PineRefSeq

Reference Genome Sequencing Of Conifers

PineRefSeq:

An adaptive approach to the sequencing of large conifer genomes

Nicholas Wheeler: University of California, Davis

University of California, Davis

Children's Hospital Oakland Research Institute

Johns Hopkins University

University of Maryland

Indiana University

Texas A&M University

Washington State University

USDA United States Department of Agriculture National Institute of Food and Agriculture

PineRefSeq

pinegenome.org/pinerefseq

PineRefSeq

Our Guiding Principles

EMPOWERMENT. Our goal is to develop the technologies, platforms and bioinformatics infrastructures required to rapidly and inexpensively sequence large and complex genomes of coniferous forest trees. This will allow the forestry community to begin sequencing the many genomes of economic and ecological importance without a dependence on centralized genome centers.

ADAPTIVE. We recognize that sequencing technologies are developing rapidly and that we must have the expertise and flexibility to rapidly adopt new approaches into our overall sequencing strategy.

COMPARATIVE. We recognize the power of comparative genomics approaches in assembling and annotating genome sequences and will use this approach throughout the project.

OPEN ACCESS. We have a policy of sharing all data generated from this project with the research community

PineRefSeq

Reference Genome Sequence:

For any given organism (species), the complete and ordered “assembly” of DNA, as denoted by the nucleotides A, T, C, and G.

Genome Sequencing – A Short History

A 5-year plan (FY 1991 to 1995) detailing the goals of the U.S. Human Genome Project was presented to members of congressional appropriations committees in mid-February, 1990.

According to the document, "a centrally coordinated project, focused on specific objectives, is believed to be the most efficient and least expensive way" to obtain the 3-billion-bp map of the human genome. In the course of the project, especially in the early years, the plan states that "much new technology will be developed that will facilitate biomedical and a broad range of biological research, bring down the cost of many [mapping and sequencing] experiments, and find application in numerous other fields."

Human Genome News, May 1990; 2(1) Five-Year Plan Goes to Capitol Hill

James Watson

Source: WikiPedia

Francis Collins

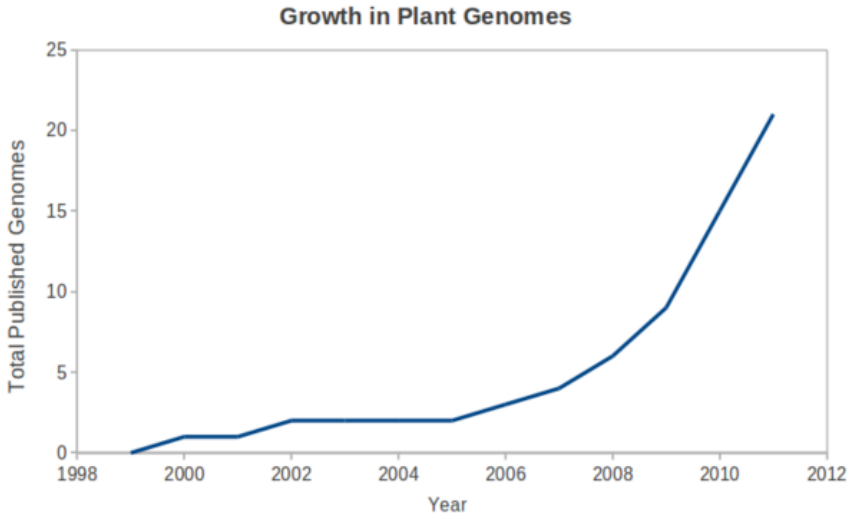
Source: Michigan St University

Craig Venter

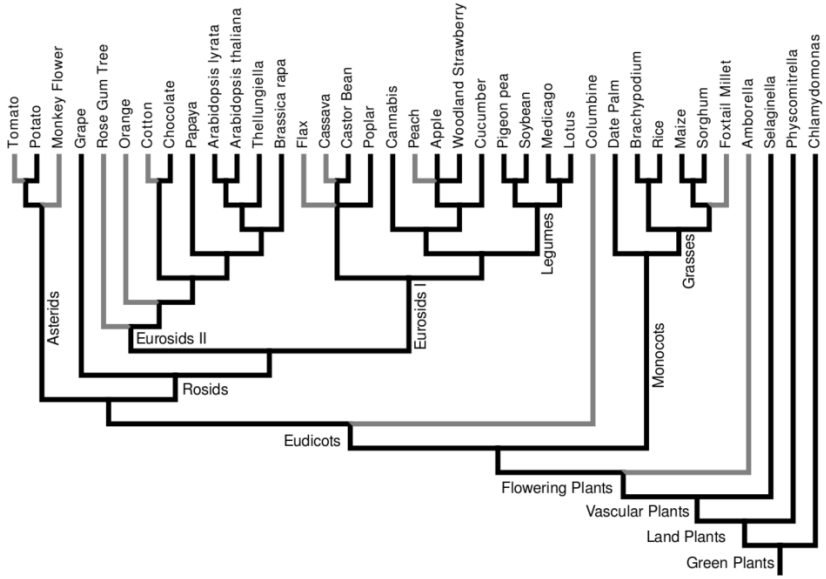
Source: WikiPedia

Genome Sequencing – A Short History (continued)

The rate of publication of plant genomes, updated in late 2011



A phylogenetic tree of all plants with published full genomes as of May 13, 2012



Figures courtesy of the CoGePedia and Phytozome.org websites.

PineRefSeq

Existing & Planned Angiosperm Tree Genome Sequences

As of mid-2012

Species		Genome Size[1] (Mbp)	# of Genes[2]	Status[3]
In Progress with Draft Assemblies				
Populus trichocarpa	Black Cottonwood	500	~40,000	2.0 / 2.2
Eucalyptus grandis	Rose Gum	691	~36,000	1.0 / 1.1
Malus domestica	Apple	881	~26,000	1.0 / 1.0
Prunus persica	Peach	227	~28,000	1.0 / 1.0
Citrus sinensis	Sweet Orange	319	~25,000	1.0 / 1.0
Carica papaya	Papaya	372	-	
Amborella trichopoda	Amborella	870	-	
Betula nana	Dwarf Birch	450	-	1.0 / -
In Progress or Planned – No Published Assemblies				
Castanea mollissima	Chinese Chestnut	800	-	
Salix purpurea	Purple Willow	327	-	
Quercus robur	Pedunculate Oak	740	-	
Populus spp. and ecotypes	Various	Various	-	
Azadirachta indica	Neem	384	-	

1 Genome size: Approximate total size, not completely assembled.
2 Number of Genes: Approximate number of loci containing protein coding sequence.
3 Status: Assembly / Annotation versions

PineRefSeq

Existing and Planned Gymnosperm Tree Genome Sequences

As of mid-2012

Species		Genome Size[1] (Mbp)	# of Genes[2]	Status[3]
Gymnosperms				
Picea abies	Norway Spruce	20,000	?	Pending
Picea glauca	White Spruce	22,000	?	Pending
Pinus taeda	Loblolly Pine	24,000	?	Pending
Pinus lambertiana	Sugar Pine	33,500	?	Pending
Pseudotsuga menziesii	Douglas-fir	18,700	?	Pending
Larix sibirica	Siberian Larch	12,030	?	Pending
Pinus pinaster	Maritime Pine	23,810	?	Pending
Pinus sylvestris	Scots Pine	23,000	?	Pending

1 Genome size: Approximate total size, not completely assembled.
2 Number of Genes: Approximate number of loci containing protein coding sequence.
3 Status: Assembly / Annotation versions; See http://www.phytozome.net for all publicly released tree genomes. Conifer genomes will also be posted here as they are completed.

Technological Advances Facilitate Sequence Acquisition

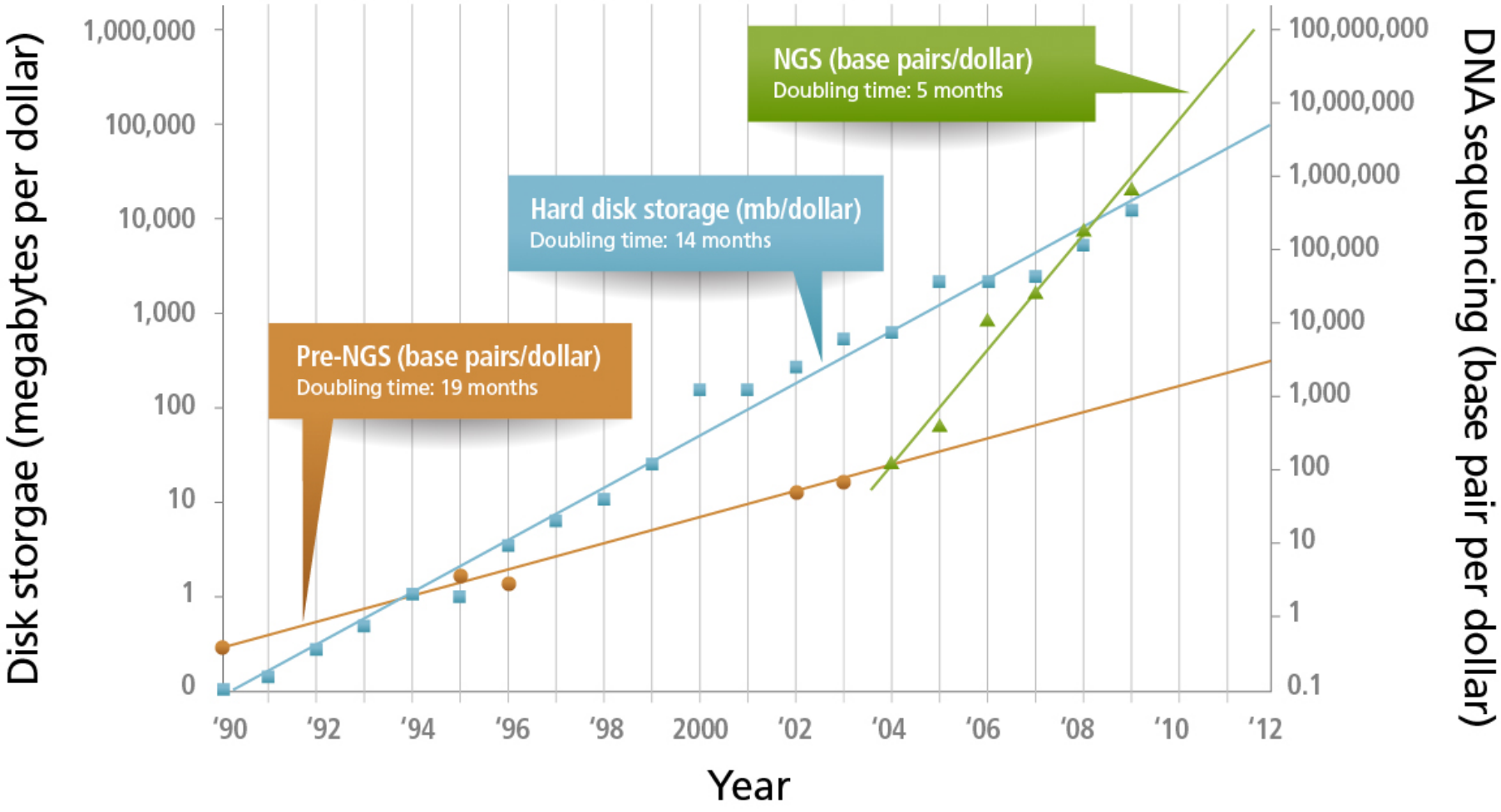


Figure credit: modified from Jill Wegrzyn, UC Davis

PineRefSeq

Why Do We Need a Conifer Genome Sequence?

Fundamental Genetic Information

Phylogenetic Representation

Ecological Representation

Development of Genomic Technologies

Economic Importance

Photo credit: Nicholas Wheeler, UC Davis

Challenges to Sequencing a Conifer Genome

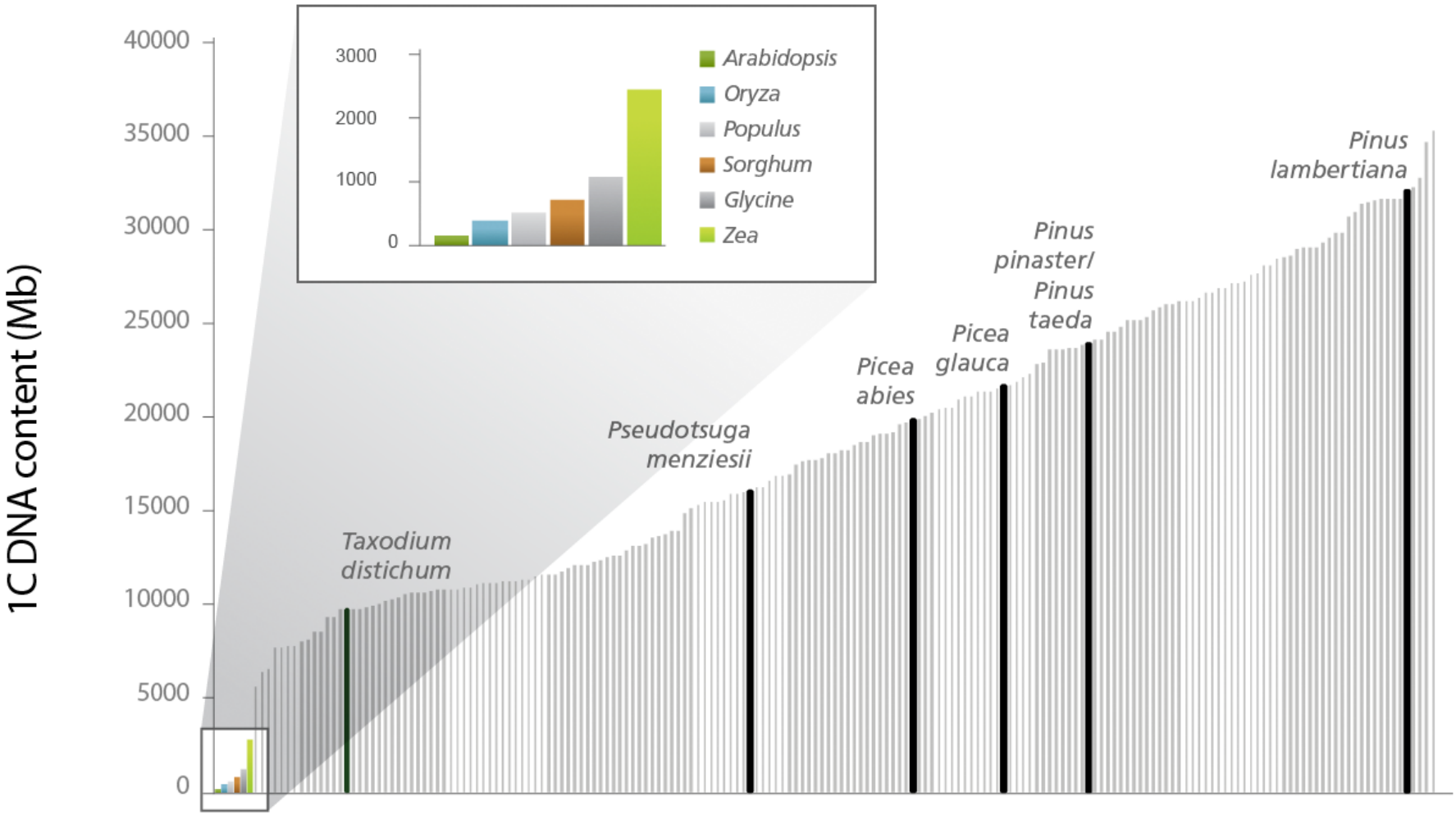


Image Credit: Modified from Daniel Peterson, Mississippi State University

PineRefSeq

Elements of a Conifer Genome Sequencing Project

Approaches to Resolving Challenges

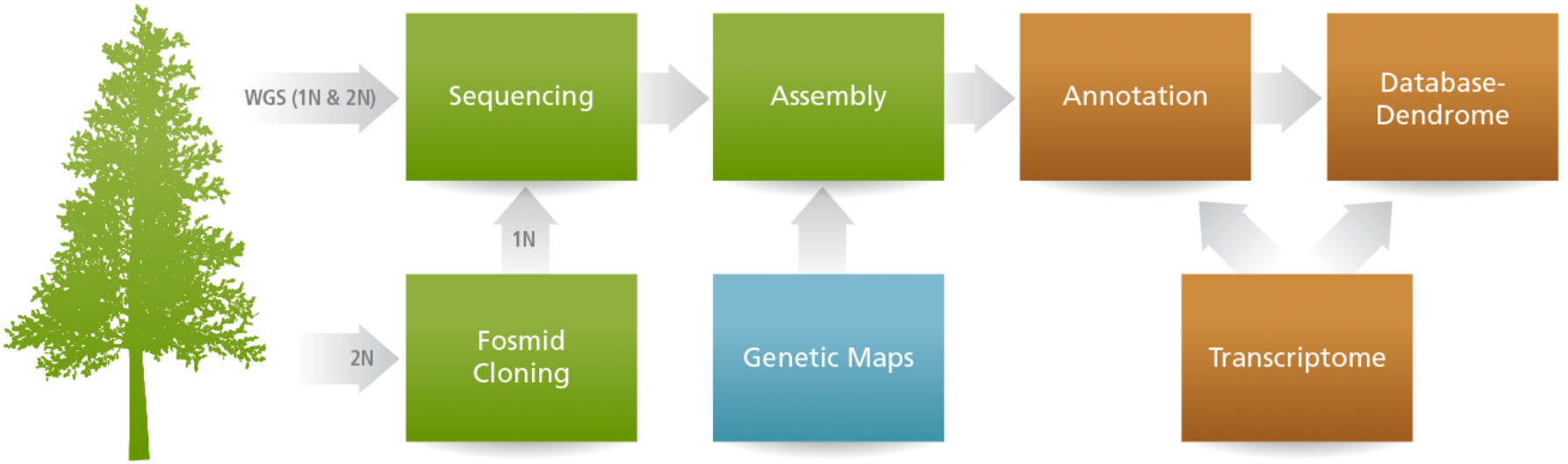


Figure Credit: Nicholas Wheeler, University of California, Davis

PineRefSeq

Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing

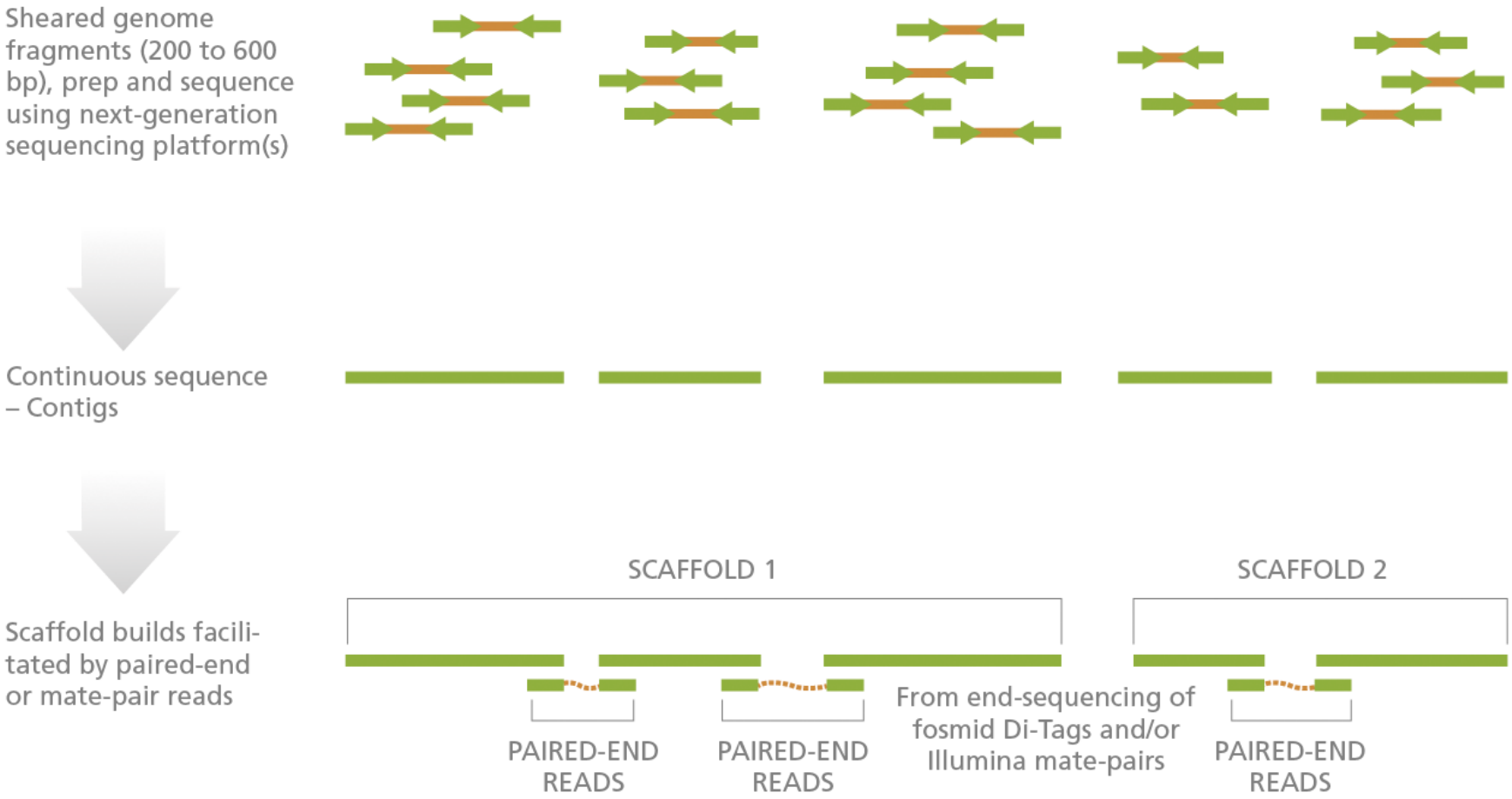


Figure Credit: Nicholas Wheeler, University of California, Davis

Acquiring the Sequence

Target Genome, Appropriate Tissues for DNA & RNA

Haploid

Haploid megagametophyte tissue
1N
Shotgun sequenced

Diploid

Diploid needle tissue
2N
40 Kb cloned fosmids, pooled and sequenced

Figure Credit: Nicholas Wheeler, University of California, Davis

Sequencing Strategy

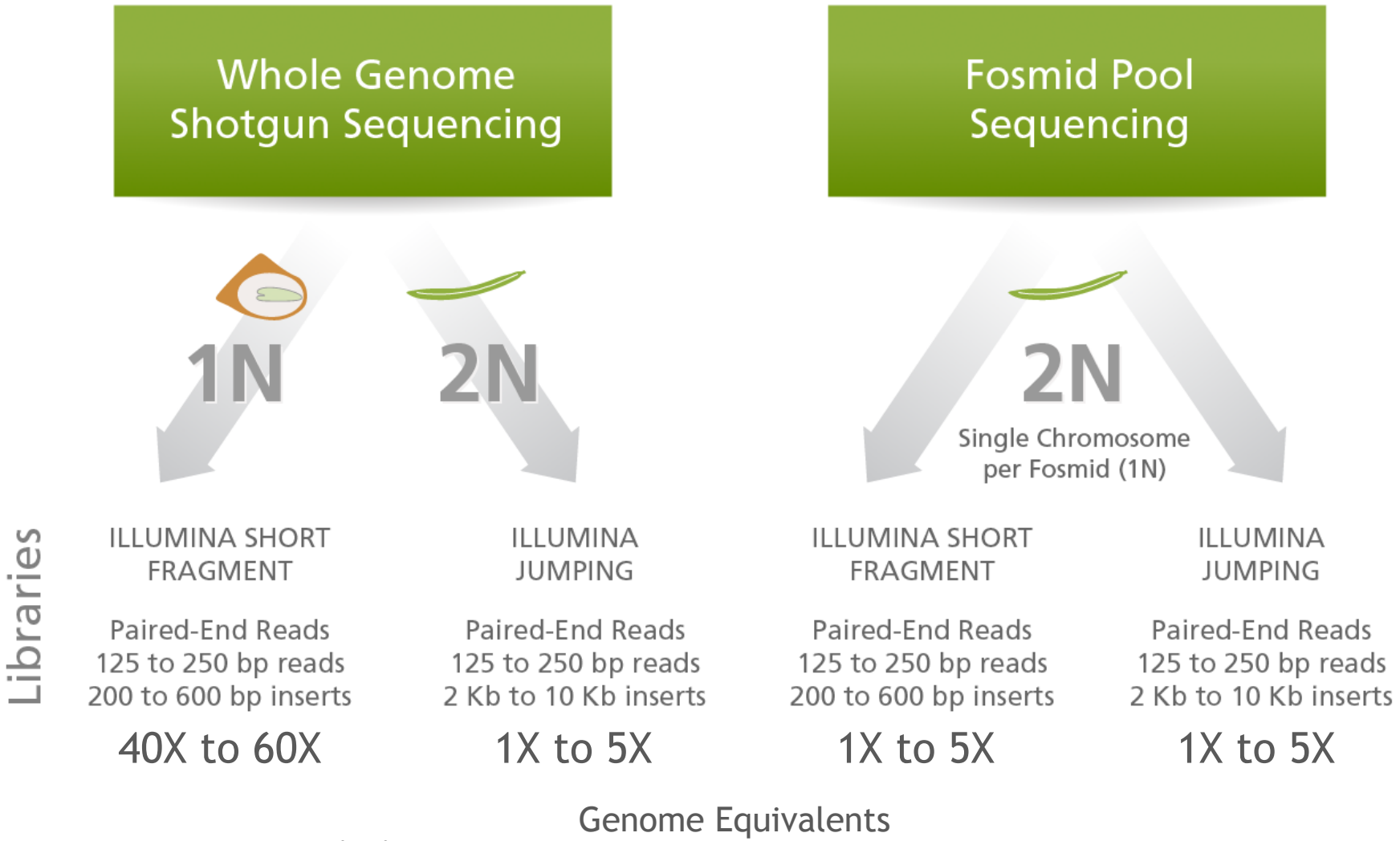
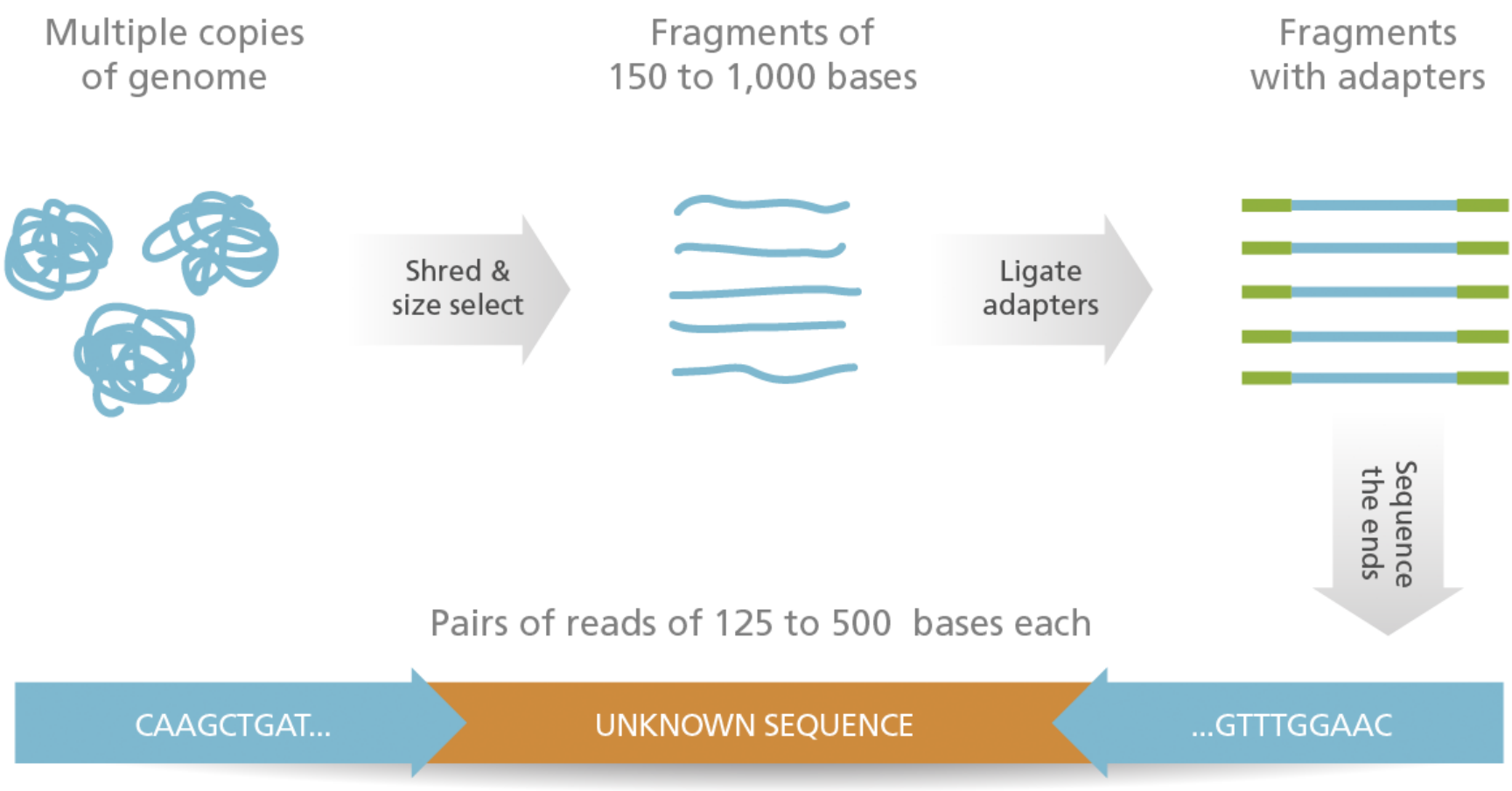


Figure Credit: Nicholas Wheeler, University of California, Davis

PineRefSeq

Whole Genome Shotgun Sequencing

Millions of Short "Reads"



Visit the Broad Institute for details on DNA preparation, library construction, and sequencing technology of Illumina HiSeq

Figure Credit: Nicholas Wheeler, University of California, Davis

PineRefSeq

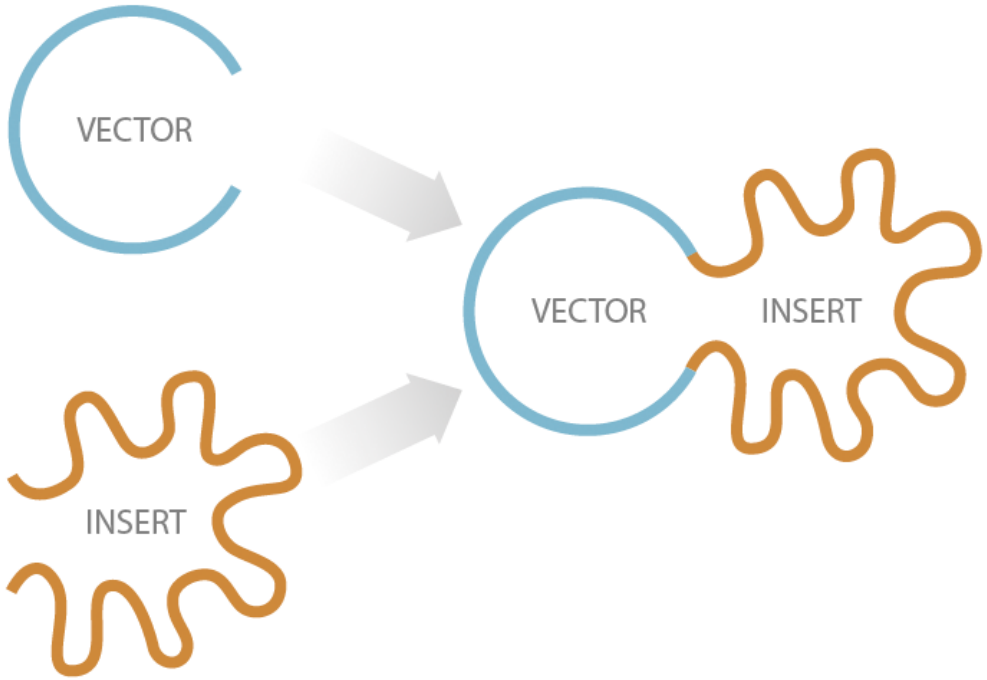
Sequencing DNA from Pools of Fosmid Clones

Fosmid Library Construction

Genomic DNA cloned into fosmid vectors represents a source of stable genomic fragments of approximately 30 to 40 kb.

Cloning vector linearized with blunt end cutter and dephosphorylated to prevent self ligation.

DNA is sheared and ends blunted. DNA molecules are separated according to size and only a fraction that can form fosmids is excised from gel.



Genomic DNA and vector are ligated together.

Fosmids are packaged into Lambda phage particles for tighter size selection and trans-fected into appropriate bacterial cells for propagation.

Figure Credit: Modified from Maxim Koriabine, Children's Hospital Oakland Research Institute

PineRefSeq

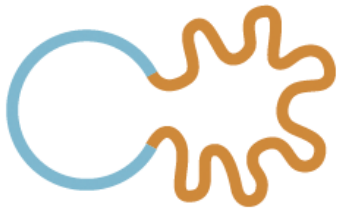
Sequencing DNA from Pools of Fosmid Clones

Preparing Fosmids for Sequencing

Assembly of complex genomes with a high level of repetitive DNA is facilitated by reducing the complexity of the “puzzle”.

Fosmid library is plated at a density of approx. 1000 colonies per petri dish.

Fosmid DNA is prepared from each pool.



DNA is sheared and used for next-generation sequencing.

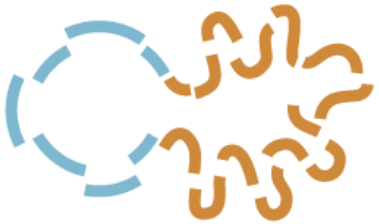
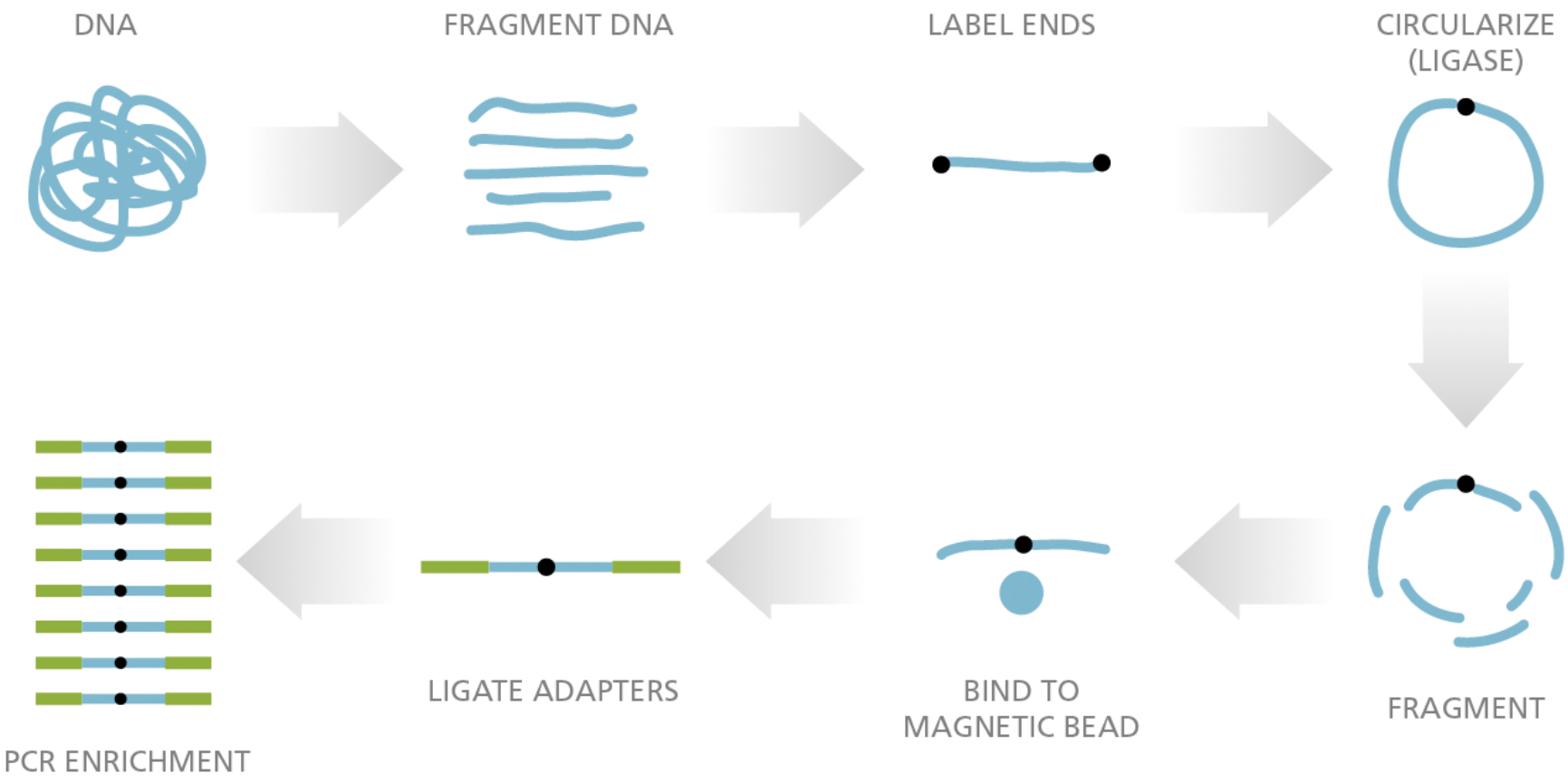


Figure Credit: Modified from Maxim Koriabine, Children’s Hospital Oakland Research Institute

PineRefSeq

Jumping Libraries

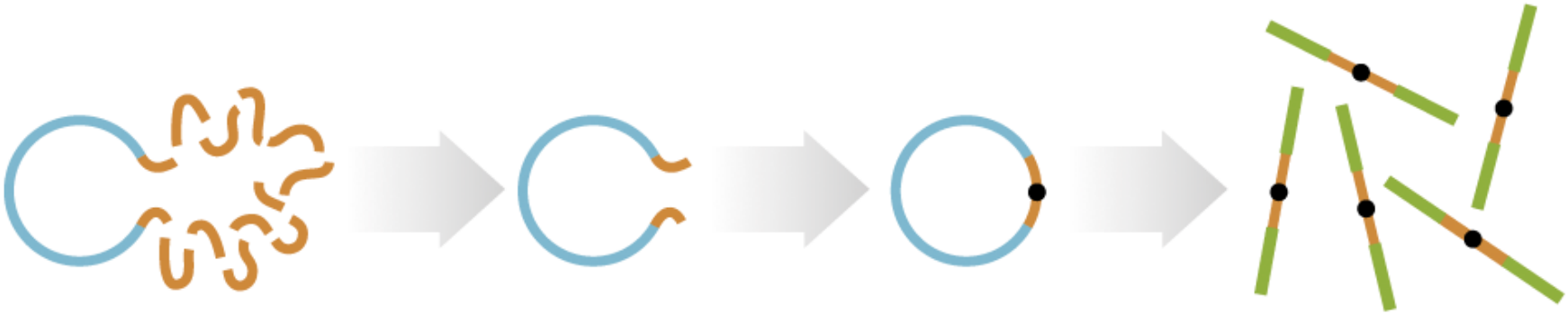
Illumina Mate-Pair, Clone-Free



PineRefSeq

Jumping Libraries

Fosmid Di-Tag Cloned



Most of the insert is removed, only short terminal portions directly ajacent to the vector remain.

Vector is recircular-ized, DNA fragments that were ~40 kb apart in a genome are now brought together.

Remaining insert DNA is PCR ampli-fied, adapters for the Next-Generation sequencing are attached

Figure Credit: Modified from Maxim Koriabine, Children's Hospital Oakland Research Institute

PineRefSeq

Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing

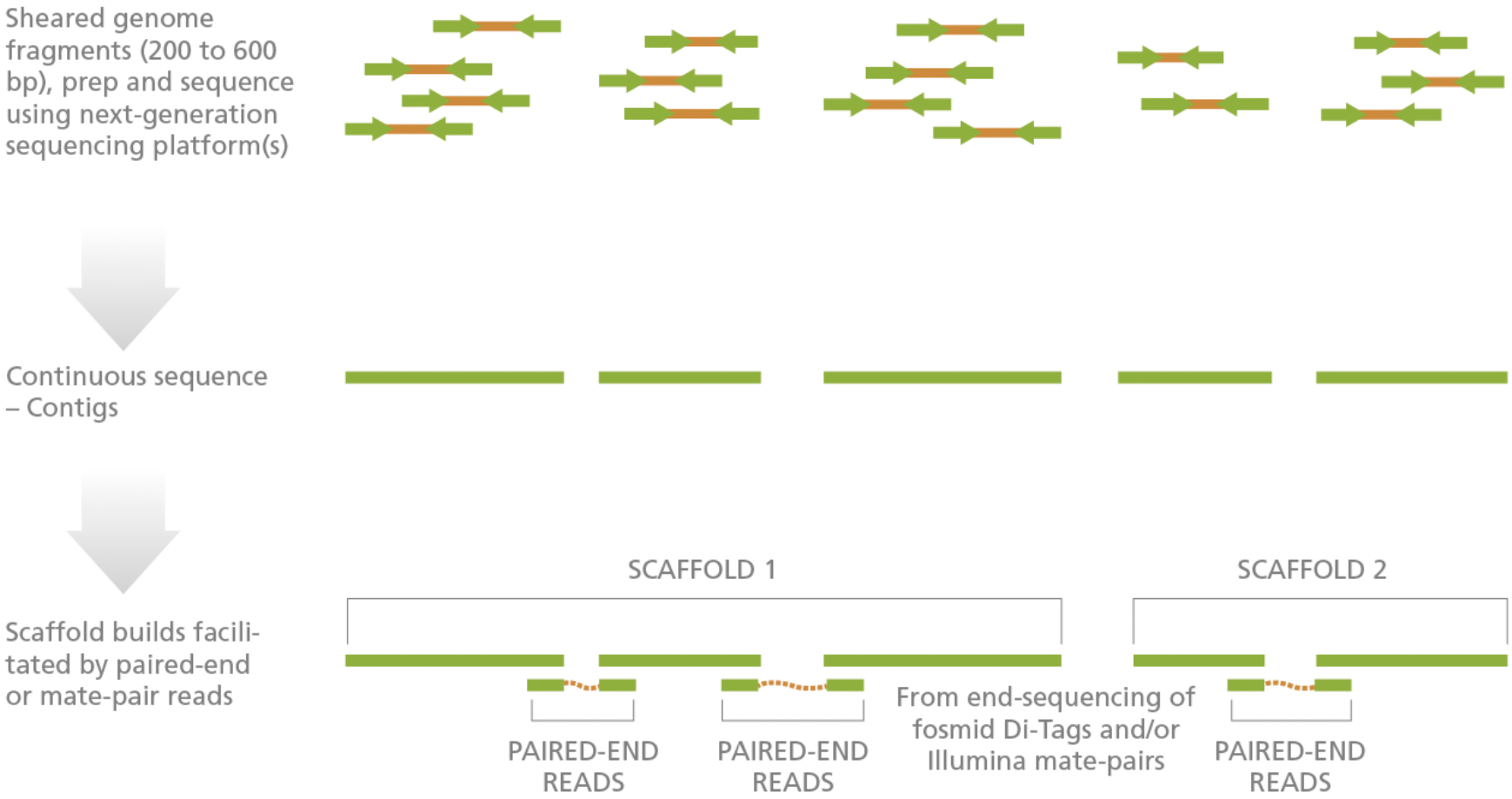


Figure Credit: Nicholas Wheeler, University of California, Davis

PineRefSeq

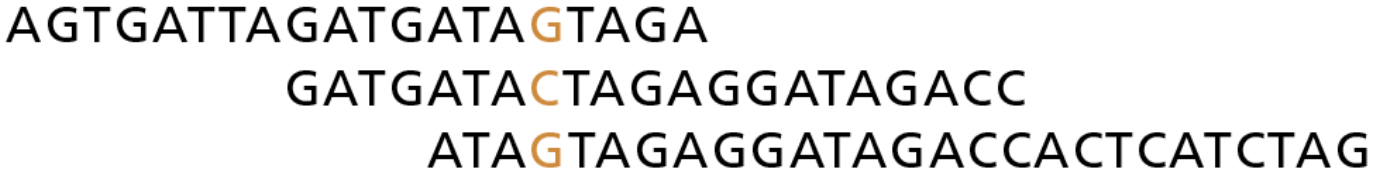
Assembling the Reference Sequence

The Essence of Assembly

This general approach is called OLC or Overlap-Layout-Consensus.

Find pieces that fit together: Compute overlaps of reads

Connect the pieces: Create layout of numerous overlapping reads



Create consensus sequence of contiguous nucleotides (i.e., contigs)

Figure Credit: Nicholas Wheeler, University of California, Davis

De Bruijn Graph Assembly Approach

Find all k-mers (short DNA sequences of length k) and build a graph.

- Every k-mer is a node
- Two nodes are linked with an edge if they share k-1 nucleotides

Figure Credit: Modified from Aleksey Zimin

PineRefSeq

Comparing OLC and Graph Assembly Approaches

OLC

Benefits

- Can deal with variable length reads and reads from different sequencing platforms
- Overlaps can be long and thus more reliable
- Overlaps do not have to be exact
- Can resolve repeats of up to read size

Drawbacks

- Computationally intensive, number of overlaps grows quadratically with the number of reads

Graph

Benefits

- Computationally efficient to find paths in the graph
- Don't have to find overlaps; they are implicit in the de Bruijn graph

Drawbacks

- The graph is very large; approximately one node per base
- Errors in the reads create spurious branches in the graph – requires error correction
- Max. size of k-mer is limited by the shortest read size

PineRefSeq

Assembly Strategy

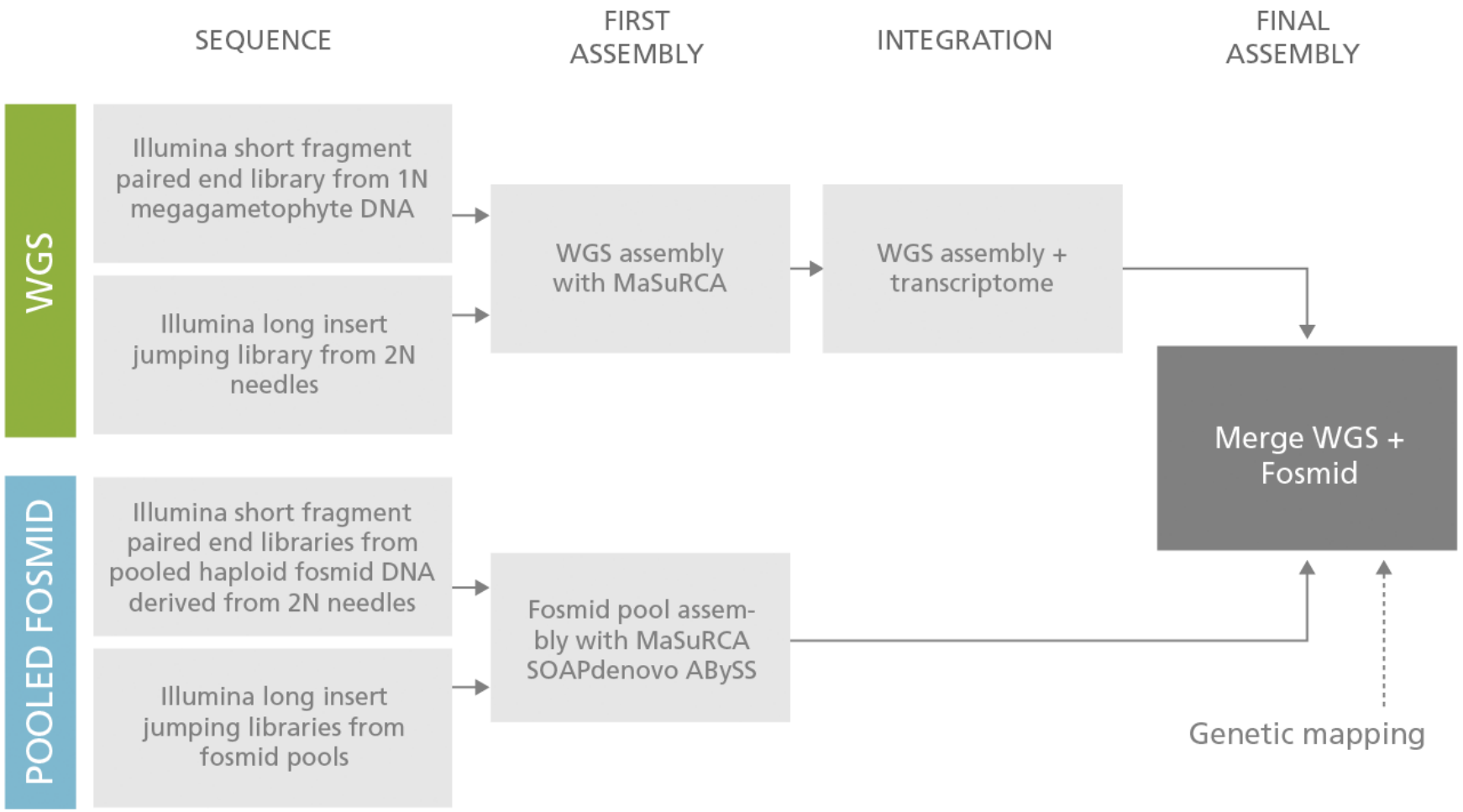


Figure Credit: Nicholas Wheeler, University of California, Davis

PineRefSeq

Dense Genetic Maps Aid Assembly

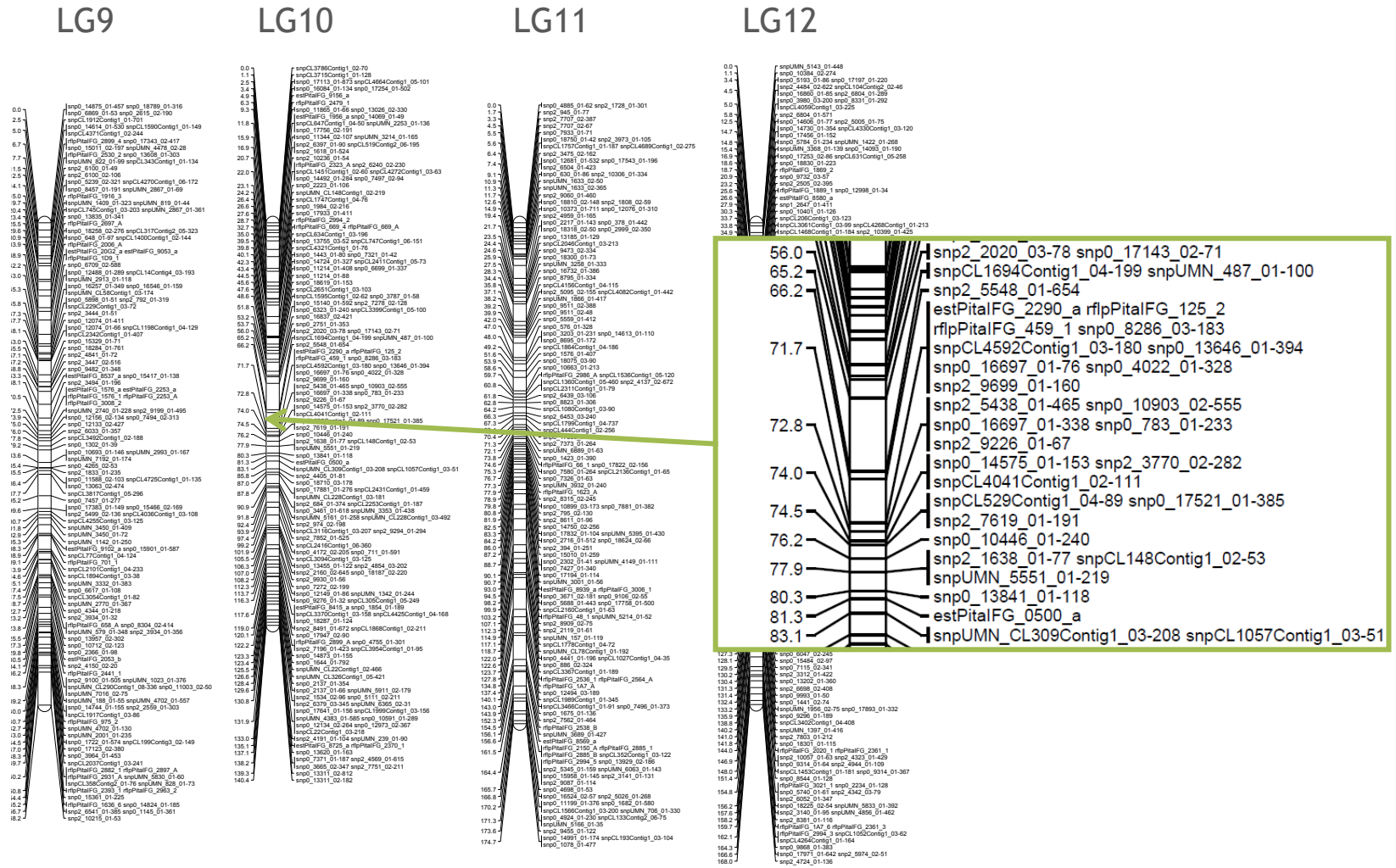


Figure Credit: Courtesy of Andrew Eckert and Pedro Martinez-Garcia, University of California, Davis

PineRefSeq

Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing, Jumping Libraries, & Genetic Maps

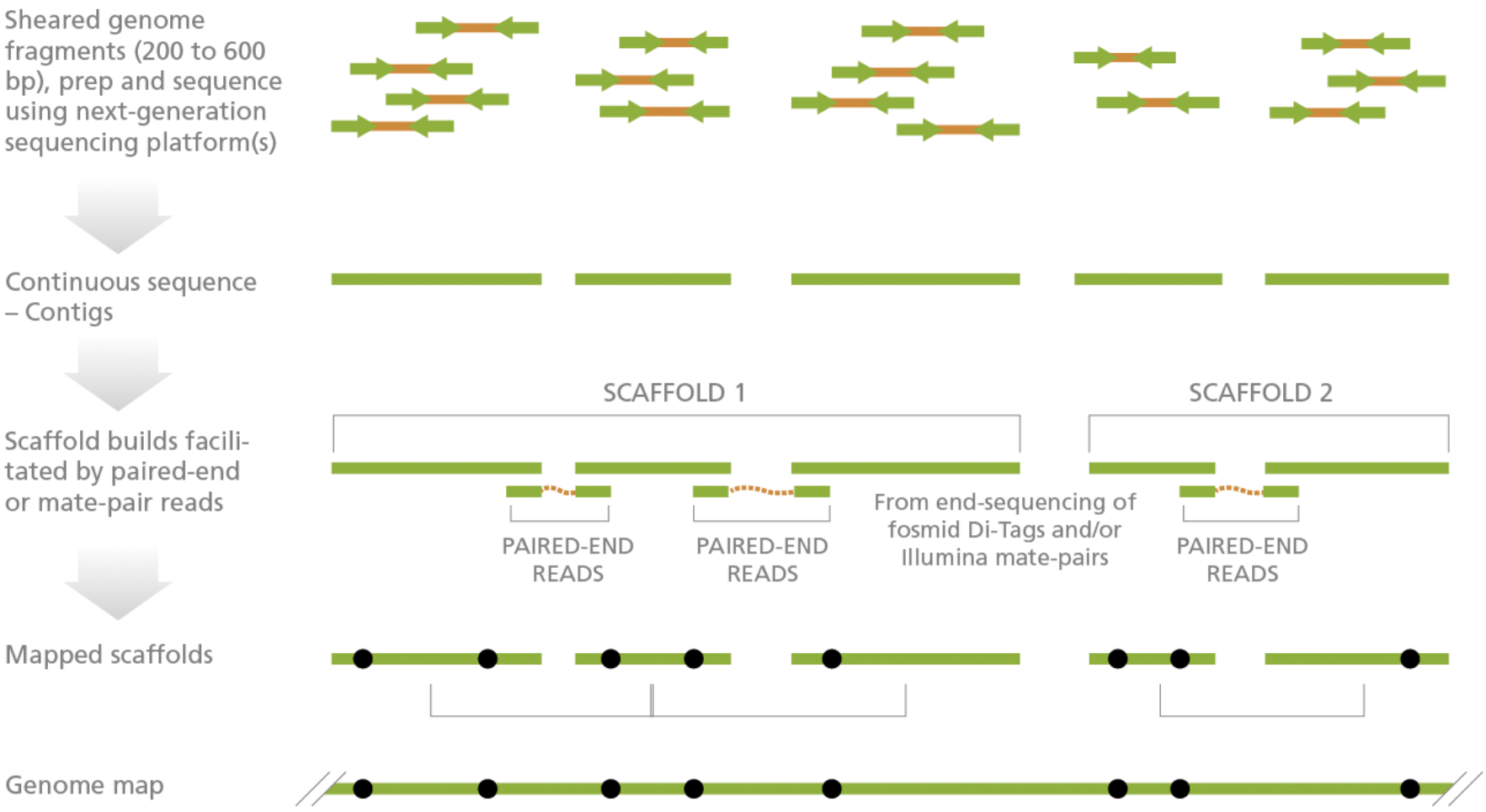
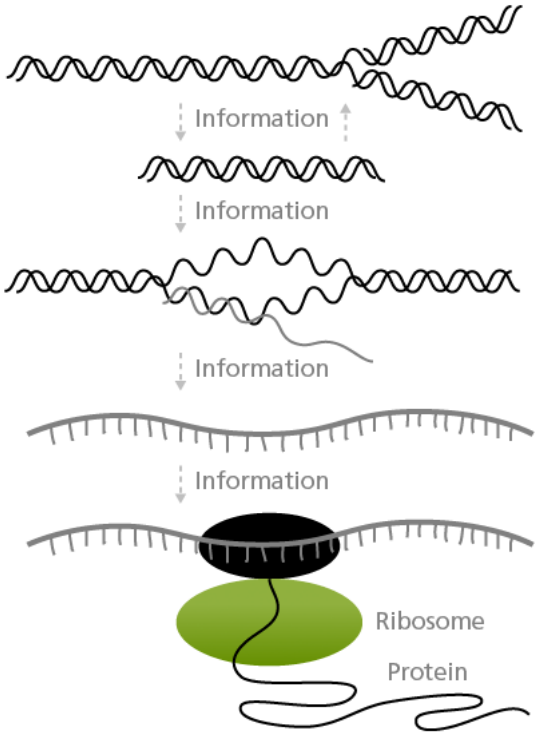


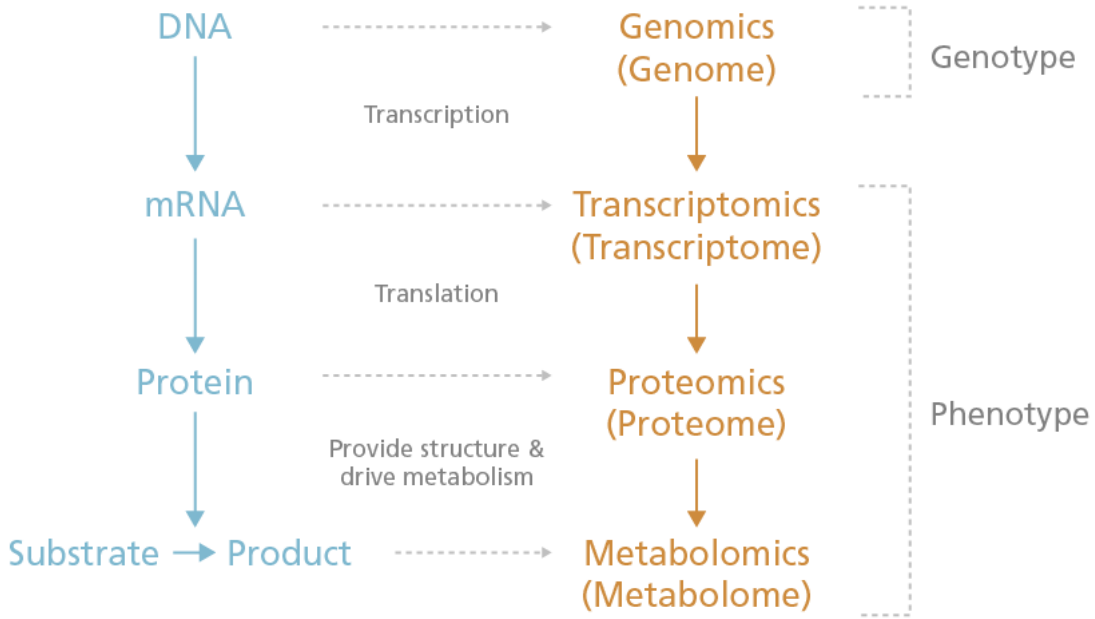
Figure Credit: Nicholas Wheeler, University of California, Davis

Transcriptome (RNA) sequencing defines the genes expressed in different pine tissues

PineRefSeq

Figure Credit: Modified from Keithanne Mockaitis, Indiana University





▼ Transcriptome libraries from many tissues and conditions are needed

How Does the Transcriptome Inform the Reference Genome?

PineRefSeq

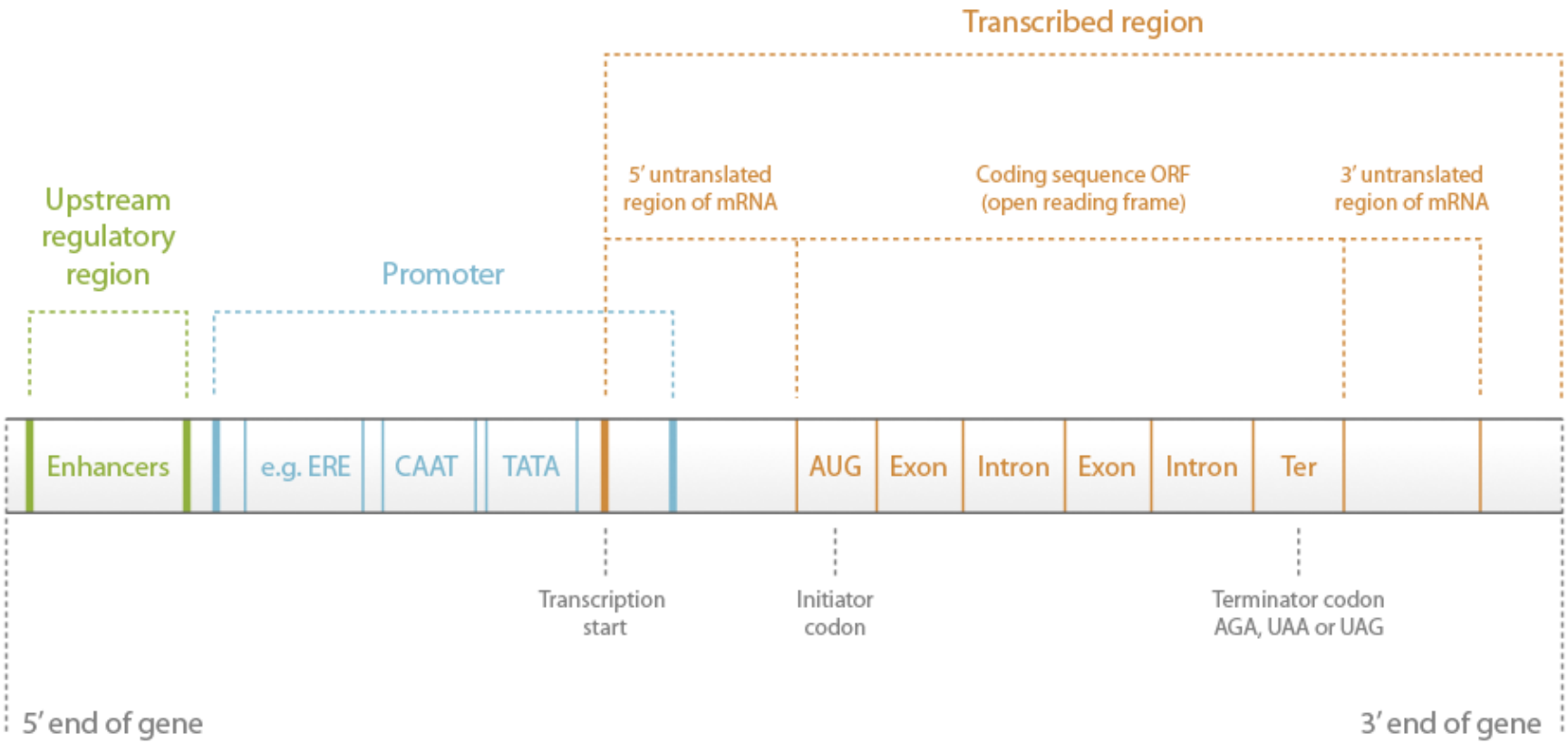


Figure Credit: Modified from Jill Wegrzyn, University of California, Davis

What Else Can We Learn From Transcriptomics ?

- Genes and Pathways
- Gene Function
- Diagnostics
- Gene Regulation
- Proxy for other Omics

Preliminary Results of Transcriptome Reference Sequencing for Three Conifer Species

PineRefSeq

Updated summary of transcriptome assemblies from 454 (CCG, JGI) and RNASeq (FS) in Psme (Douglas-fir), Pila (sugar pine), and Pita (loblolly pine).

Library	N, quality filtered	Nucleotides	Transcripts Assembled	Mean Contig Length	Unique Transcripts
Pila Needles & Candles 454 (Newbler)	1,096,017	387,174,063	28,910	955	49,035
Pila Needle RNASeq (Trinity)			33,961		
Psme Needles and Candles 454 (Newbler)	1,216,156	419,643,998	25,041	961	92,897
Psme Needle RNASeq (Trinity)			99,936		
Pita Shoot 454 (Newbler)	874,971	205,284,775	62,342	1,124	48,842
Pita Callus 454 (Newbler)	882,199	344,842,307	37,322		
Pita Stem 454 (Newbler)	934,760	310,498,816	43,234		

PineRefSeq

Genome Annotation

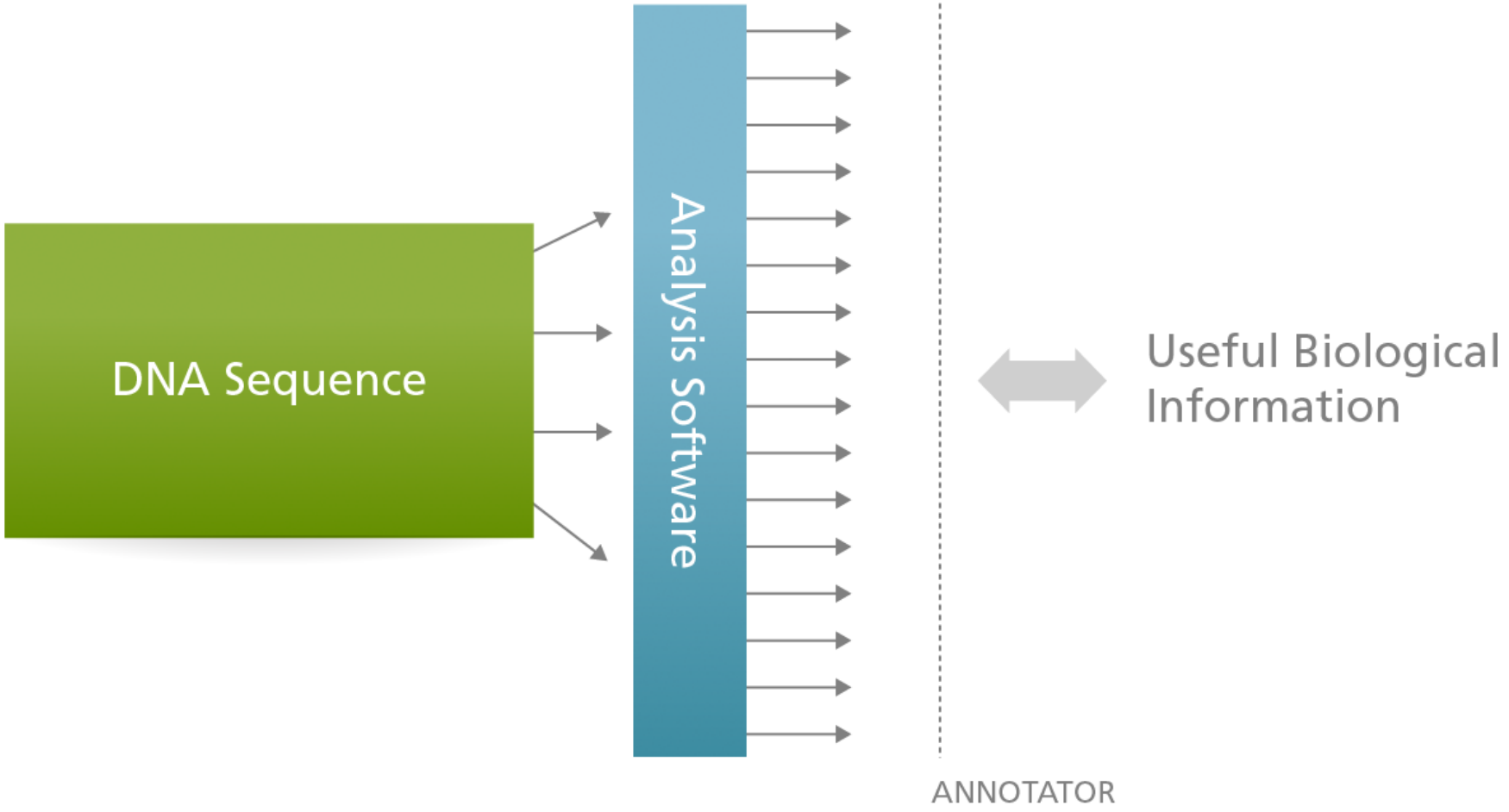


Figure Credit: Modified from Jill Wegrzyn, University of California, Davis

PineRefSeq

The Functional Annotation Process

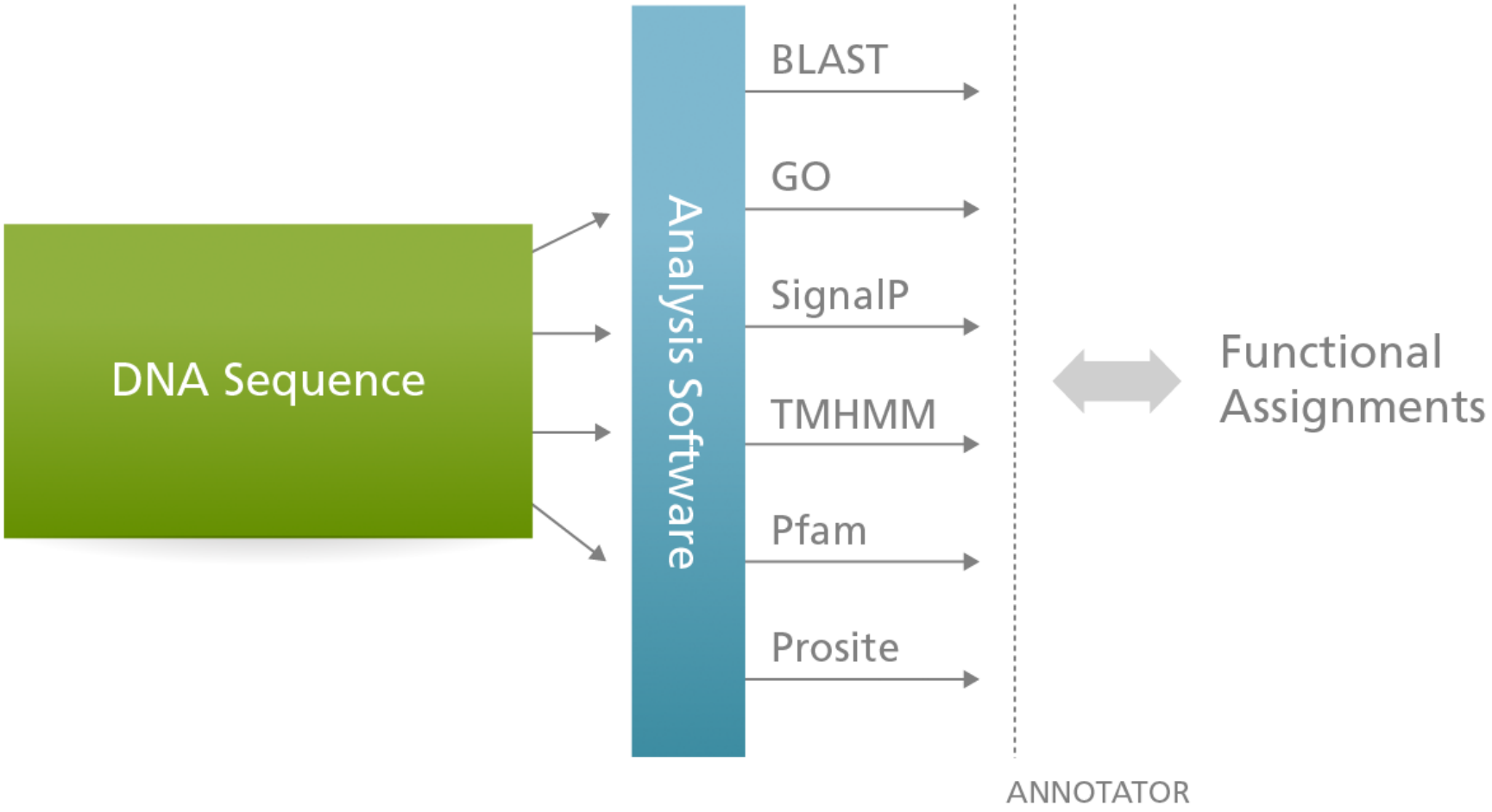
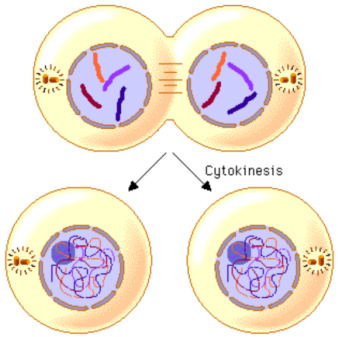


Figure Credit: Modified from Jill Wegrzyn, University of California, Davis

Gene Ontology

Biological Process

A commonly recognized series of events



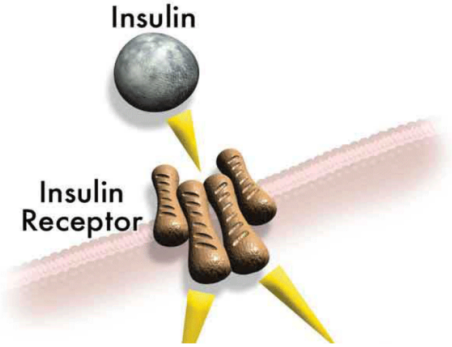
Cell division

Mitosis

Organelle fission

Molecular Function

An elemental activity or task or job



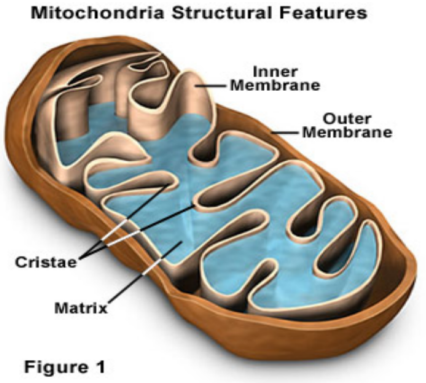
Protein kinase activity

Insulin binding

Insulin receptor activity

Cellular Component

Where a gene product is located



Mitochondrion

Mitochondrial matrix

Mitochondrial membrane

Figure Credit: Modified from Jill Wegrzyn, University of California, Davis

Transcriptome Ontology

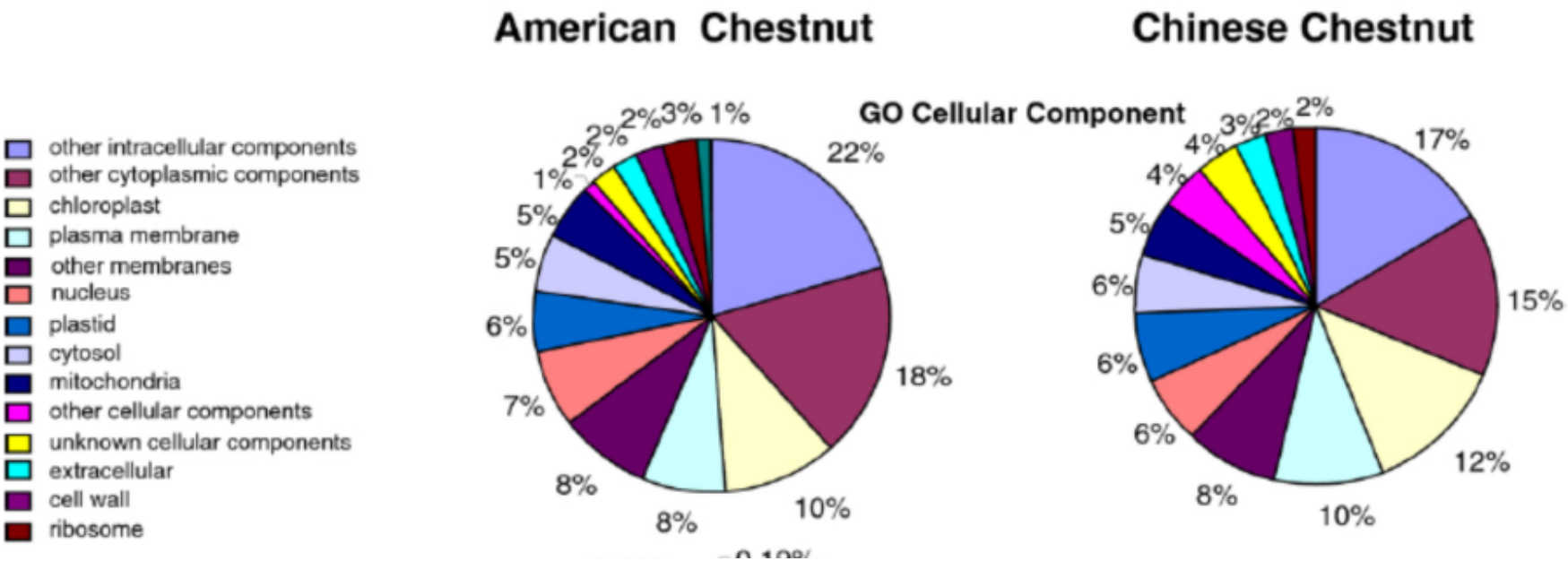


Figure Credit: Modified from Barakat et al. BMC Plant Biology 2012, 12:38 - http://www.biomedcentral.com/1471-2229/12/38

PineRefSeq

More Functional Annotation

Identifying and classifying regulatory sequences

Identifying and classifying genes that produce functional RNAs

- tRNA – Protein Synthesis
- rRNA - Protein Synthesis
- U snRNA – Splicing
- snoRNA - rRNA modification
- miRNA – Gene regulation

PineRefSeq

A Complete Annotation

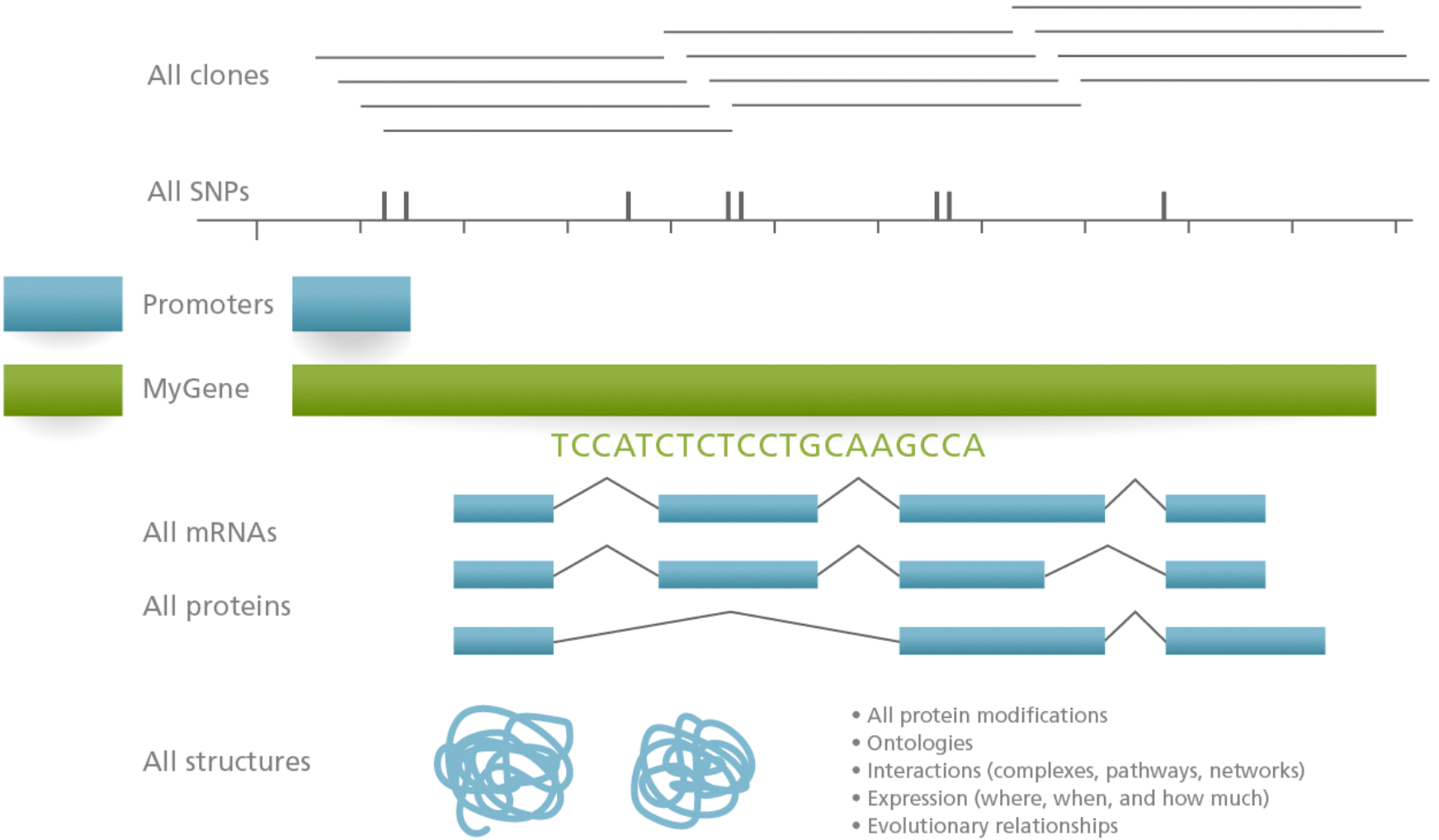
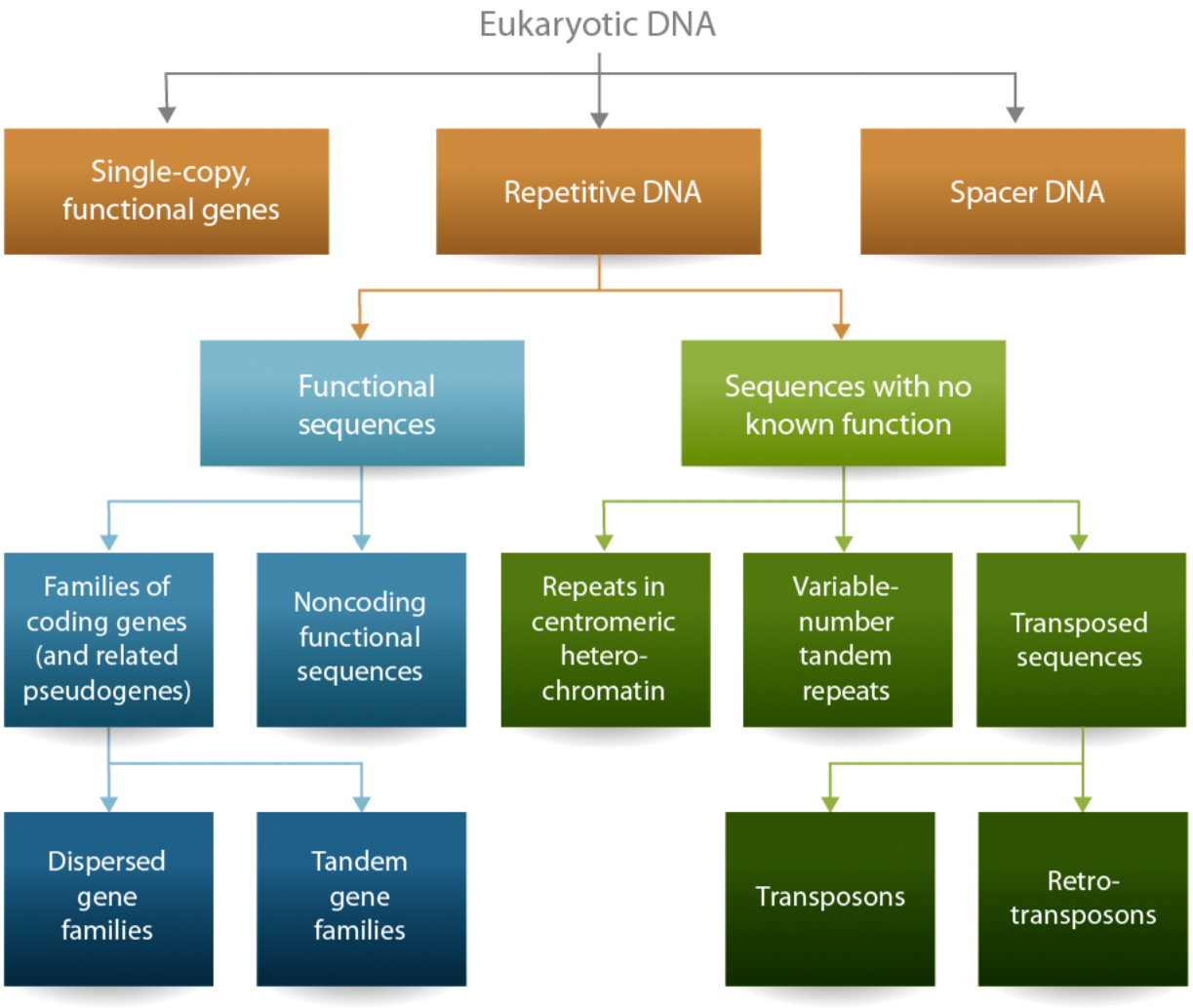


Figure Credit: Modified from Jill Wegrzyn, University of California, Davis

PineRefSeq

Annotating Structural Features of Genome Sequences



- Repetitive sequence content & distribution
- Pseudogenes/gene families
- GC content
- Segmental duplication
- Centromere and telomere structure

PineRefSeq

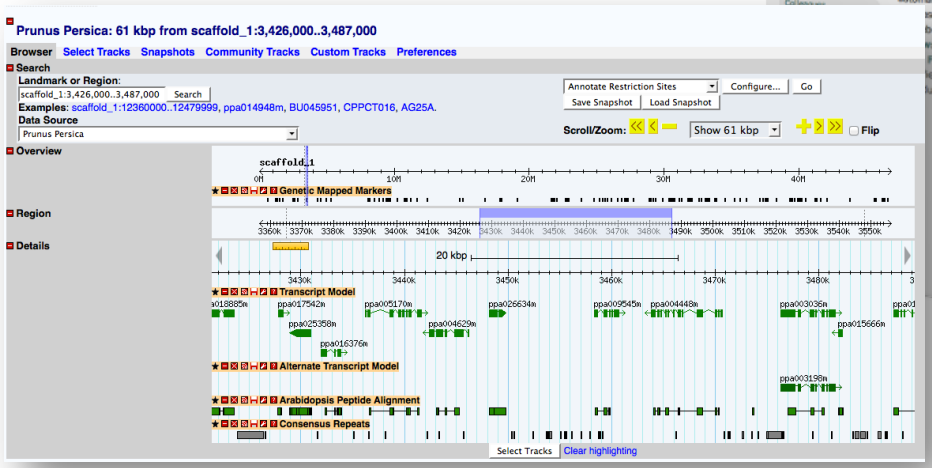
Database Resource Requirements

Project Level Genome Browsers

Goal: Provide capacity to capture, archive, curate, distribute, and analyze genomic information.

[TreeGenes]

[Genome Database for Rosaceae]



Public Databases

Nucleic Acid Sequence Databases

ENA / EBI – European Bioinformatics Institute

DDBJ – DNA Data Bank of Japan

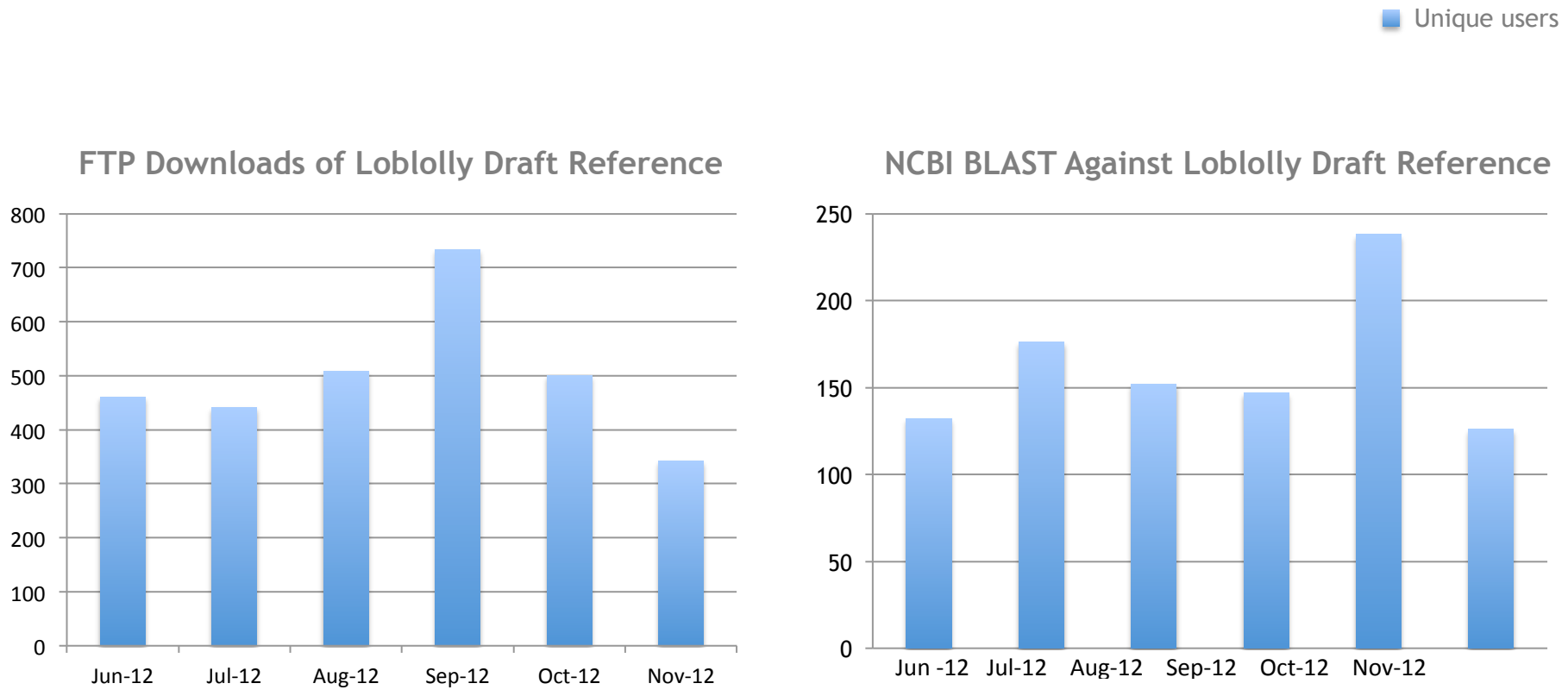
NCBI /GenBank – National Center for Biotechnology Info

NCBI Entrez

[NCBI Entrez](#) – The Life Sciences Search Engine

PineRefSeq

The Pine Reference Sequence in Use



Version 0.6 Release: June, 2012
Version 0.8 Release: January, 2013

Figure Credit: Modified from Jill Wegrzyn, University of California, Davis

Genome Assembly Statistics for Recently Sequenced Species

PineRefSeq

Year	Common Name	Scientific Name	Assembly Size (GB)	Predicted Size (GB)	N50 Contig (KB)	N50 Scaffold (KB)
2011	Potato	*Solanum tuberosum L.*	0.7	0.8	31.4	1320.0
2011	Orangutan	*Pongo abelii/ pygmaeus*	3.1	3.1	15.5	740.0
2011	Nake Mole Rat	*Heterocephalus glaber*	2.7		19.3	1590.0
2011	Atlantic Cod	*Gadus morhua*	0.8		2.8	690.0
2011	Coral Reef	*Acropora digitifera*	0.4	0.4	10.7	190.0
2012	Gorilla	*Gorilla gorilla gorilla*	2.9		11.9	914.0
2012	Oyster	*Crassostrea gigas*	0.6	0.6	19.4	400.0
2013	Radish	*Raphanus sativus L*	0.4	0.5	25.0	
2012	Wheat	*Triticum aestivum*	5.5	17.0	0.6	0.6
2013	**Loblolly Pine**	***Pinus taeda***	**20.1**	**22.0**	**8.2**	**30.7**

Key Website Resources

Human Genome Sites
http://www.ornl.gov/sci/techresources/Human_Genome/project/hgp.shtml
http://www.nature.com/nature/supplements/collections/humangenome/index.html
http://www.sciencemag.org/content/300/5617/286.abstract

Photo Credits: Slide 5
http://en.wikipedia.org/wiki/Craig_Venter
http://www.humanitiesandhealth.wordpress.com
http://en.wikipedia.org/wiki/James_D._Watson

Figure Credits: Slide 6
http://www.phytozome.net/
http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes

Resource – Broad Institute for Illumina Sequencing Technology
http://www.broadinstitute.org/scientific-community/science/platforms/genome-sequencing/broadillumina-genome-analyzer-boot-camp

Key Website Resources

Slide 38
http://dendrome.ucdavis.edu/treegenes/
http://www.rosaceae.org/

Slide 39
http://www.ebi.ac.uk/ena/
http://ddbj.sakura.ne.jp/
http://www.ncbi.nlm.nih.gov/genbank/

Slide 40
http://www.ncbi.nlm.nih.gov/sites/gquery

Side 43
Download:
http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v0.9/
BLAST: http://dendrome.ucdavis.edu/resources/blast/
Data Release Policy: http://www.pinegenome.org/pinerefseq/Data_Use_Policy.pdf

References Cited

Gibson, G., and S. V. Muse. 2004. A primer of genome science. (2nd Ed). Sinauer Associates, Sunderland, MA.

Lesk, A. M. 2012. Introduction to genomics. (2nd Ed). Oxford University Press, New York.