

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



United States
Department of
Agriculture

National Institute
of Food and
Agriculture



pinegenome.org/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

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

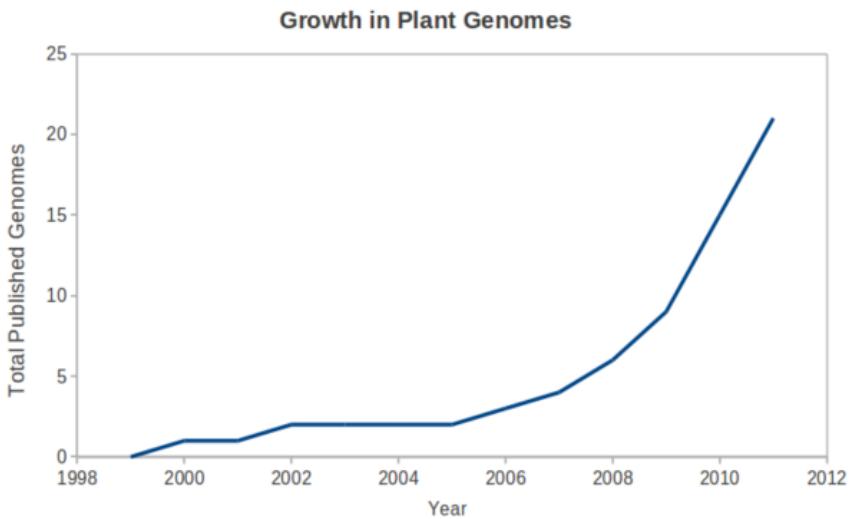
Craig Venter



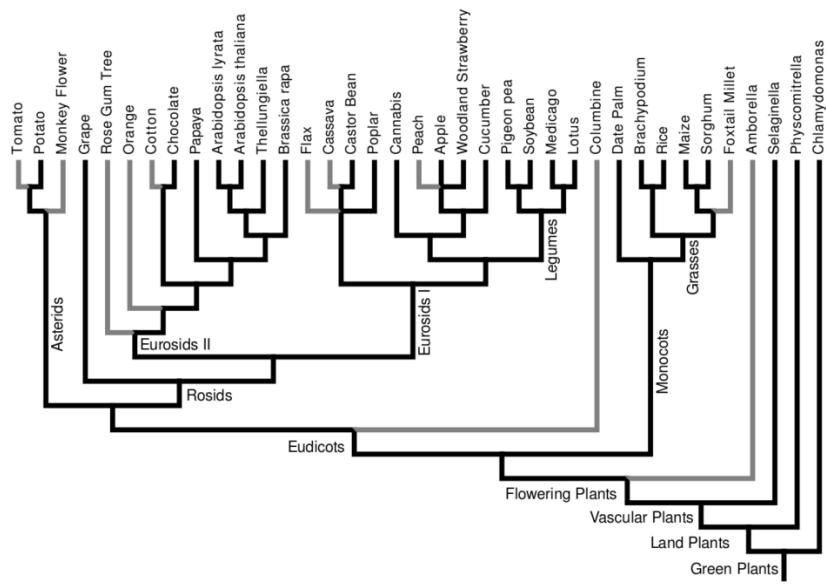
Source: Wikipedia

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



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				
<i>Populus trichocarpa</i>	Black Cottonwood	500	~40,000	2.0 / 2.2
<i>Eucalyptus grandis</i>	Rose Gum	691	~36,000	1.0 / 1.1
<i>Malus domestica</i>	Apple	881	~26,000	1.0 / 1.0
<i>Prunus persica</i>	Peach	227	~28,000	1.0 / 1.0
<i>Citrus sinensis</i>	Sweet Orange	319	~25,000	1.0 / 1.0
<i>Carica papaya</i>	Papaya	372	-	
<i>Amborella trichopoda</i>	Amborella	870	-	
<i>Betula nana</i>	Dwarf Birch	450	-	1.0 / -
In Progress or Planned - No Published Assemblies				
<i>Castanea mollissima</i>	Chinese Chestnut	800	-	
<i>Salix purpurea</i>	Purple Willow	327	-	
<i>Quercus robur</i>	Pedunculate Oak	740	-	
<i>Populus</i> spp. and ecotypes	Various	Various	-	
<i>Azadirachta indica</i>	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

Existing and Planned Gymnosperm Tree Genome Sequences

As of mid-2012

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

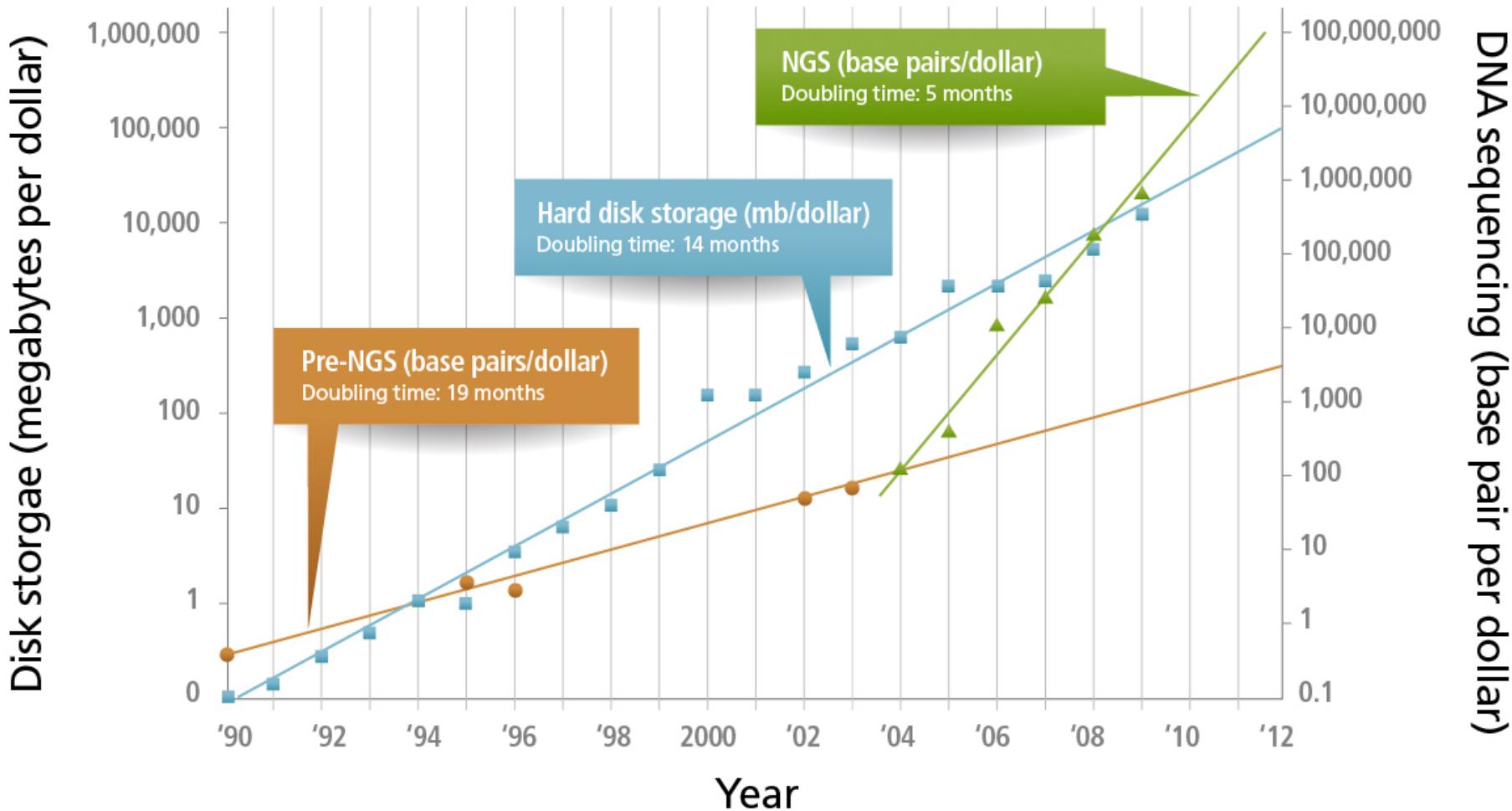


Figure credit: modified from Jill Wegrzyn, UC Davis

Why Do We Need a Conifer Genome Sequence?

Fundamental Genetic Information

Phylogenetic Representation

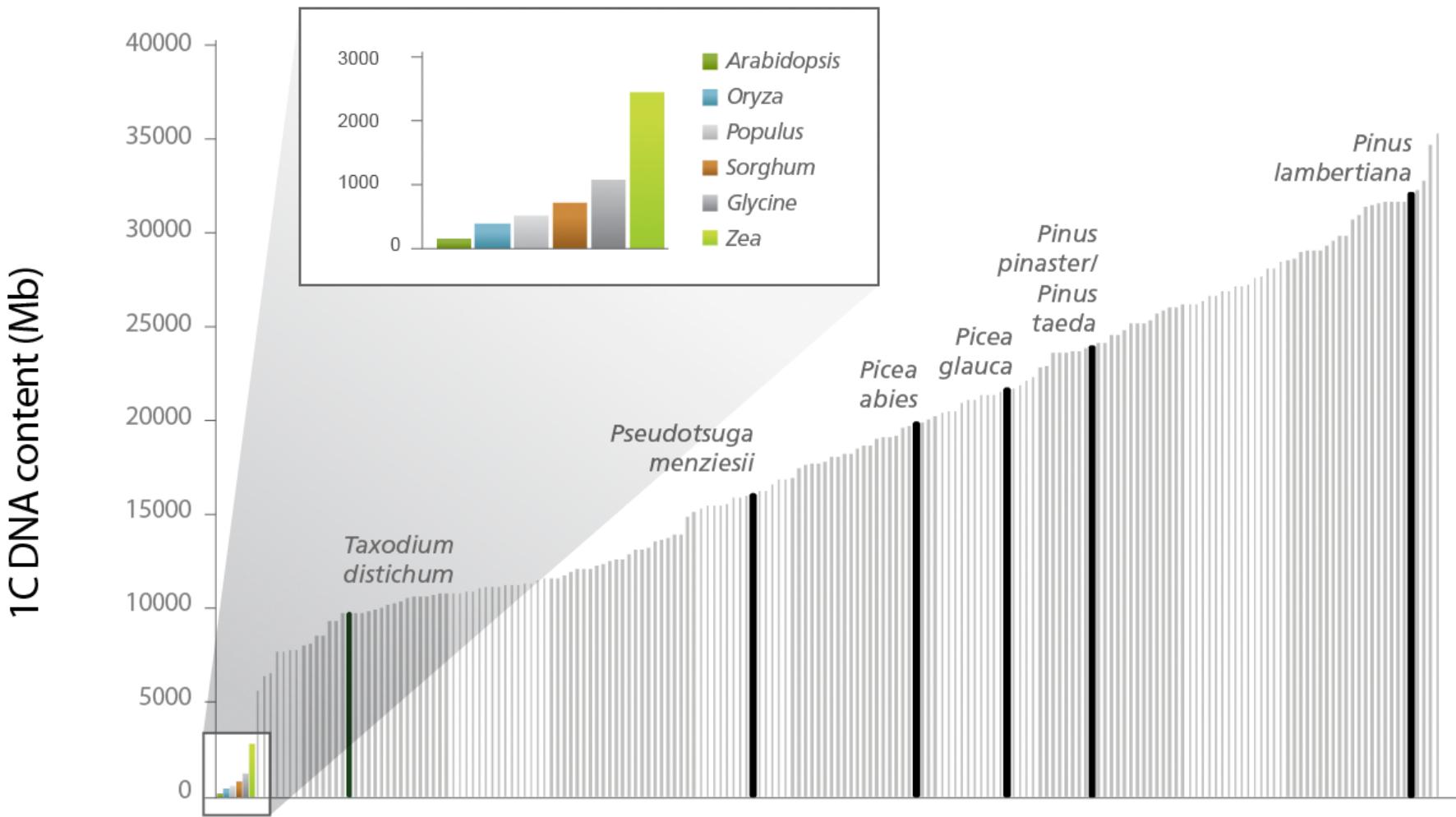
Ecological Representation

Development of Genomic Technologies

Economic Importance

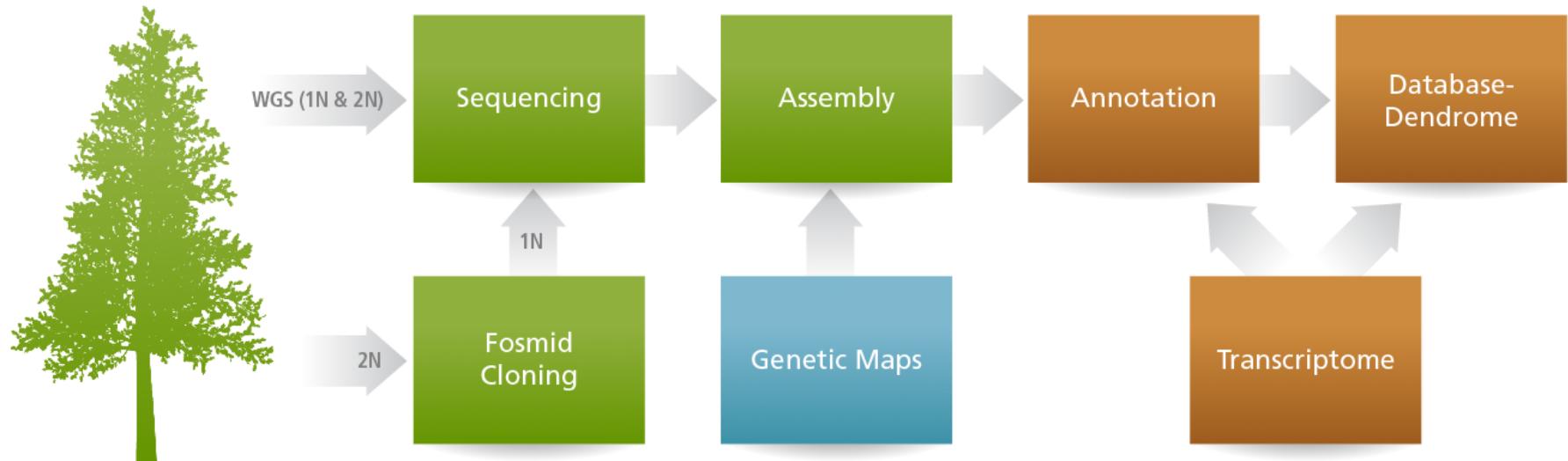


Challenges to Sequencing a Conifer Genome



Elements of a Conifer Genome Sequencing Project

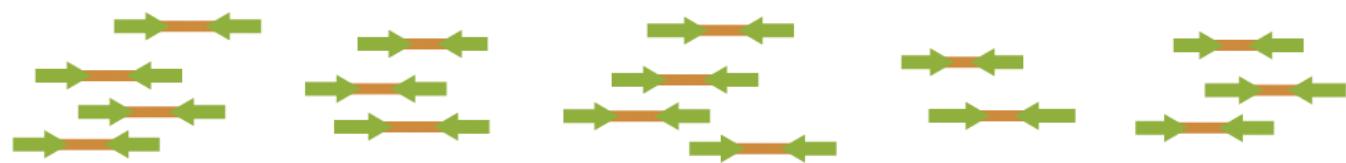
Approaches to Resolving Challenges



Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing

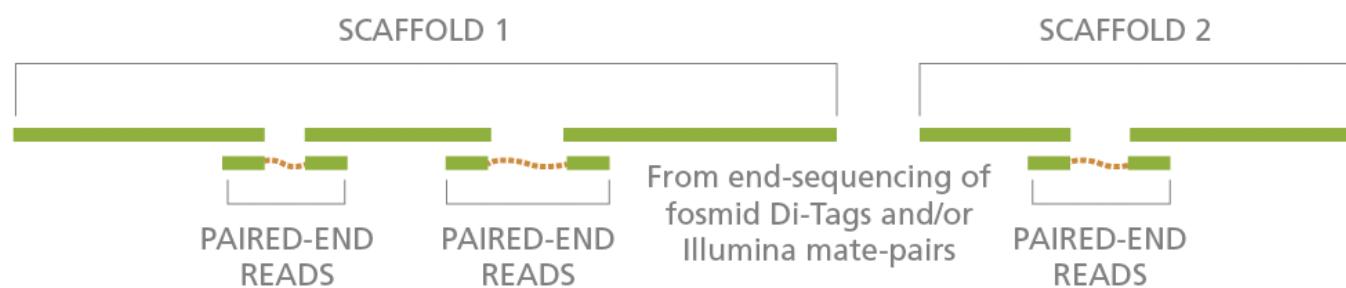
Sheared genome
fragments (200 to 600
bp), prep and sequence
using next-generation
sequencing platform(s)



Continuous sequence
– Contigs

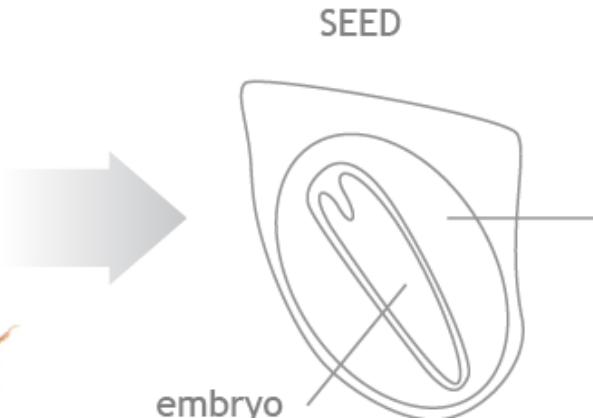


Scaffold builds facilitated by paired-end or mate-pair reads



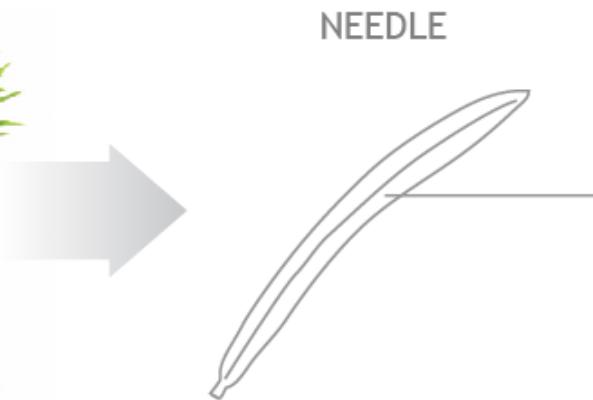
Acquiring the Sequence

Target Genome, Appropriate Tissues for DNA & RNA



Haploid

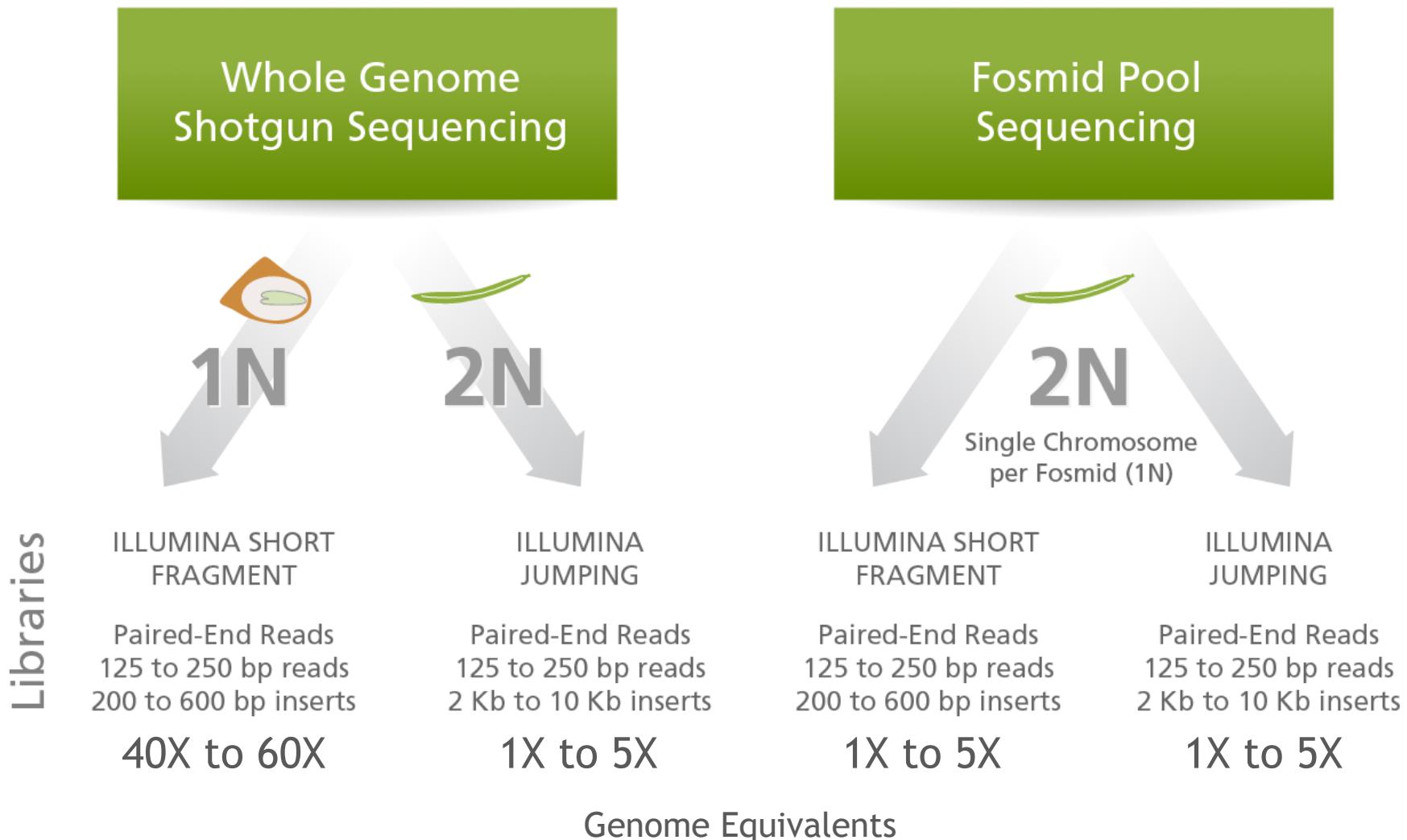
Haploid megagametophyte tissue
1N
Shotgun sequenced



Diploid

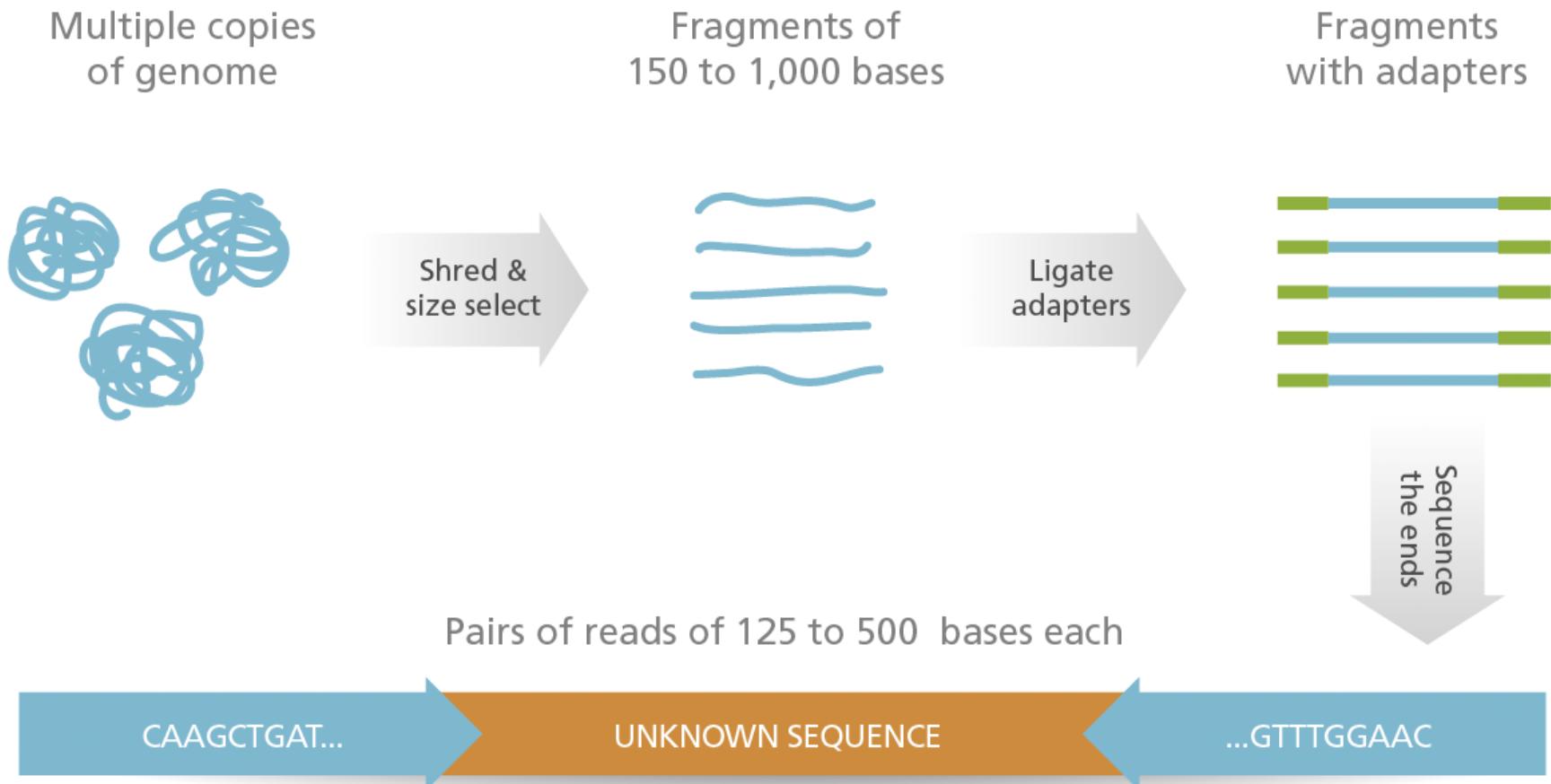
Diploid needle tissue
2N
40 Kb cloned fosmids, pooled
and sequenced

Sequencing Strategy



Whole Genome Shotgun Sequencing

Millions of Short “Reads”



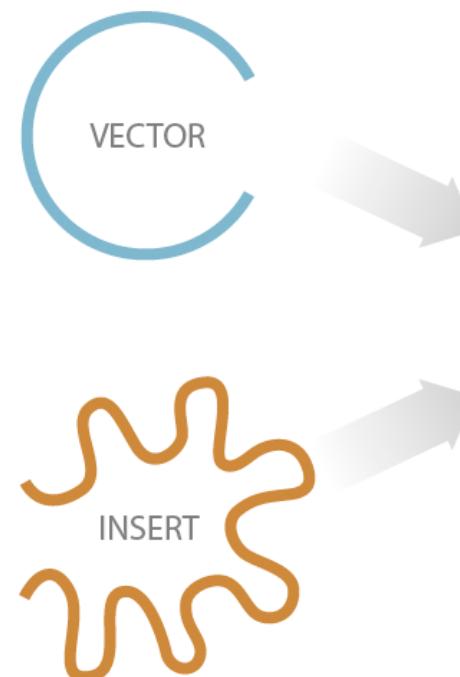
Visit the [Broad Institute](#) for details on DNA preparation, library construction, and sequencing technology of Illumina HiSeq

Sequencing DNA from Pools of Fosmid Clones

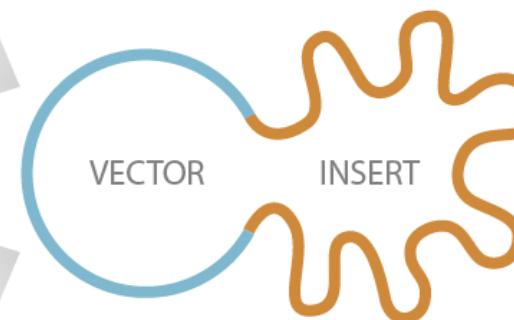
Fosmid Library Construction

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 transfected into appropriate bacterial cells for propagation.

Sequencing DNA from Pools of Fosmid Clones

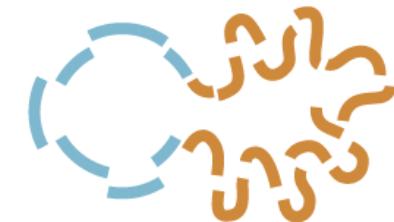
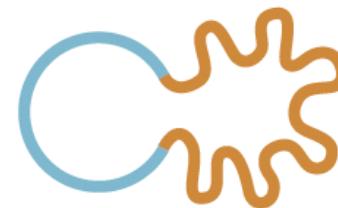
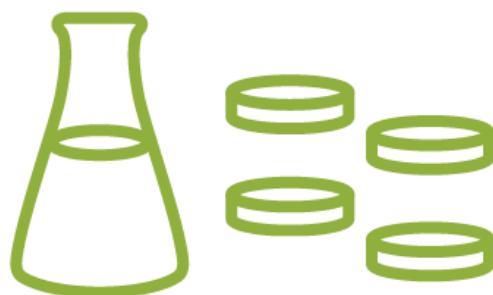
Preparing Fosmids for Sequencing

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.

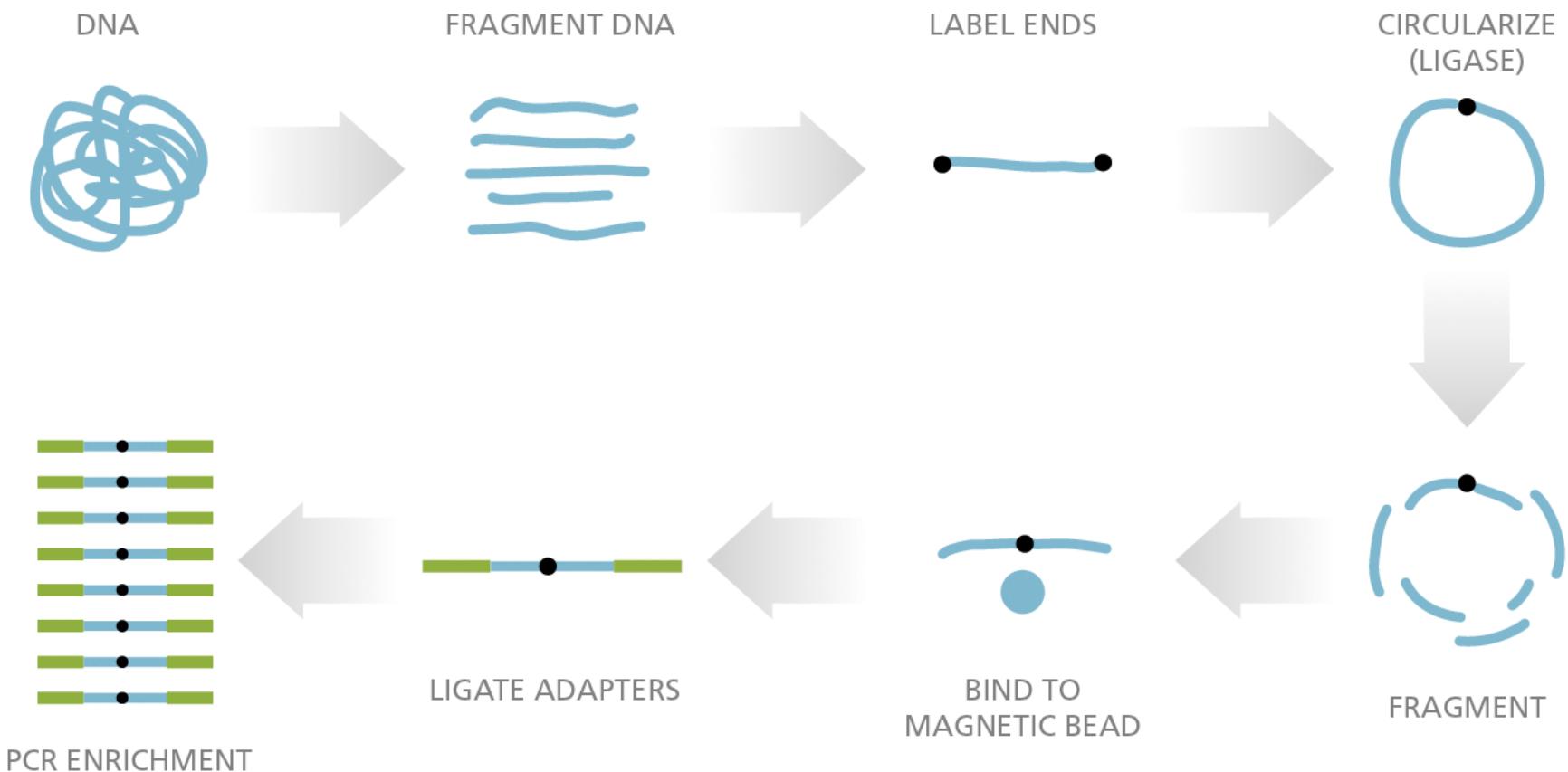
Fosmid DNA is prepared from each pool.

DNA is sheared and used for next-generation sequencing.



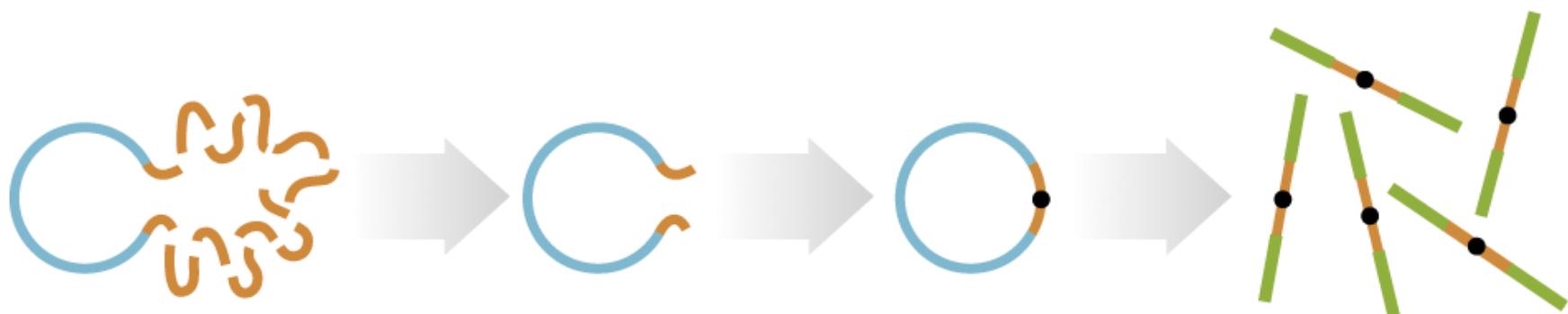
Jumping Libraries

Illumina Mate-Pair, Clone-Free



Jumping Libraries

Fosmid Di-Tag Cloned



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

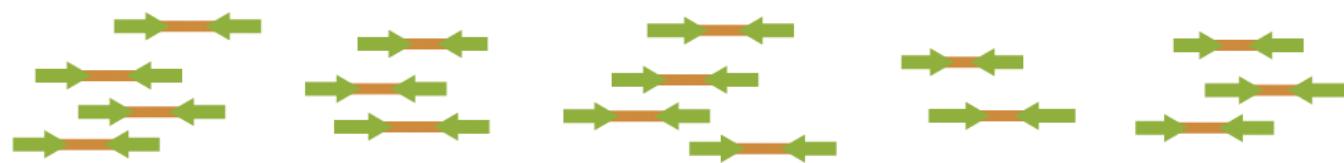
Vector is recircularized, DNA fragments that were ~40 kb apart in a genome are now brought together.

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

Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing

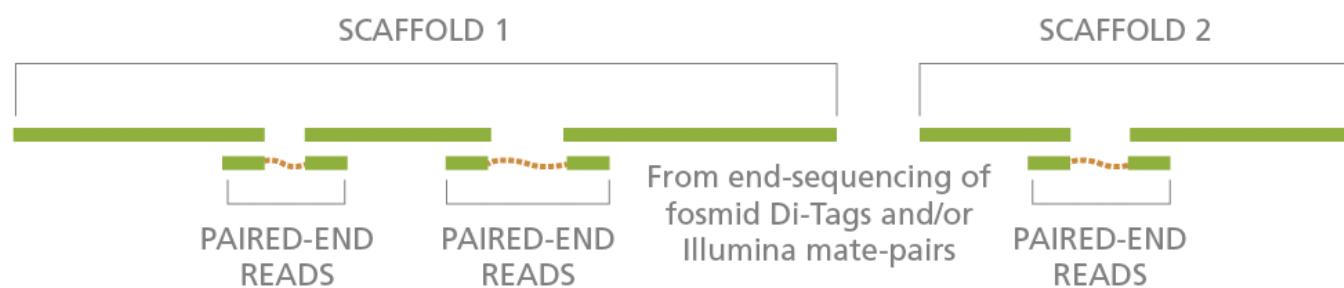
Sheared genome
fragments (200 to 600
bp), prep and sequence
using next-generation
sequencing platform(s)



Continuous sequence
– Contigs



Scaffold builds facilitated by paired-end or mate-pair reads



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

AGTGATTA **GATGATACTAGA**
||| ||| | | | |
GATGATA **GTAGA** GGATAGATTAA

Connect the pieces:
Create layout of numerous overlapping reads

AGTGATTAGATGATA**GTAGA**
GATGATA**C**TAGAGGGATAGACC
ATAG**T**AGAGGGATAGACCACACTCATCTAG

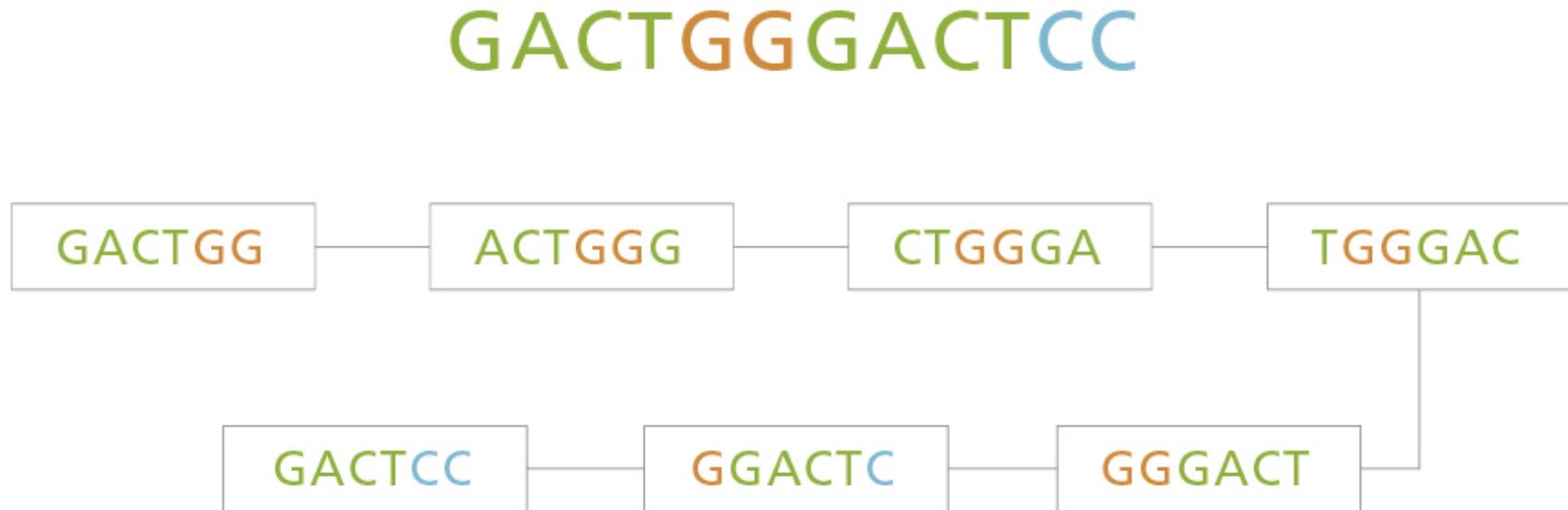
Create consensus sequence of contiguous nucleotides (i.e., contigs)

AGTGATTAGATGATA**GTAGAGGGATAGACCACACTCATCTAG**

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



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

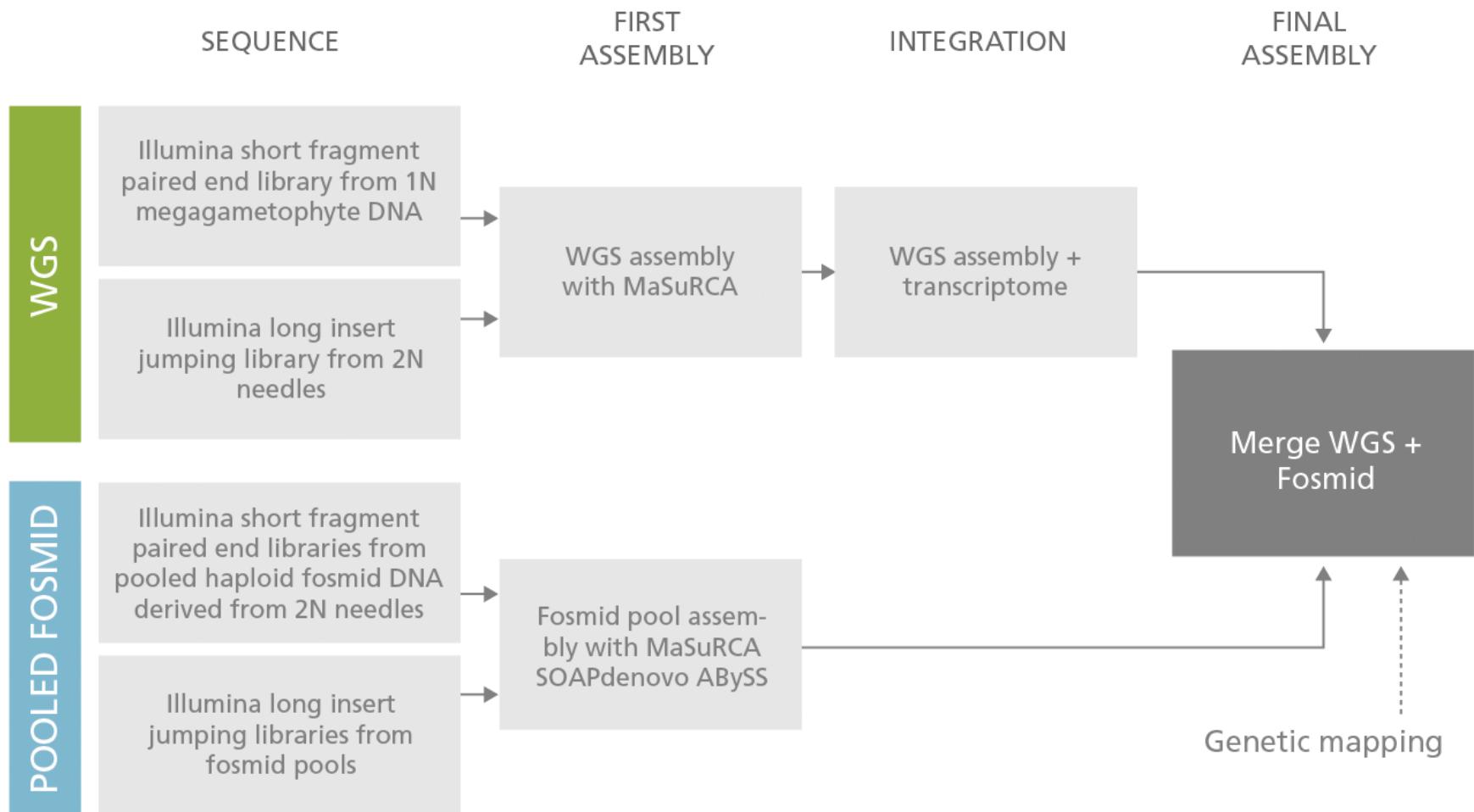
Benefits

- 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

Assembly Strategy



Dense Genetic Maps Aid Assembly

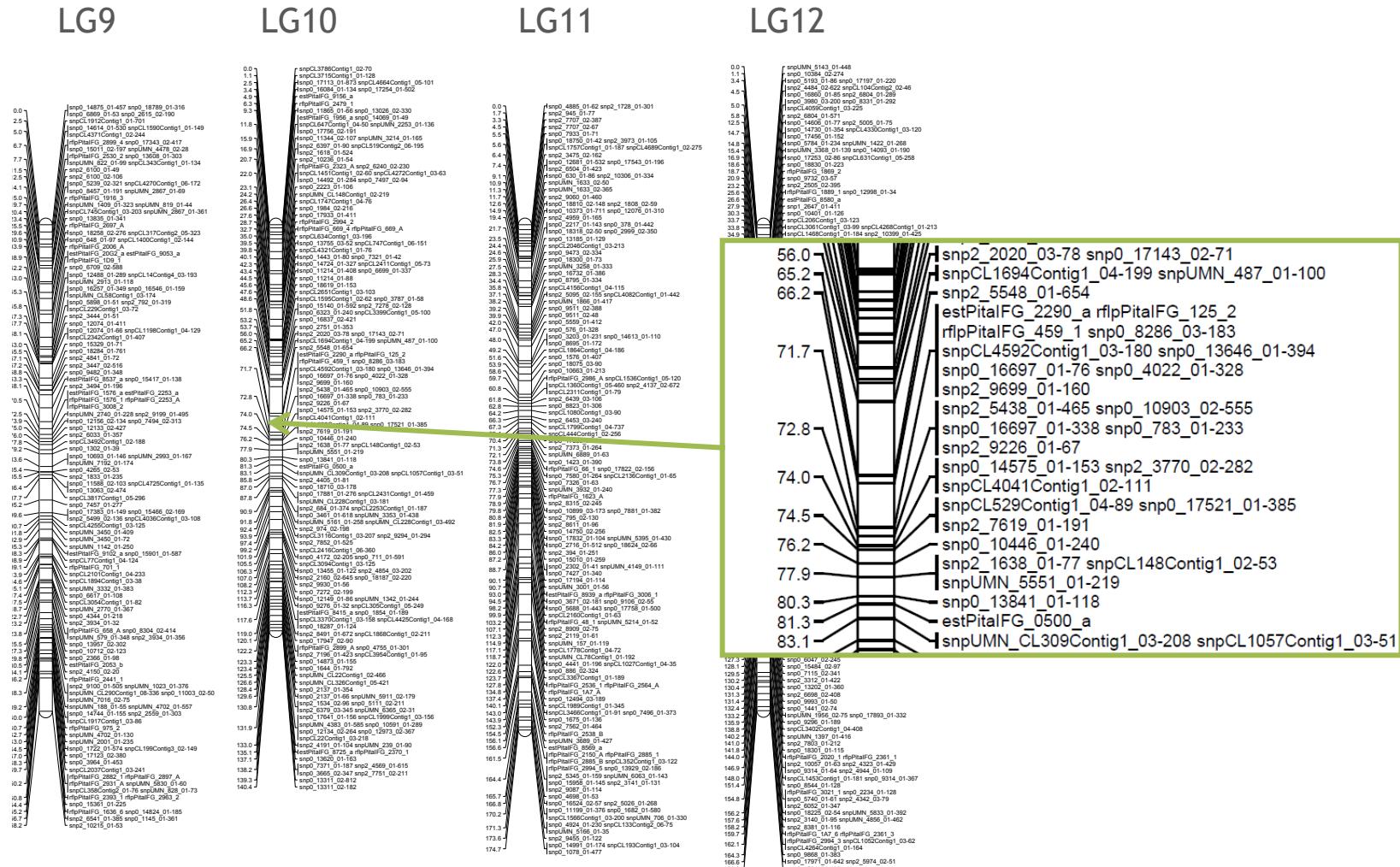
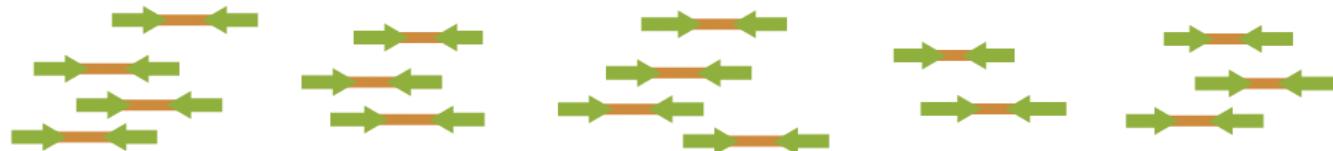


Figure Credit: Courtesy of Andrew Eckert and Pedro Martinez-Garcia, University of California, Davis

Assembling the Reference Sequence

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

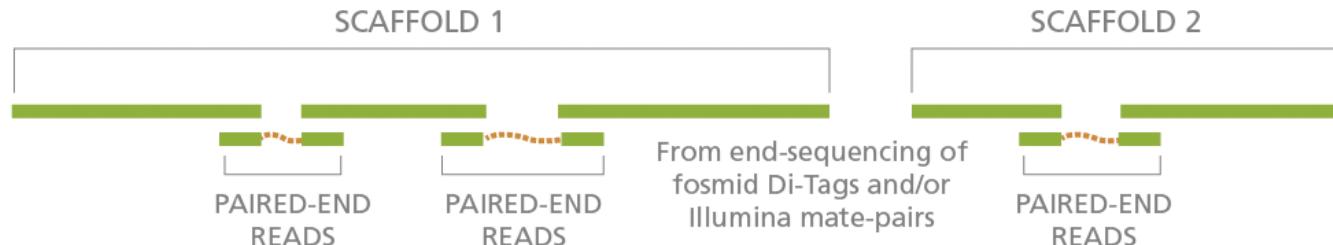
Sheared genome
fragments (200 to 600
bp), prep and sequence
using next-generation
sequencing platform(s)



Continuous sequence
– Contigs



Scaffold builds facili-
tated by paired-end
or mate-pair reads



Mapped scaffolds

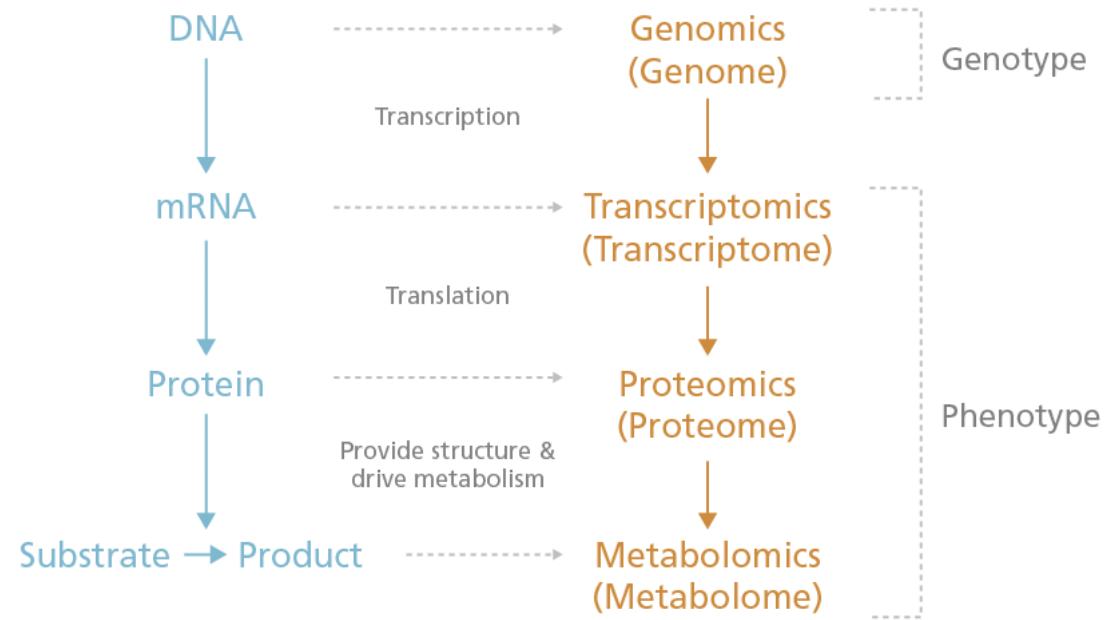
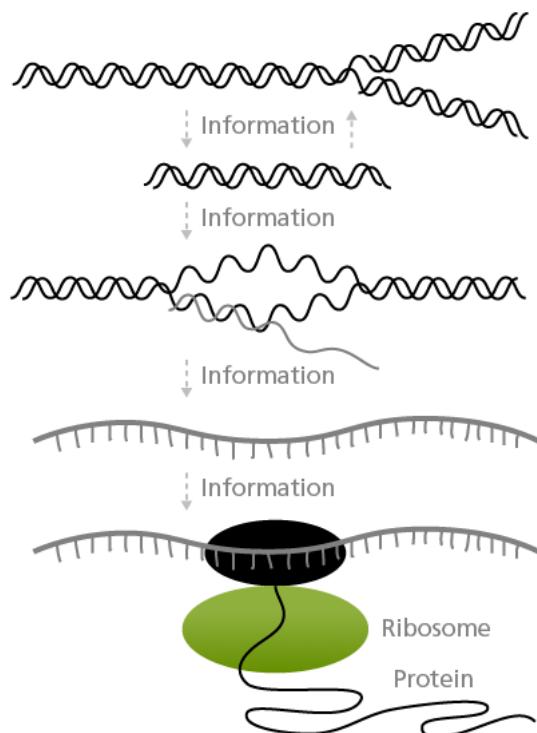


Genome map



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

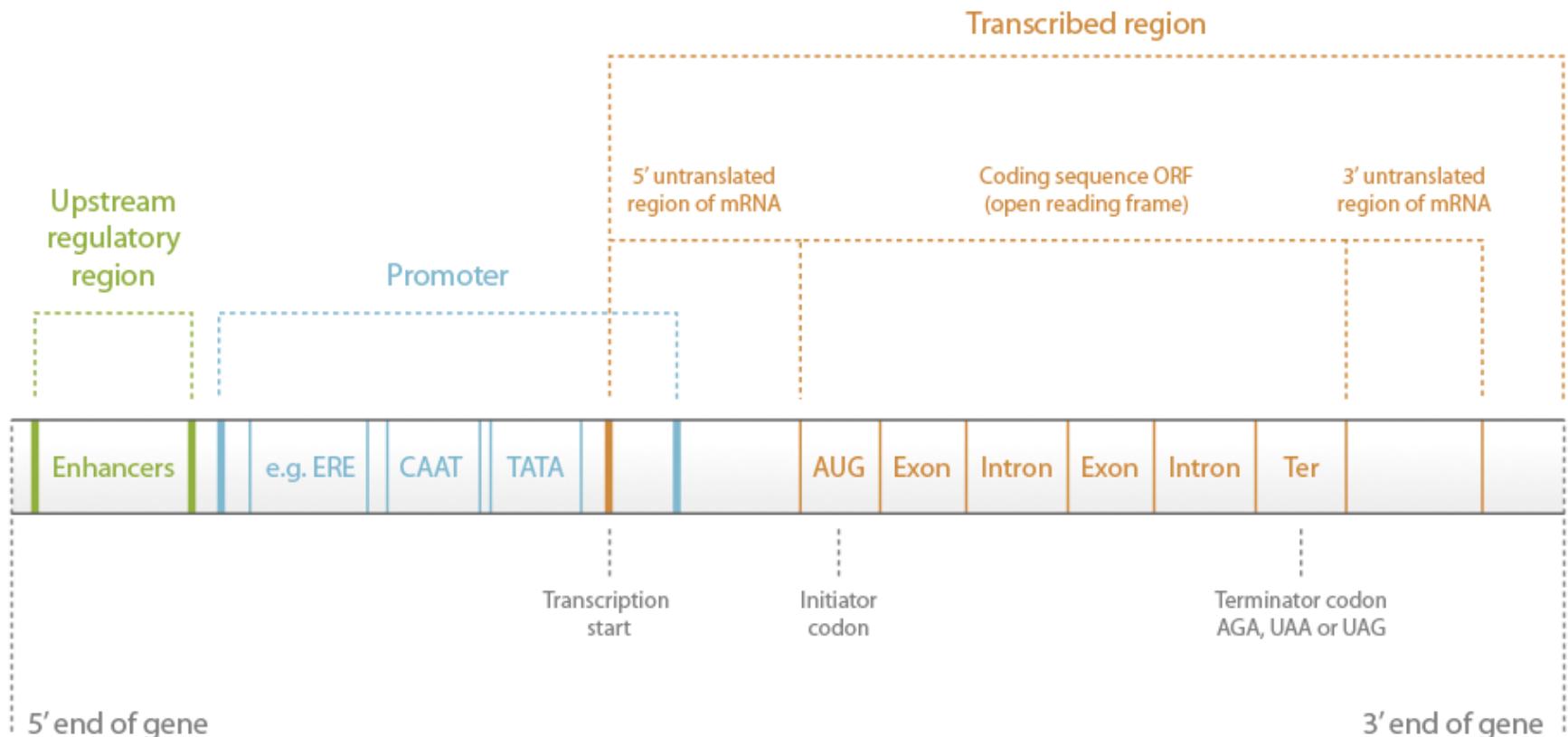
Figure Credit: Modified from Keithanne Mockaitis, Indiana University



▼ Transcriptome libraries from many tissues and conditions are needed



How Does the Transcriptome Inform the Reference Genome?



What Else Can We Learn From Transcriptomics ?

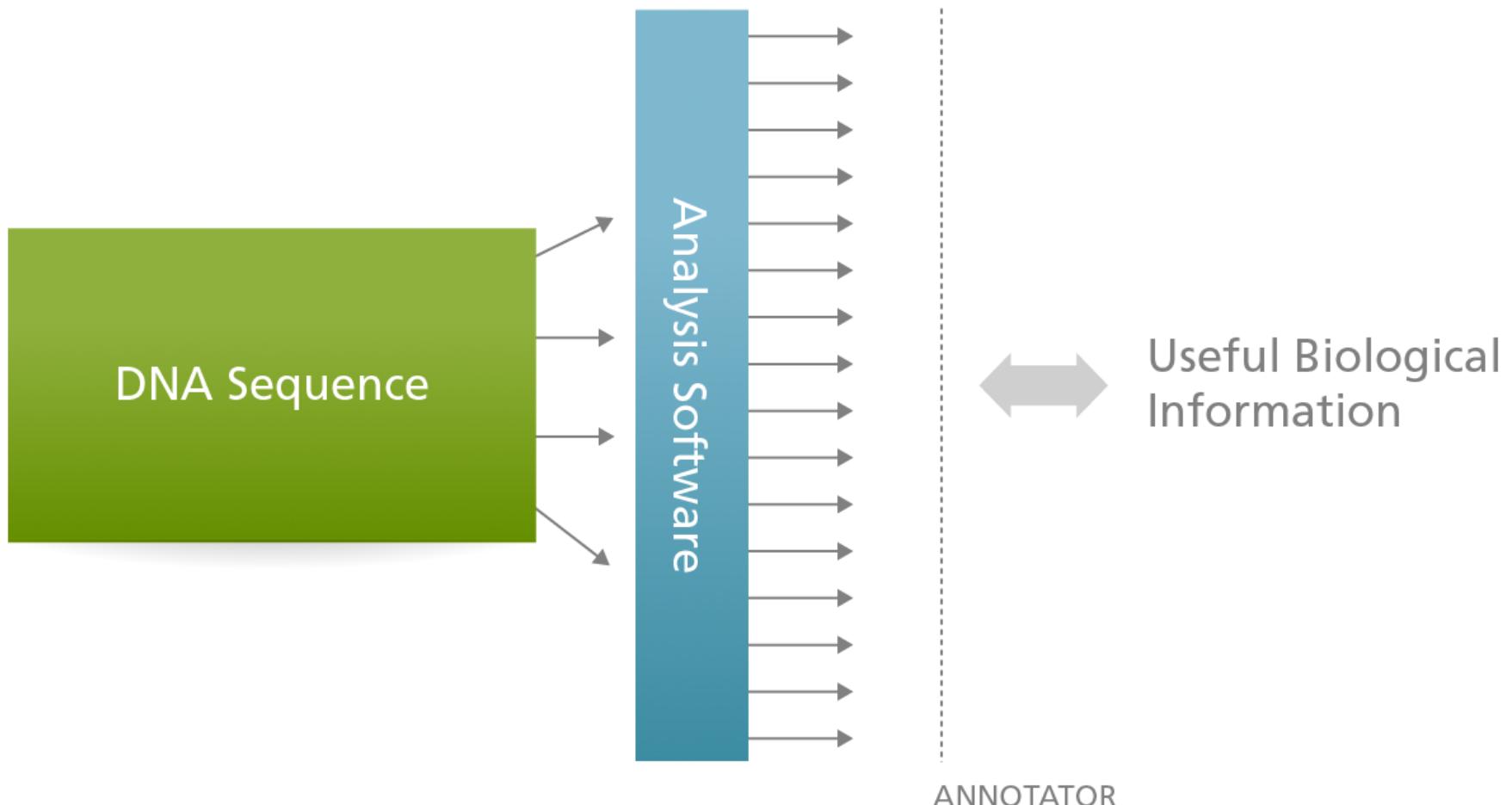
- Genes and Pathways
- Gene Function
- Diagnostics
- Gene Regulation
- Proxy for other Omics

Preliminary Results of Transcriptome Reference Sequencing for Three Conifer Species

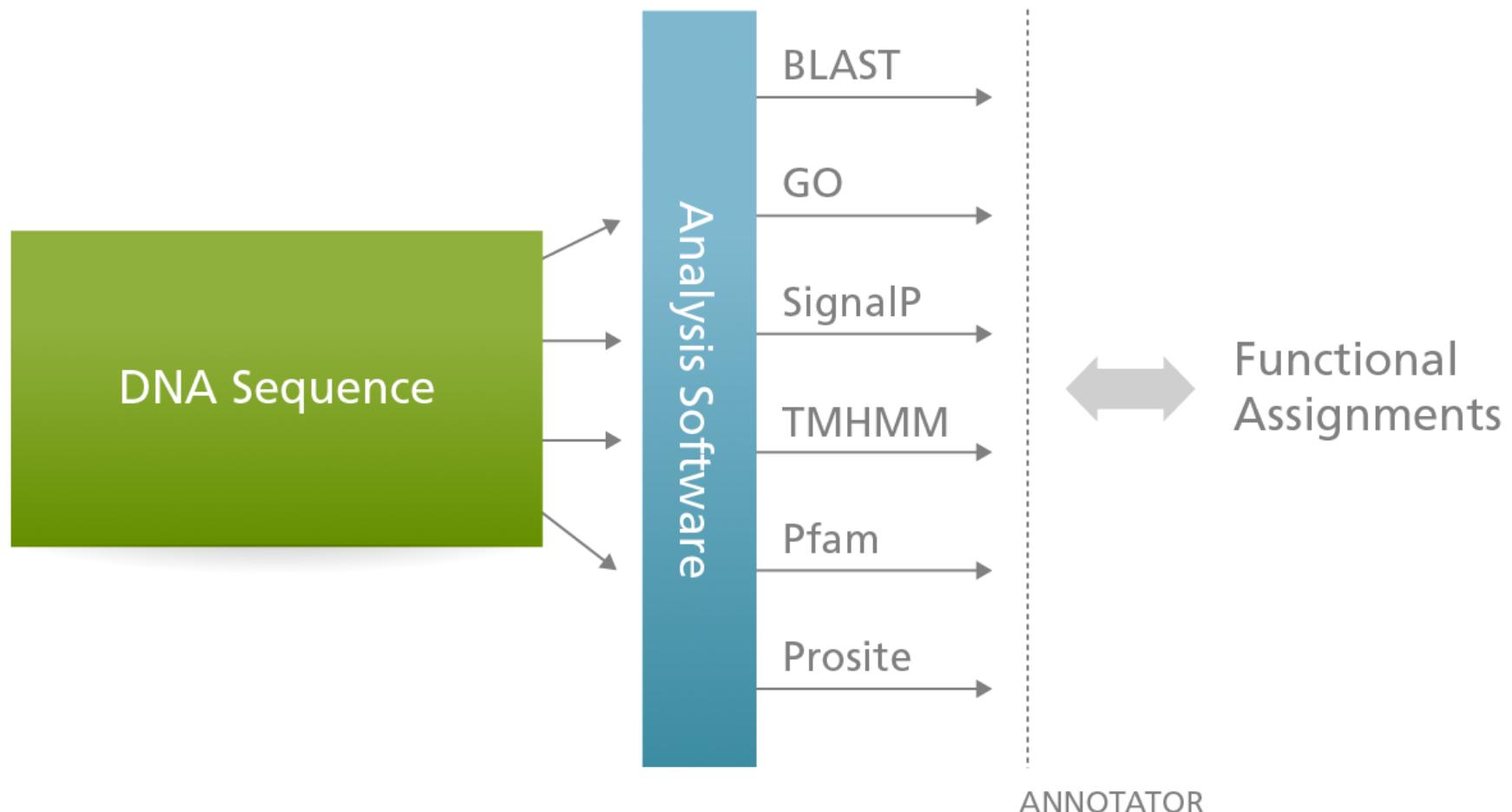
Updated summary of transcriptome assemblies from 454 (CCG, JGI) and RNASeq (FS) in Psme (Douglas-fir), Pila (sugar pine), and Pita (loblolly pine).

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

Genome Annotation



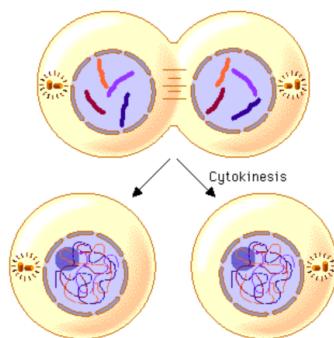
The Functional Annotation Process



Gene Ontology

Biological Process

A commonly recognized series of events



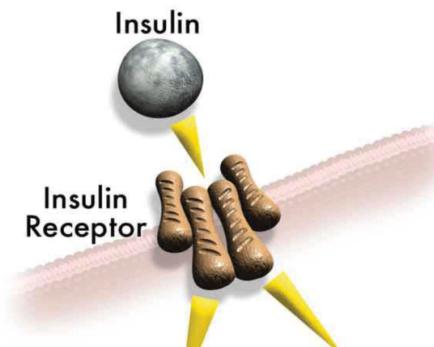
Cell division

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

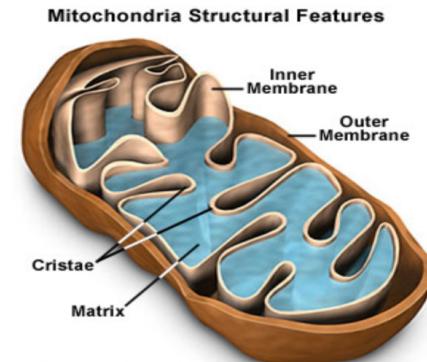


Figure 1

Mitochondrion

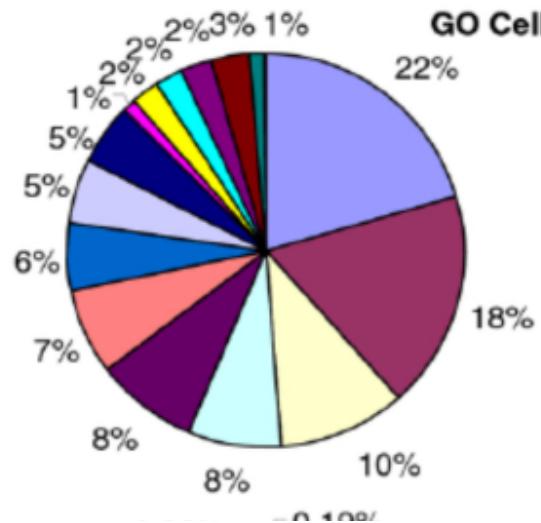
Mitochondrial matrix

Mitochondrial membrane

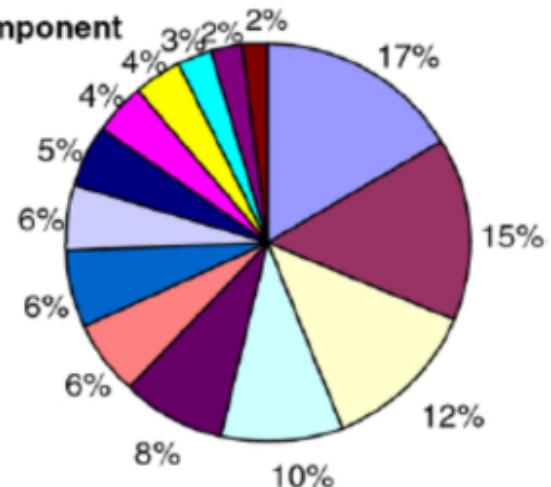
Transcriptome Ontology

- █ other intracellular components
- █ other cytoplasmic components
- █ chloroplast
- █ plasma membrane
- █ other membranes
- █ nucleus
- █ plastid
- █ cytosol
- █ mitochondria
- █ other cellular components
- █ unknown cellular components
- █ extracellular
- █ cell wall
- █ ribosome

American Chestnut



Chinese Chestnut



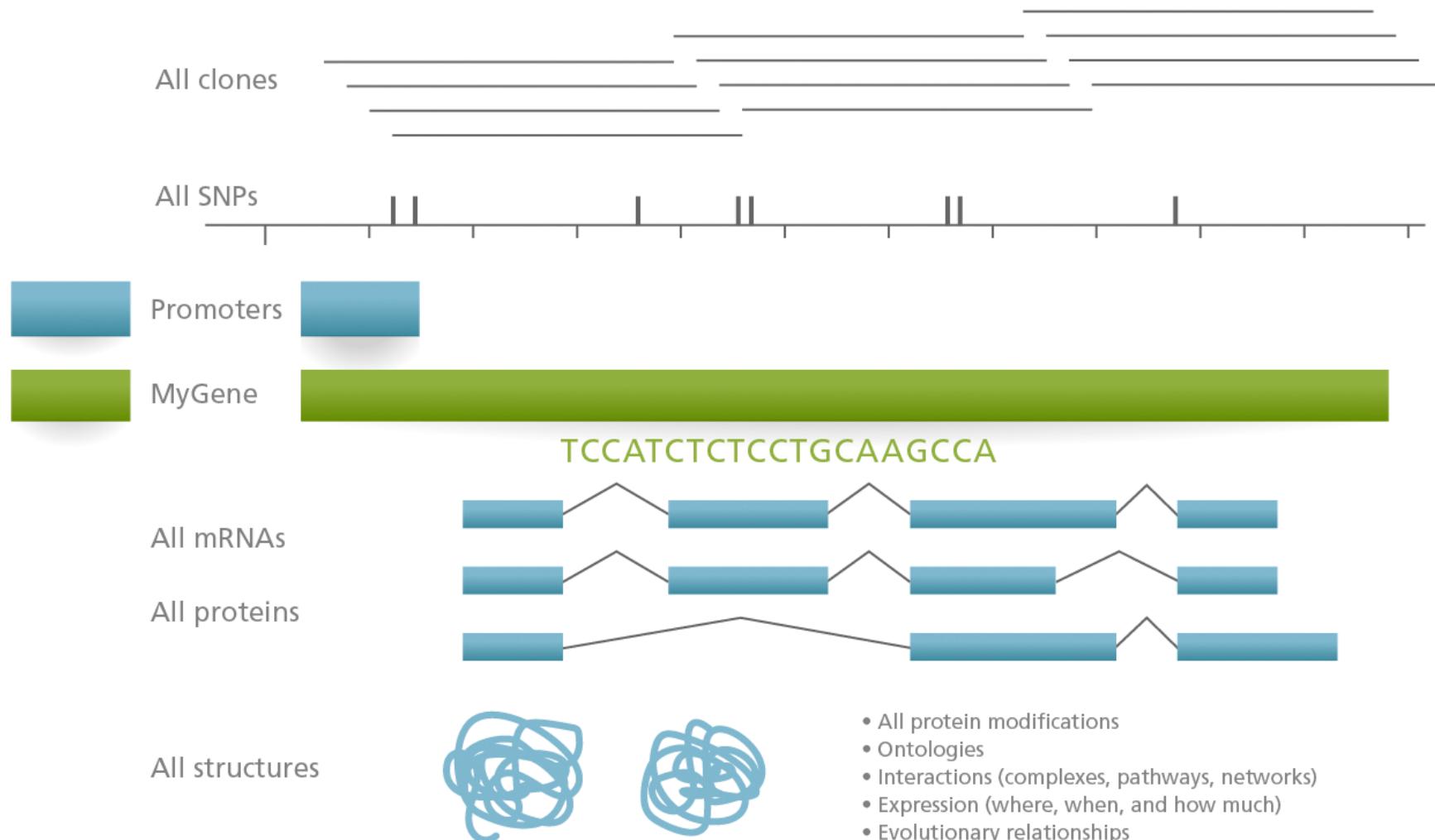
More Functional Annotation

Identifying and classifying regulatory sequences

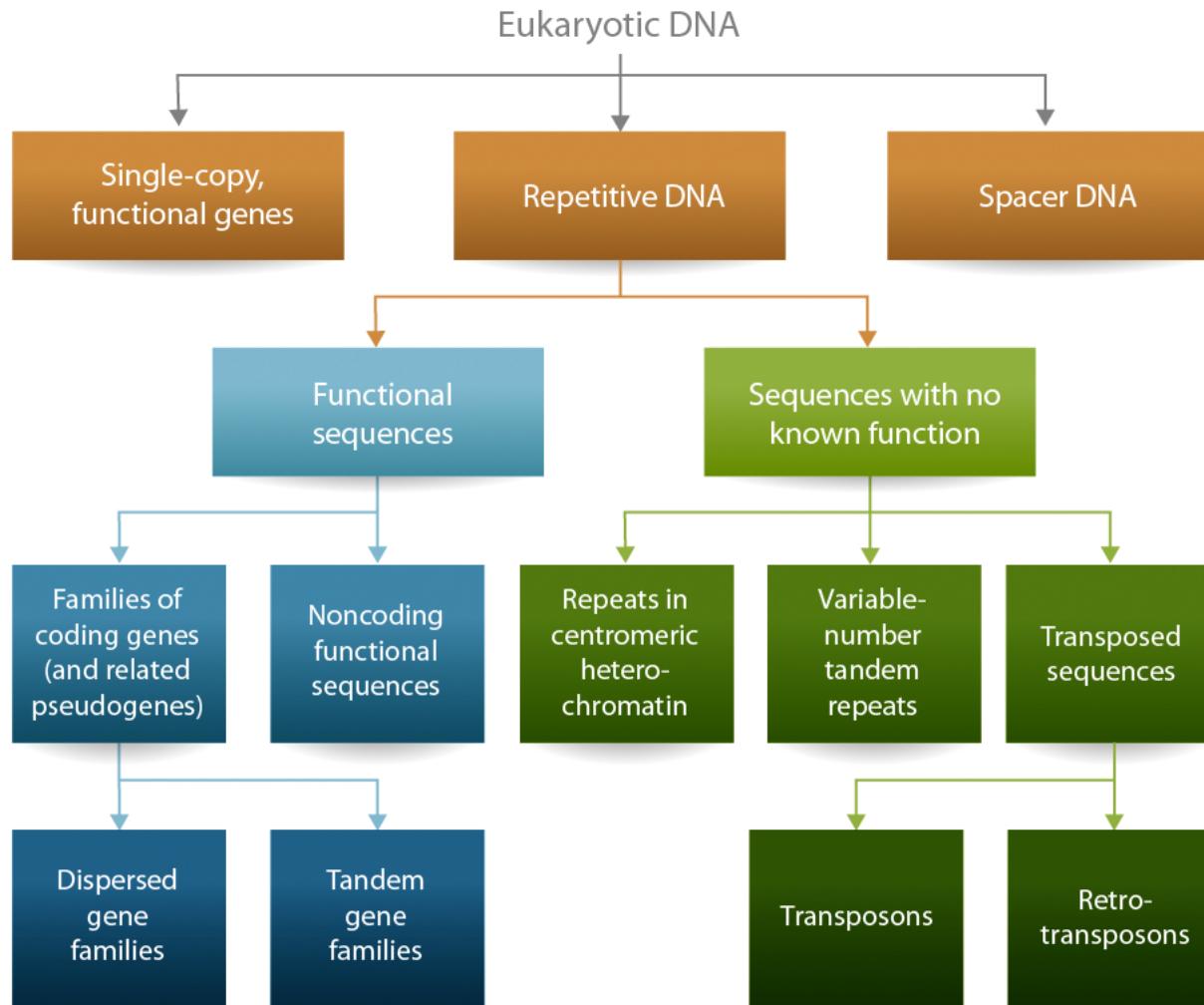
Identifying and classifying genes that produce functional RNAs

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

A Complete Annotation



Annotating Structural Features of Genome Sequences



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

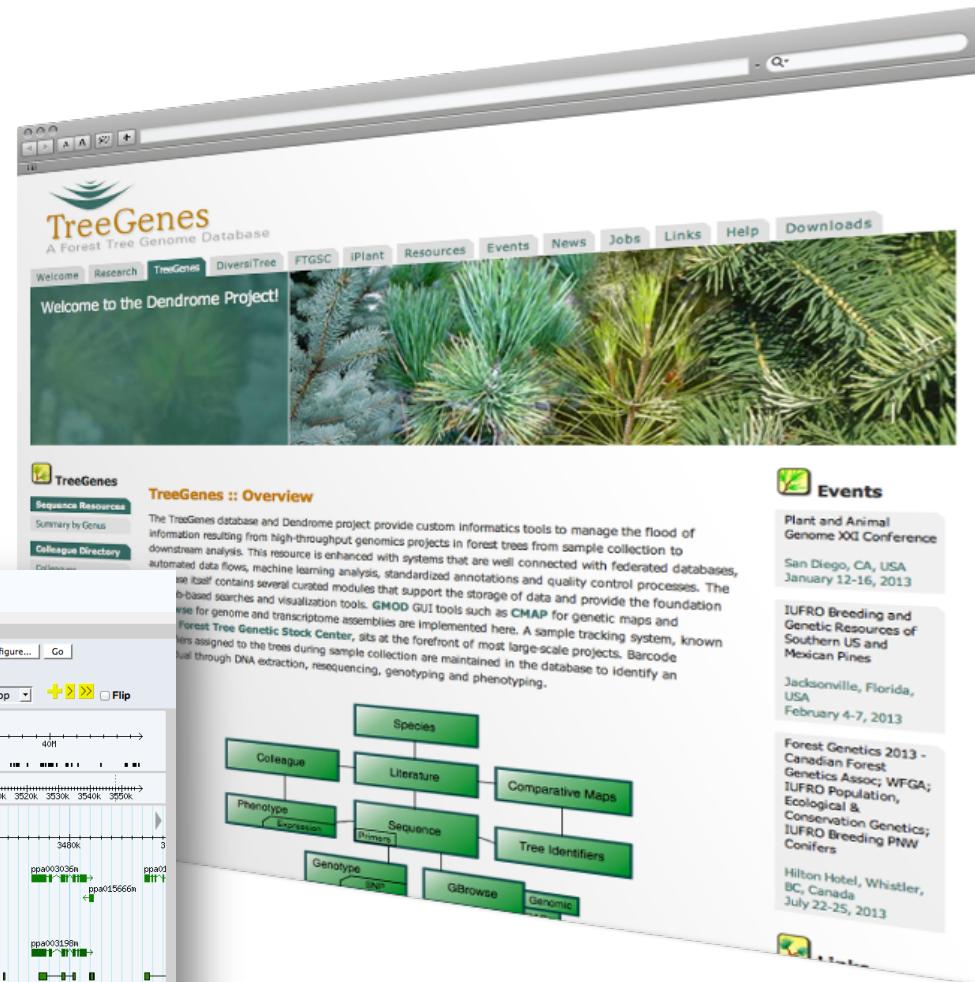
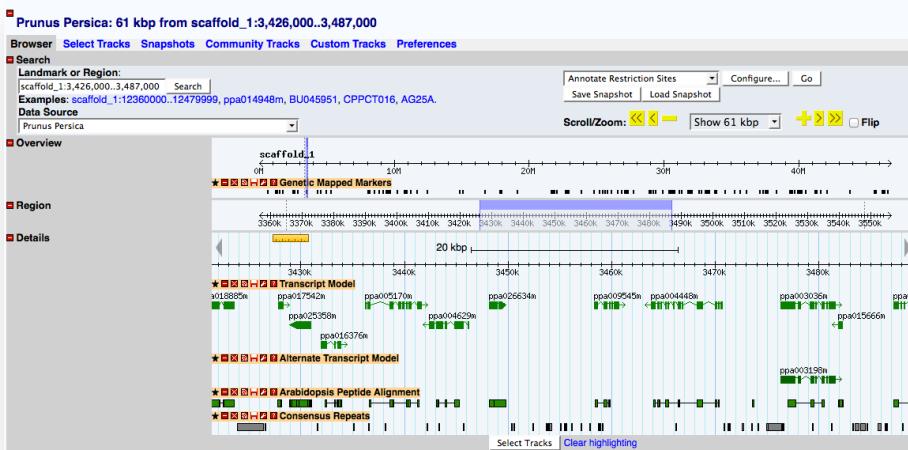
Database Resource Requirements

Project Level Genome Browsers

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

TreeGenes

Genome Database for Rosaceae



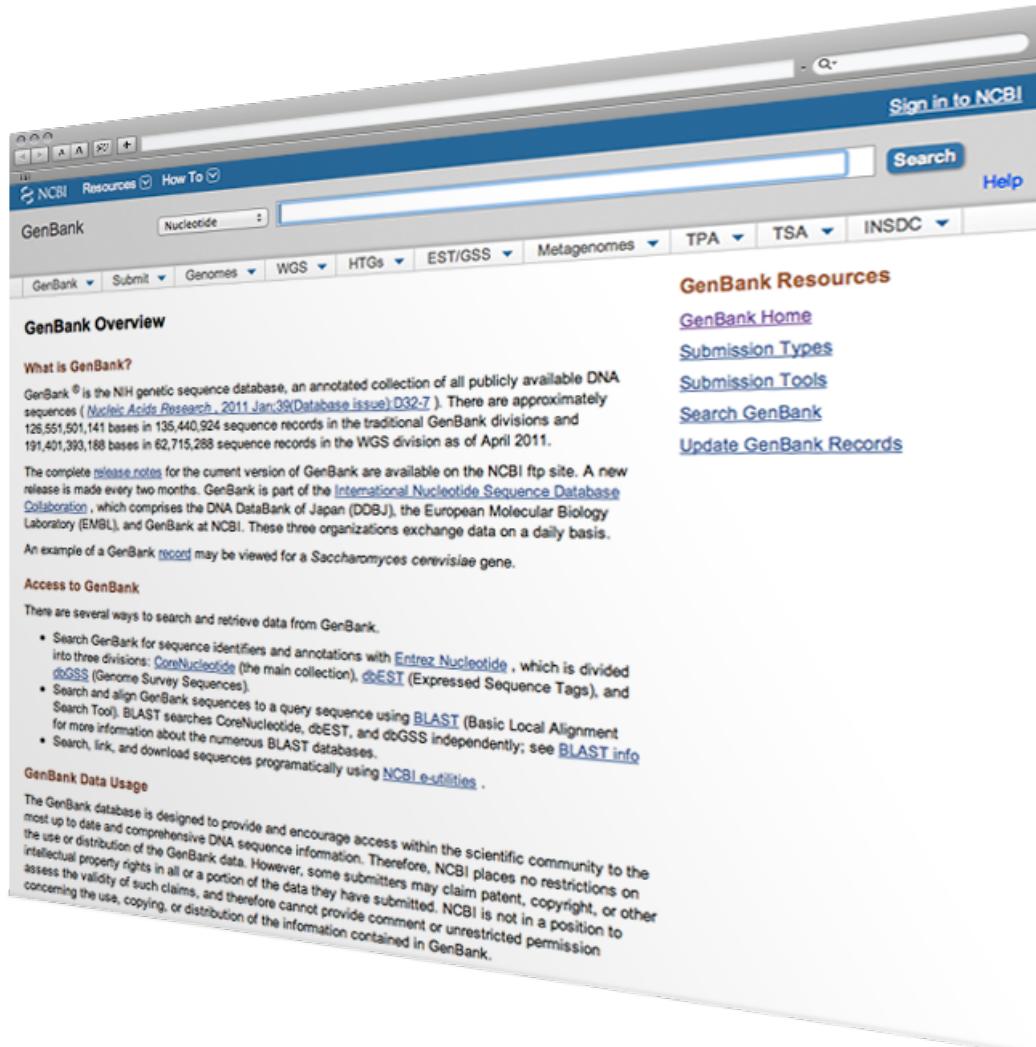
Public Databases

Nucleic Acid Sequence Databases

[ENA / EBI](#) - European Bioinformatics Institute

[DDBJ](#) - DNA Data Bank of Japan

[NCBI /GenBank](#) - National Center for Biotechnology Info



The screenshot shows the NCBI GenBank homepage. At the top, there's a navigation bar with links for "Resources", "How To", "Search", and "Help". Below the navigation bar, there's a search bar and a "Sign in to NCBI" link. The main content area has a blue header "GenBank Resources" with links to "GenBank Home", "Submission Types", "Submission Tools", "Search GenBank", and "Update GenBank Records". The main content starts with a section titled "GenBank Overview" which includes a brief description of what GenBank is, its history, and some statistics about the number of sequence records. It also mentions the International Nucleotide Sequence Database Collaboration. Below this, there's a section titled "Access to GenBank" with instructions on how to search and retrieve data from GenBank. A bulleted list provides several ways to search: Entrez Nucleotide (CoreNucleotide, dbEST, dbGSS), BLAST (Basic Local Alignment Search Tool), and NCBI eUtilities. At the bottom, there's a section titled "GenBank Data Usage" with a detailed explanation of the database's design and the lack of restrictions on its use or distribution.

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2011 Jan 39(Database issue) D32-7). There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,168 bases in 62,715,288 sequence records in the WGS division as of April 2011.

The complete [release notes](#) for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank [record](#) may be viewed for a *Saccharomyces cerevisiae* gene.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

NCBI Entrez

NCBI Entrez - The Life Sciences Search Engine

[NCBI](#) Resources How To

BioProject BioProject [Limits](#) [Advanced](#)

Display Settings: [FASTA](#) Send to:

Pinus lambertiana (sugar pine)
DOE Joint Genome Institute Pinus lambertiana EST project

DOE Joint Genome Institute Pinus lambertiana EST project

Project Data Type: Transcriptome or Gene expression

Attributes: Scope: Monosolate; Material: Transcriptome; Capture: Whole; Method type: Sequencing

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	
OTHER DATASETS	
BioSample	
► SRA Data Details	
Parameter	
Data volume, Gbases	
Data volume, Mbytes	

See Genome Information for Pinus lambertiana

Navigate Across 1 additional project is related by organism.

[FASTA](#)

Pinus taeda temperature-induced lipocalin (TIL) mRNA, complete cds

GenBank: DQ222992.1
[GenBank](#) [Graphics](#) [PopSet](#)

```
>gi|77744880|gb|DQ222992.1| Pinus taeda temperature-induced lipocalin (TIL) mRNA, complete cds
ATGGCTAAAGGAGATTTCGAGGTTGTGAAGGGACTAGACCTGGCAGAGGTACATGGGGGTATGGTATGAA
TCGGCTTCGATGCCGTCCTTCCAGCCAAAGATGGCATTAACCCAGGGCTACGTATTGCGGAATAA
AGACAGCACTGTGATGCTGAACGAGACGTTGGACGGAAAAGCTCCATCGAGGAAGTGCG
TACAAGGTGATCCAAAAGCGAGGATGCCAAATTAAAGTTAAATTATGGTGCCTCCCTTCCCGA
TTATCCAGCTATGGAATTACTGGGTTCTGCTGGATGAAGATTACAGTGGGCTCTCATTGGAGA
ACCTCTCAAACTACCTATGGGCTCTGCAAGGCAACGACGCTAGACGAAGCAATTATAATGCTTA
TTGGAACATGCTGCCAAGAAGGGTAGCACGCTGGGCTCTCATAAAATCAGCAAAATGACGATCCAG
AGACTGAAGCTCCAAGGATAAAGGGGTTCTGGTGGATTAAGGCTCTGCTTGGAAAATAG
```

Welcome to the Entrez cross-database search page

Entrez, The Life Sciences Search Engine

NCBI PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases

PubMed: biomedical literature citations and abstracts
 PubMed Central: free, full text journal articles
 Site Search: NCBI web and FTP sites

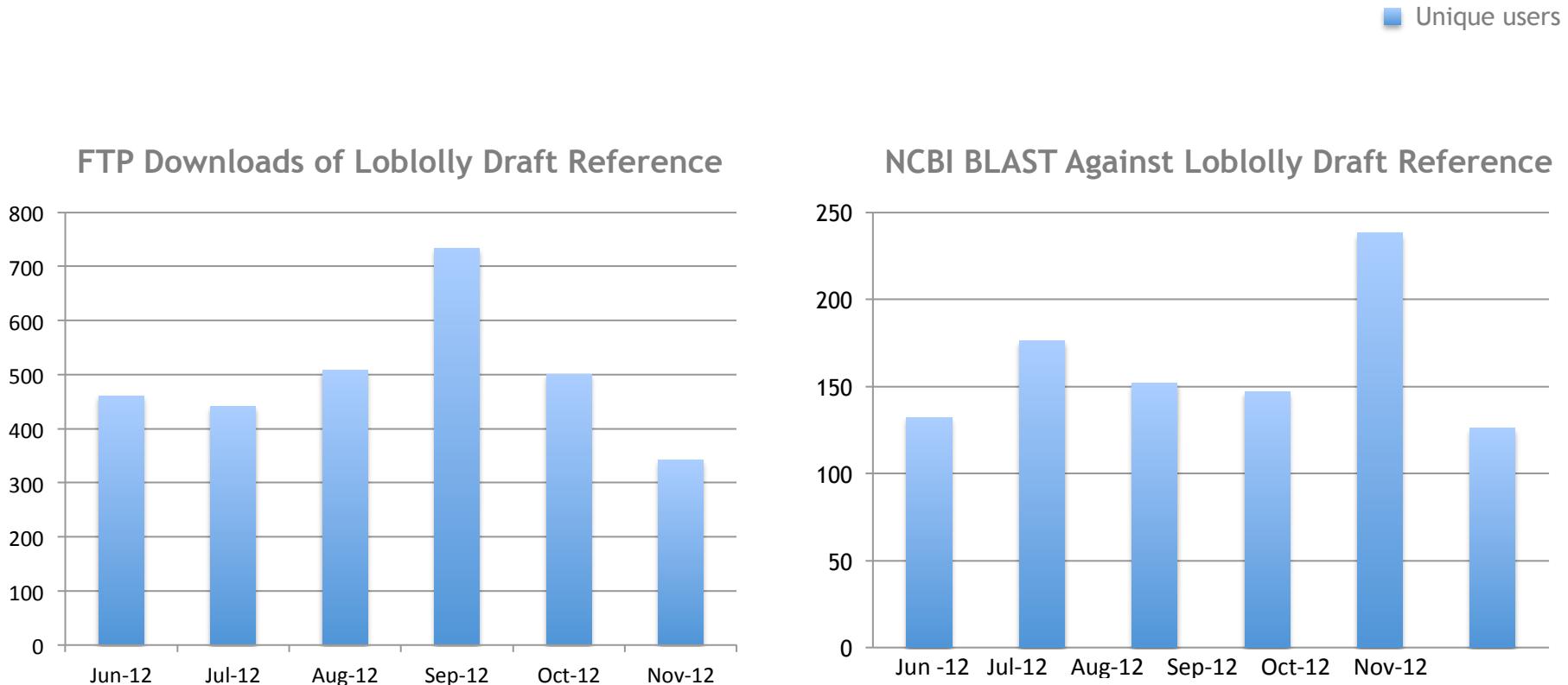
Books: online books
 OMIM: online Mendelian Inheritance in Man

Nucleotides: Core subset of nucleotide sequence records
 EST: Expressed Sequence Tag records
 GSS: Genome Survey Sequence records
 Protein: sequence database
 Genome: whole genome sequences

dbGaP: genotype and phenotype
 UniGene: gene-oriented clusters of transcript sequences
 CDD: conserved protein domain database
 Clone: integrated data for clone resources
 UniSTS: markers and mapping data
 PopSet: population study data sets
 GEO Profiles: expression and molecular abundance profiles
 GEO DataSets: experimental sets of GEO data
 Epigenomics: Epigenetic maps and data sets
 PubChem BioAssay: bioactivity screens of chemical substances
 PubChem Compound: unique small molecule chemical structures
 PubChem Substance: deposited chemical substance records
 Protein Clusters: a collection of related protein sequences

OMIM: online Mendelian Inheritance in Animals

The Pine Reference Sequence in Use



Version 0.6 Release: June, 2012

Version 0.8 Release: January, 2013

Genome Assembly Statistics for Recently Sequenced Species

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