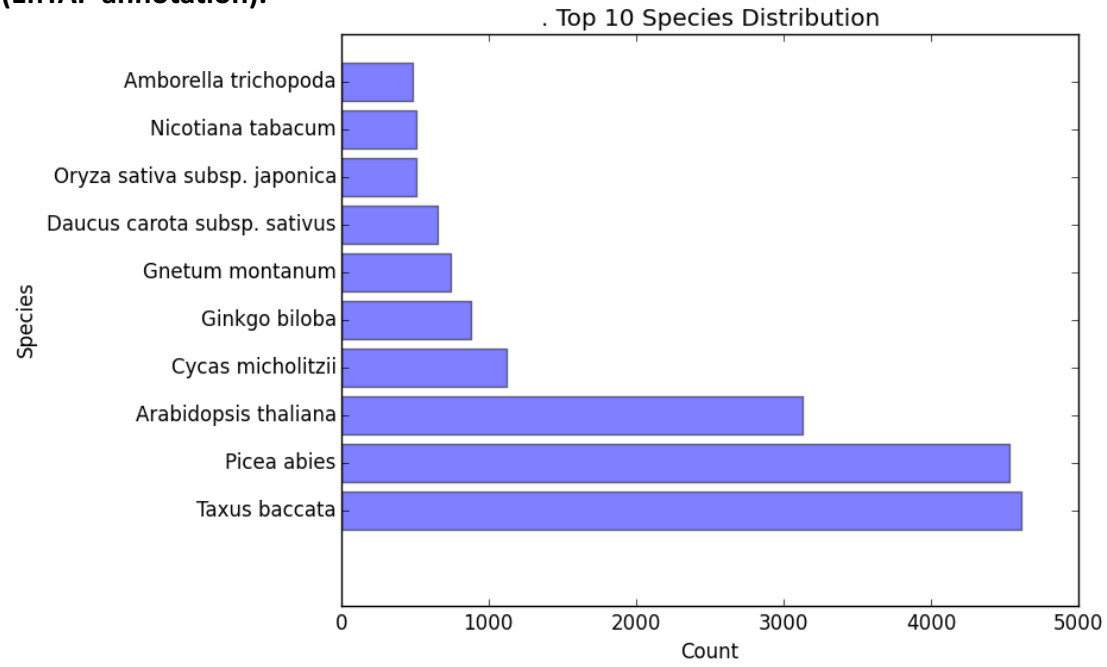
Consensus Across Various Methods



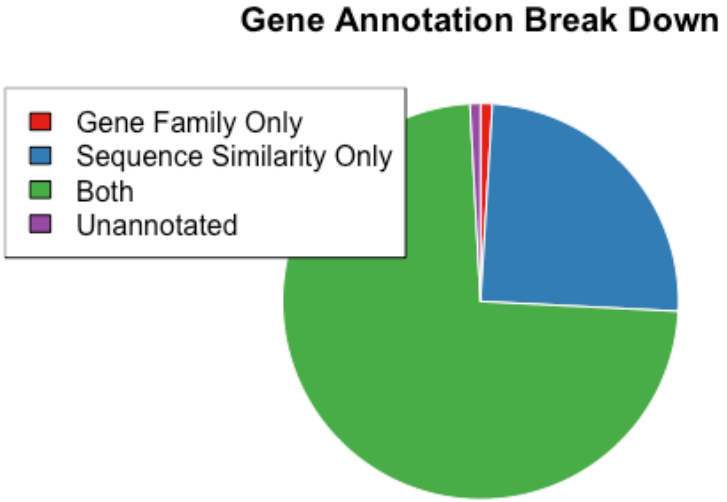
Validation Through Functional Annotation



Total genes annotated by species for the top 10 representatives (EnTAP annotation).



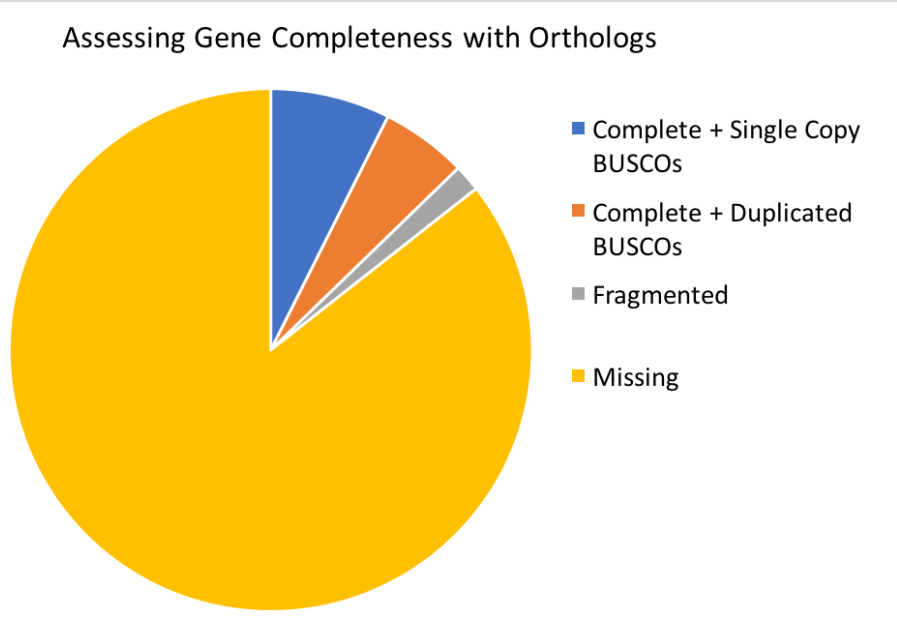
Total genes annotated via sequence similarity and/or gene family assignment via EnTAP.



Assessing Gene Completeness

Using Orthofinder to discover orthogroups that share Orthlogous genes across species.

	BRAKER	Illumina	MAKER	Pacbio
BRAKER	689	303	251	320
Illumina	303	329	105	101
MAKER	251	105	266	41
Pacbio	320	101	41	339



Using BUSCO to discover universal single copy orthologs in embryophyta.

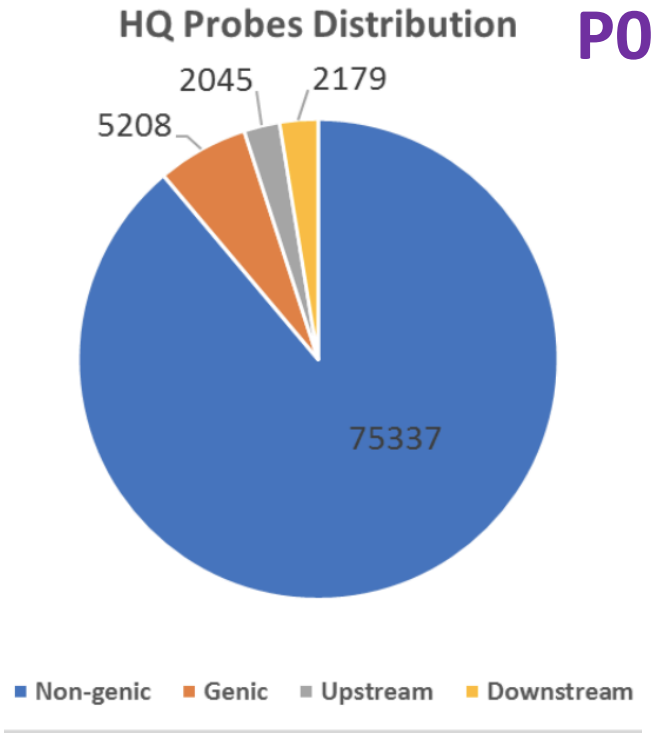
Current & Future Direction

- BRAKER (no long intron models)
  - Merging Gene Models
  - Using Proteomic Evidence
- Annotation of other conifers
  - Sugar Pine (New Genome Assembly)
  - Douglas fir (New RNA data)

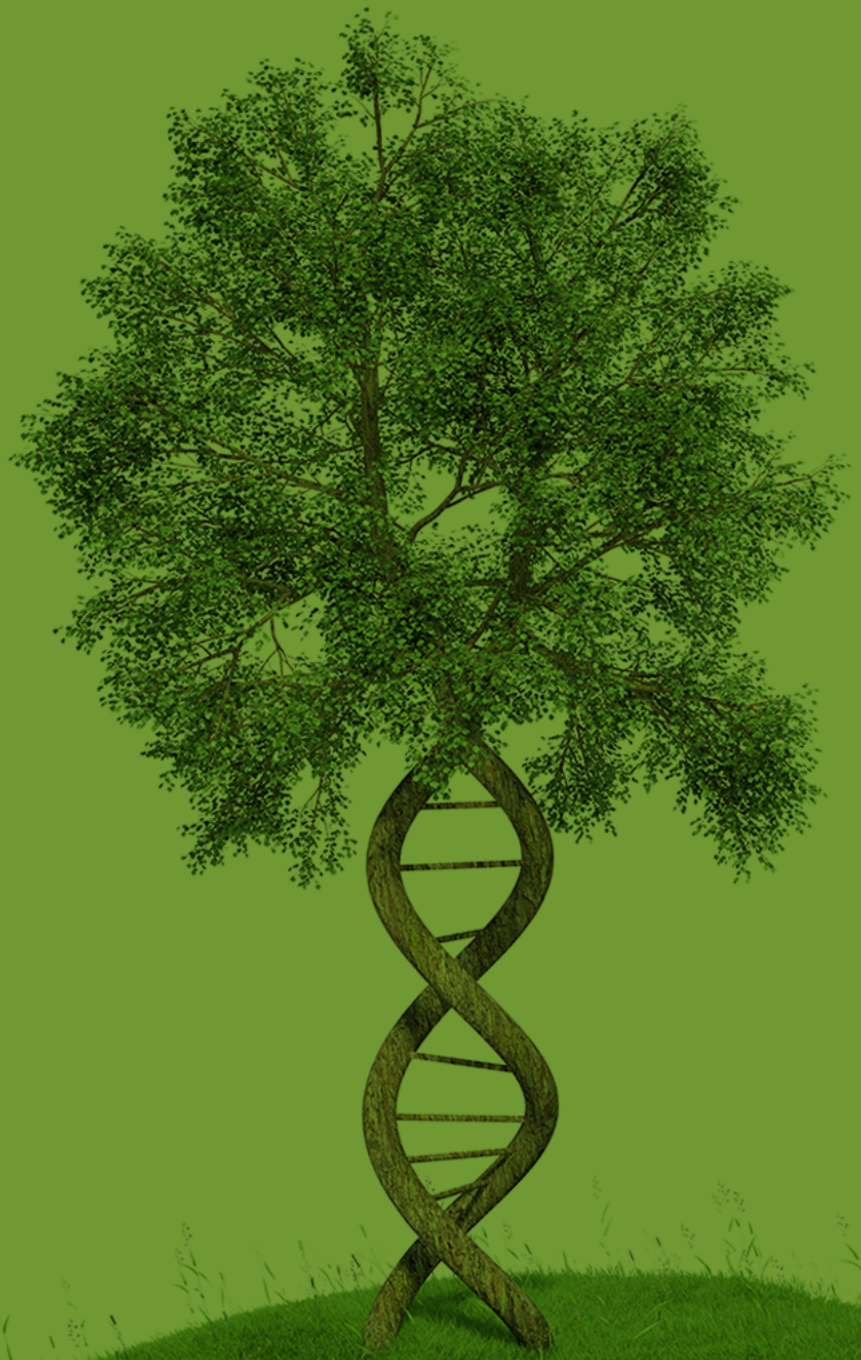
Applications of Current Annotation

Problem:
Discover genes that are responsible for fighting fusiform rust and other desirable phenotypic traits i.e. height.

Solution:
Predict and utilize variants within the Loblolly pine to apply phenotypic selection.



P0617



Madison Caballero UCONN
Jill Wegrzyn UCONN