PineRefSeq: Sequencing Strategies in Conifer

Kristian Stevens

Dept. of Evolution and Ecology
University of California, Davis

with
Marc Crepeau and Daniela Puiu,
Ann Holtz-Morris, Maxim Koriabine
Charis Cardeno, Aleksey Zimin
Pieter de Jong, Charles Langley, Steven Salzberg

PAG XXI, 12th January 2013

PineRefSeq

Goal

To provide the benefits of conifer reference genome sequences to the research, management and policy communities.

Specific Objectives

- *Provide a high-quality reference genome sequence of loblolly pine looking toward sugar pine and Douglas-fir.*
- Provide a complete transcriptome resource for gene discovery, reference building, and aids to genome assembly
- Provide annotation, data integration, and data distribution through Dendrome and TreeGenes databases.

The Large, Complex Conifer Genomes Present a Formidable Challenge

- Challenges
  - The estimated 24 Gigabase loblolly pine genome is 8 times larger than the human genome, and far exceeds any genome sequenced to date.
  - Conifer genomes generally possess large gene families (duplicated and divergent copies of a gene), and abundant pseudo-genes.
  - The vast majority of the genome appears to be moderately or highly repetitive DNA
- Approaches to Resolving Challenges
  - Complementary sequencing strategies that seek to simplify the process through use of actual or functional haploid genomes and reduced size of individual assemblies.

Plant Genome Size Comparisons

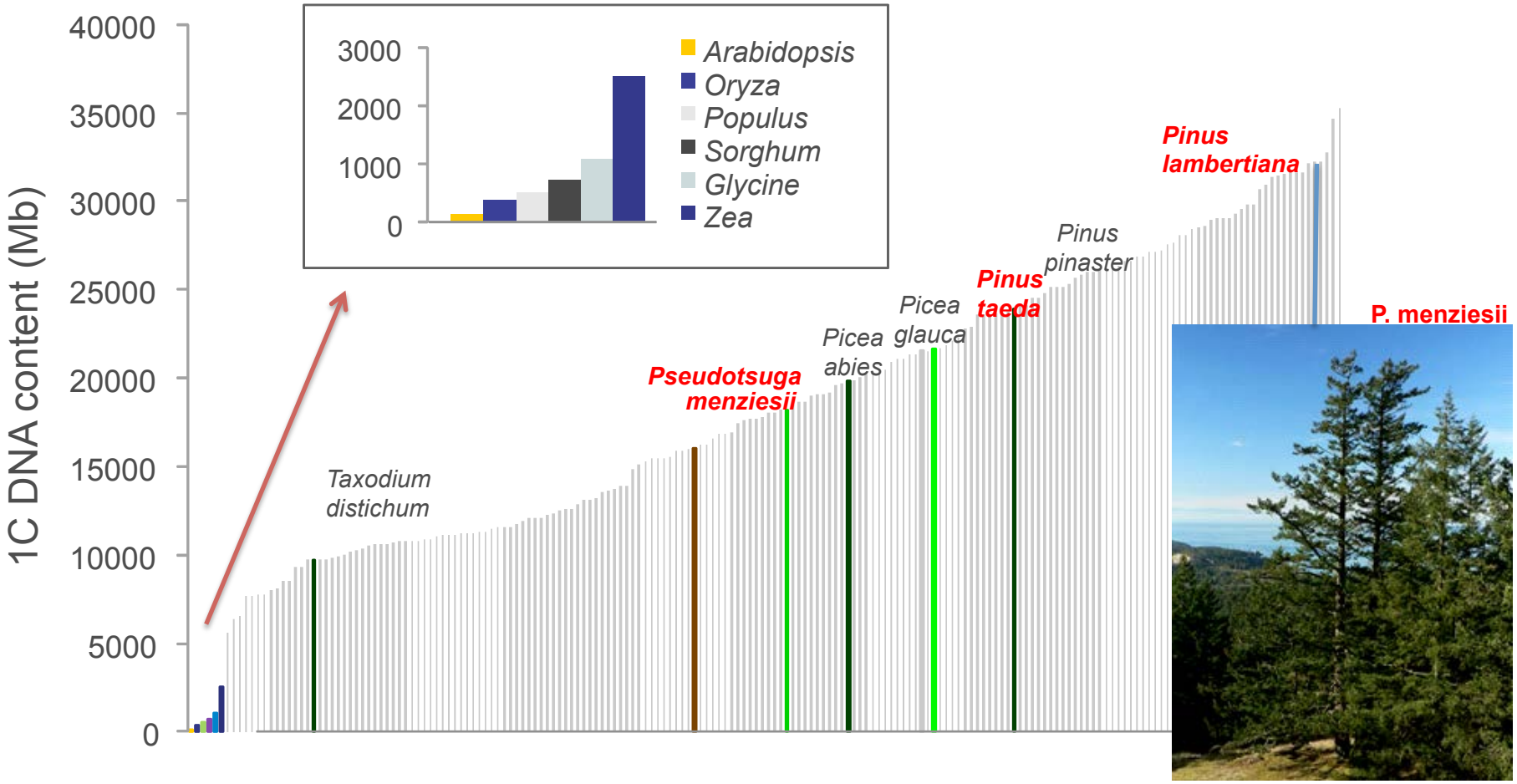


Image Credit: Modified from Daniel Peterson, Mississippi State University

Existing and Planned Angiosperm Tree Genome Sequences

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

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; http://www.phytozome.net/ ; http://asgpb.mhpcc.hawaii.edu/papaya/ ; http://www.amborella.org ;

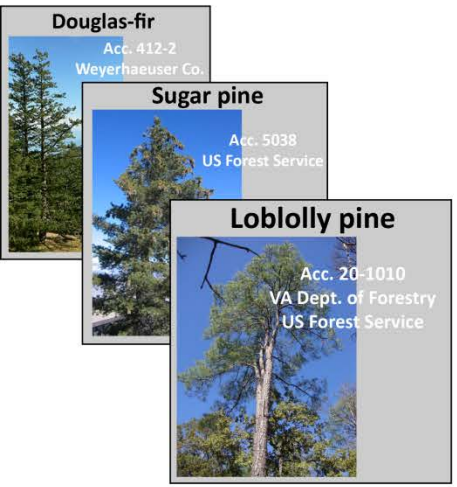
(purple willow – Http://www.poplar.ca/pdf/edomonton11smart.pdf ; Neem - (http://www.strandls.com/viewnews.php?param=5&param1=68

Existing and Planned Gymnosperm Tree Genome Sequences

Species		Genome Size[1]	Number of Genes[2]	Status[3]
Gymnosperms				
Picea abies	Norway Spruce	20,000 Mbp	?	Pending
Picea glauca	White Spruce	22,000 Mbp	?	Pending
Pinus taeda	Loblolly Pine	24,000 Mbp	?	Pending
Pinus lambertiana	Sugar Pine	33,500 Mbp	?	Pending
Pseudotsuga menziesii	Douglas-fir	18,700 Mbp	?	Pending
Larix sibirica	Siberian Larch	12,030 Mbp	?	Pending
Pinus pinaster	Maritime Pine	23,810 Mbp	?	Pending
Pinus sylvestris	Scots Pine	~23,000 Mbp	?	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 publically released tree genomes. Conifer genomes will also be posted here as they are completed.

Elements of the Conifer Genome Sequencing Project



WGS (1N & 2N)

2N

1N

Sequencing

UC Davis — UC DAVIS Evolution & Ecology
-Libraries: short/long inserts
-Sequencing
= GAIIX (overlapping reads)
= HiSeq (125 bp reads)
-Data curation and fosmid assembly

Fosmid Cloning

CHORI BAC/PAC
- 37 kb clones
- >2X coverage
- Pool size ≥ 500
- diTags
- Illumina library construction

Assembly

Maryland-JHU Genome Assembly Group
-Assembler development (MSR-CA)
-Assembly
=WGS
=Fosmid
=Meta

Genetic Maps

UC Davis
-SNP calls from scaffolds and genetic mapping to order scaffolds

Transcriptome

IU-TAMU-UC Davis
-Multiple tissues/conditions
-Library development
-Sequencing (Roche 454/Illumina)
-Assembly

Annotation

UC Davis-WSU
-Functional Annotation
-Comparative genomics
-Anchoring to existing genetic maps
-Integration of genome and transcriptome resources

Database-Dendrome

UC Davis
-Data organization and distribution
-Ontology Development (PO&TO)
-Community Annotation (GenSAS)
-Web Services (SSWAP)
-Analysis Pipeline Integration (iPlant)

Two Approaches to Conifer Sequencing

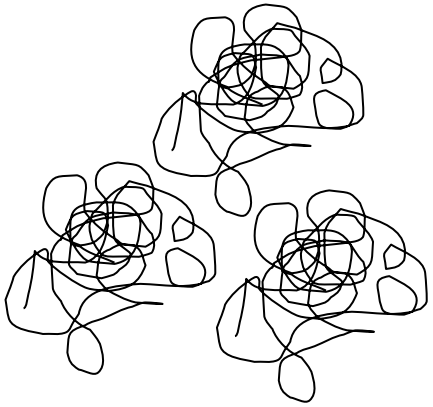
- Whole Genome Shotgun (WGS) sequencing of
  - Haploid Megagametophyte: Goal – deep representative short insert libraries from a single haploid ($1N$) segregant. Haploid genome significantly improves sequence assembly
  - Diploid Parental Genotype 20-1010
    (Released to public for this project)

botit.botany.wisc.edu

- Direct Sequencing of Pooled Fosmid libraries
  - Pools of multiple 38.5 kbp *P.taeda* fosmid clones that are well within available assembler's specs.
  - Key is to tune complexity of individual pools for both economy and assembly quality
  - Small enough to be haploid, facilitating assembly ($2N$ library but $1N$ pool).

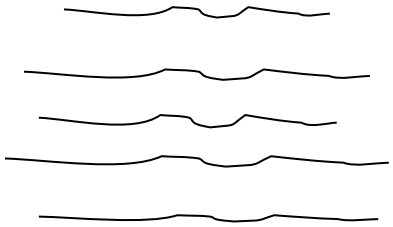
Whole Genome Shotgun Reads

Genomic DNA



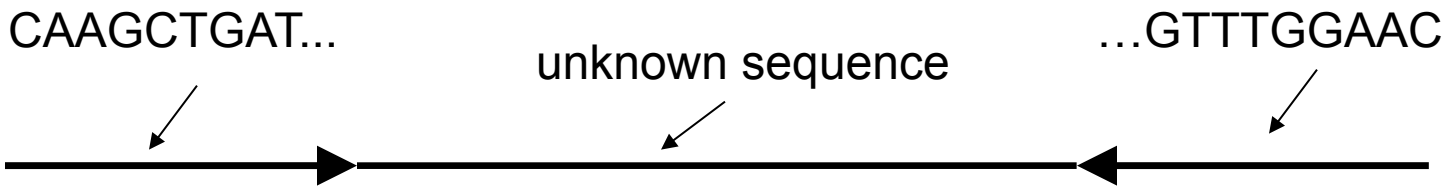
fragmented & size selected

fragments of 200 - 10,000 bases



sequence the ends

pairs of reads of 100 – 250 bases each



Technology for De Novo Sequencing of the Conifer Genomes

Parallel and Complementary Approaches

Whole Genome Shotgun Sequencing (Haploid/Diploid)

- Illumina GAIIx
- Illumina HiSeq

Sequencing of Pooled Fosmid Clones (Diploid)[1]

- Illumina GAIIx
  - Max Output: 95 Gigabases
  - Max. paired end reads - 640 million
  - Max. Read Length – 2 x 150 bp
- Illumina HiSeq
  - Max Output: 300 Gigabases
  - Max. paired end reads - 3 billion
  - Max. Read Length – 2 x 100 bp

[1] Effectively haploid

Data current as of 3/2012 (Illumina)

Megagametophyte Whole Genome Shotgun (M-WGS)

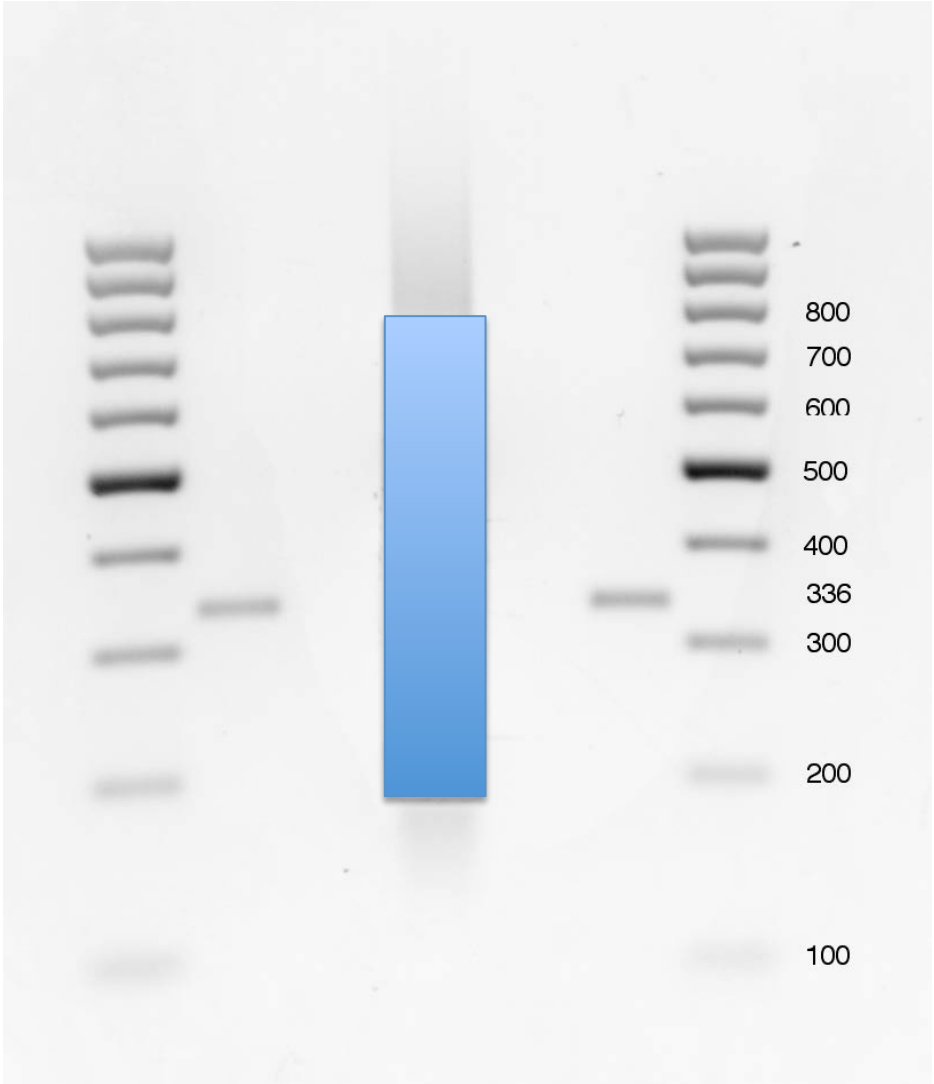
- Not enough haploid DNA in a megagametophyte to implement a complete list of WGS ingredients.
  - Obtain DNA for longer insert linking libraries (> 1kbp) from diploid needle tissue.
  - Prepare only short insert Illumina libraries from megagametophye tissue.
  - Library complexity is still a challenge
    - Many short libraries will be needed

***P. taeda* 2011 crop**

N	54
mean	1361 ng
st. dev.	675 ng
min	580 ng
max	3560 ng

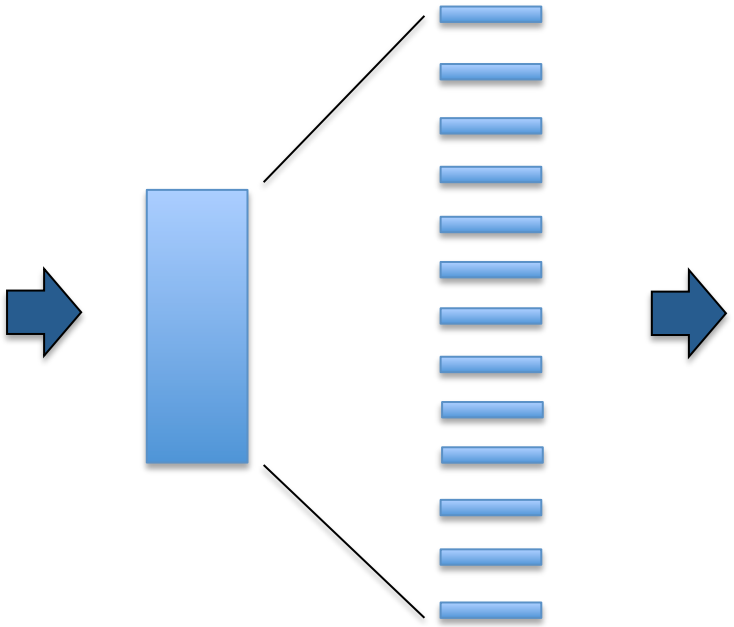
M-WGS Short Insert Libraries

DNA Fragmentation and Partitioning



[Left] The fragmented meg DNA sample is run on an agarose gel.

[Right] A target size range is extracted and partitioned for tight c.v. on the insert size distribution.

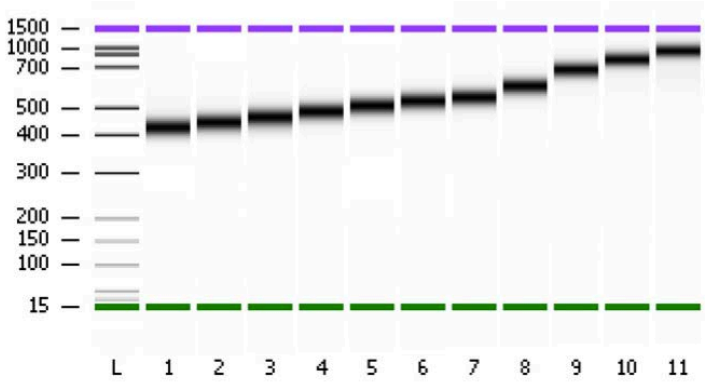


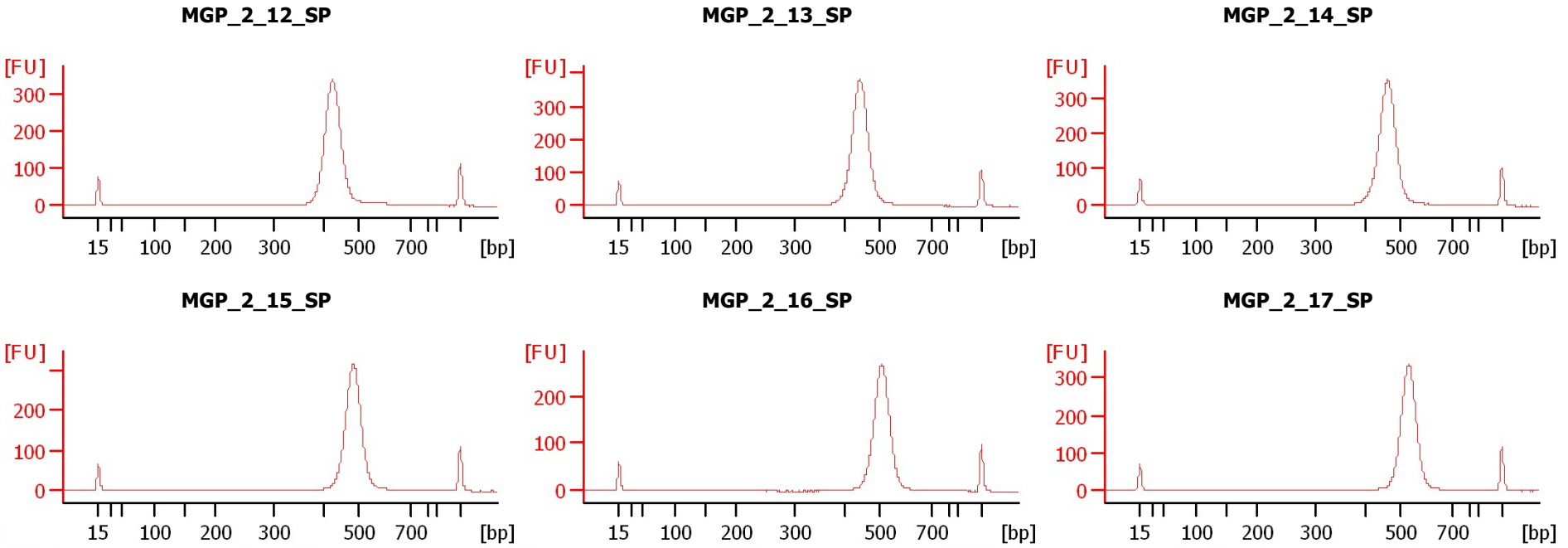
M-WGS Short Insert Libraries

Preliminary QC and Size Selection

Each DNA sample is then run on an Agilent Bioanalyzer to determine a preliminary estimate of insert size and coefficient of variation.

If within spec, selected DNA samples are converted into Illumina libraries





M-WGS Short Insert Libraries

Library QC and Titration

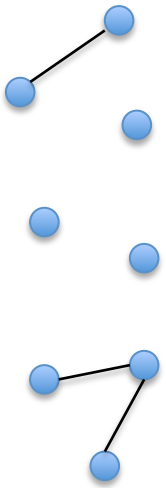
- Libraries are subsequently QCed on a MiSeq

Library ID	ecies	Median	C.V.
MGP_2_3_SP	*Pinus lambertiana*	180	0.05
MGP_2_4_SP	*Pinus lambertiana*	191	0.05
MGP_2_5_SP	*Pinus lambertiana*	201	0.04
MGP_2_6_SP	*Pinus lambertiana*	212	0.05
MGP_2_7_SP	*Pinus lambertiana*	226	0.05
MGP_2_8_SP	*Pinus lambertiana*	240	0.04
MGP_2_9_SP	*Pinus lambertiana*	254	0.04
MGP_2_10_SP	*Pinus lambertiana*	268	0.04
MGP_2_11_SP	*Pinus lambertiana*	281	0.04
MGP_2_12_SP	*Pinus lambertiana*	296	0.04
MGP_2_13_SP	*Pinus lambertiana*	313	0.04
MGP_2_14_SP	*Pinus lambertiana*	329	0.04
MGP_2_15_SP	*Pinus lambertiana*	347	0.04
MGP_2_16_SP	*Pinus lambertiana*	365	0.04
MGP_2_17_SP	*Pinus lambertiana*	380	0.04
MGP_2_18_SP	*Pinus lambertiana*	395	0.04
MGP_2_500_SP	*Pinus lambertiana*	447	0.04
MGP_2_600_SP	*Pinus lambertiana*	546	0.05
MGP_2_700_SP	*Pinus lambertiana*	633	0.05
MGP_2_800_SP	*Pinus lambertiana*	714	0.07

Metric	Definition
Sample Name	Sample name from the sample sheet
Clusters	Number of clusters sequenced for this sample
Clusters %	Percentage of successfully indexed clusters from this sample
% PF	Percentage of clusters for this sample that passed filters
% Aligned R1	Percentage of clusters for which read 1 successfully aligned
% Aligned R2	Percentage of clusters for which read 2 successfully aligned
Length Median	Median fragment length for this sample
Length Min	Low percentile (corresponding to 3 standard deviations from the median) of fragment lengths for this sample
Length Max	High percentile (corresponding to 3 standard deviations from the median) of fragment lengths for this sample
Mismatch R1	Mismatch rate for this sample in read 1
Mismatch R2	Mismatch rate for this sample in read 2

M-WGS Libraries

How deep to sequence a library?



- Determine vertices
  - k length prefixes of each DNA molecule (paired end read) are concatenated
- Determine edges
  - An edge between two vertices if they differ by one or fewer sites
- Identify and count connected components.
  - a.k.a single linkage clustering
- Compute average multiplicity
  - Number of paired end reads / number of clusters
  - Target < 1.25X

Notable Assemblers for Illumina Data

- MaSuRCA
- Allpaths-LG
- SOAPdenovo
- ABySS
- Velvet
- Contrail
- SGA

k-mers

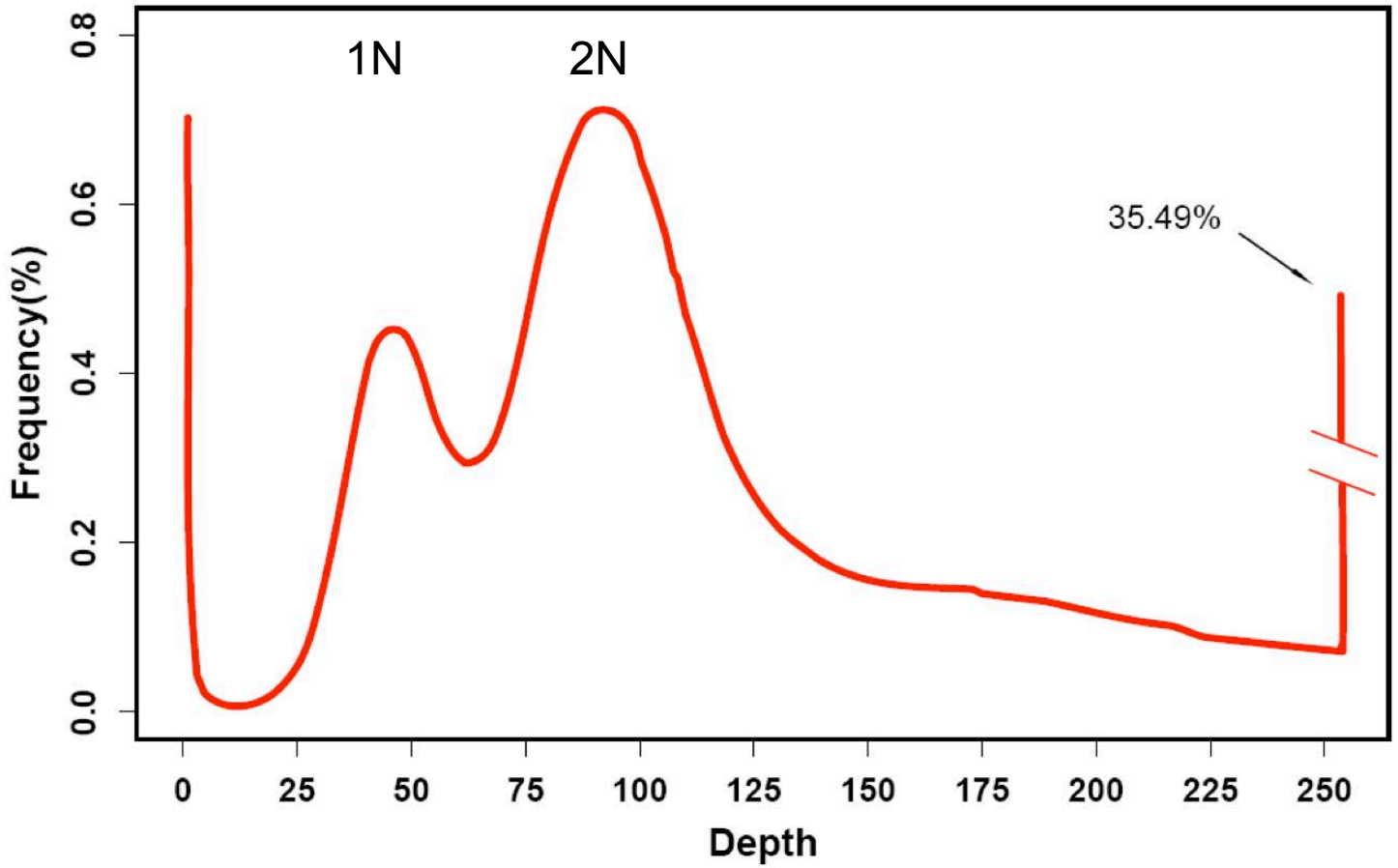
Query: Does a distinct length *k* string occur in the genome?

- Experimentally, this is often answered with hybridization
- To answer it of the WGS reads we will first chop the sequence up into all substrings of length *k*
- Each length *r* read becomes *r – k + 1* strings of size *k*

- We choose a *k* that best supports the query above allowing specific locations on the genome to be queried.
- From *k*-mers alone we can estimate genome size, repeat content, and even assemble whole genomes (SBH).

17-Mer Complexity Analysis

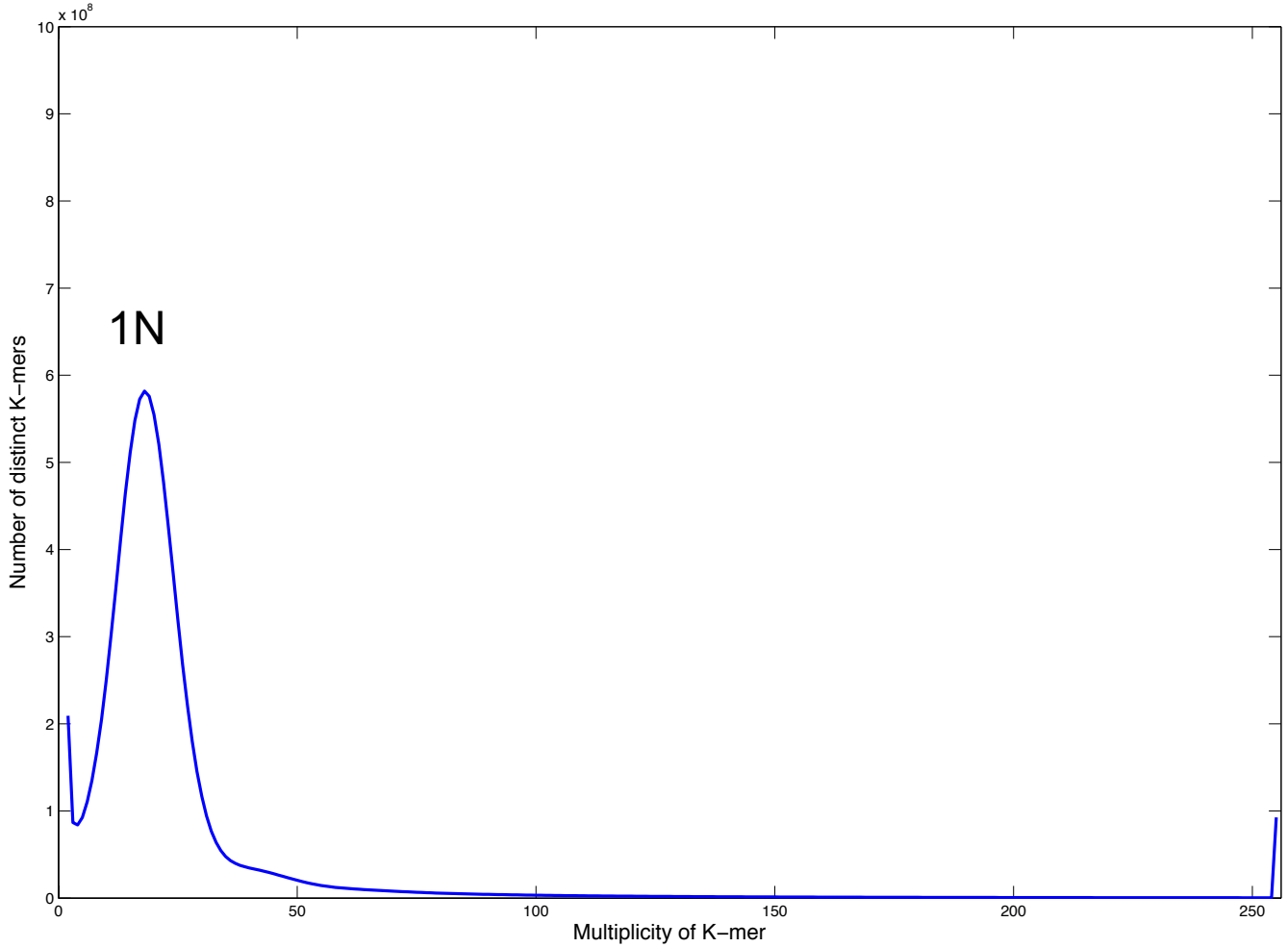
Pacific oyster - *Crassostrea gigas*



(Zhang et al. 2012)

31-Mer Complexity Analysis

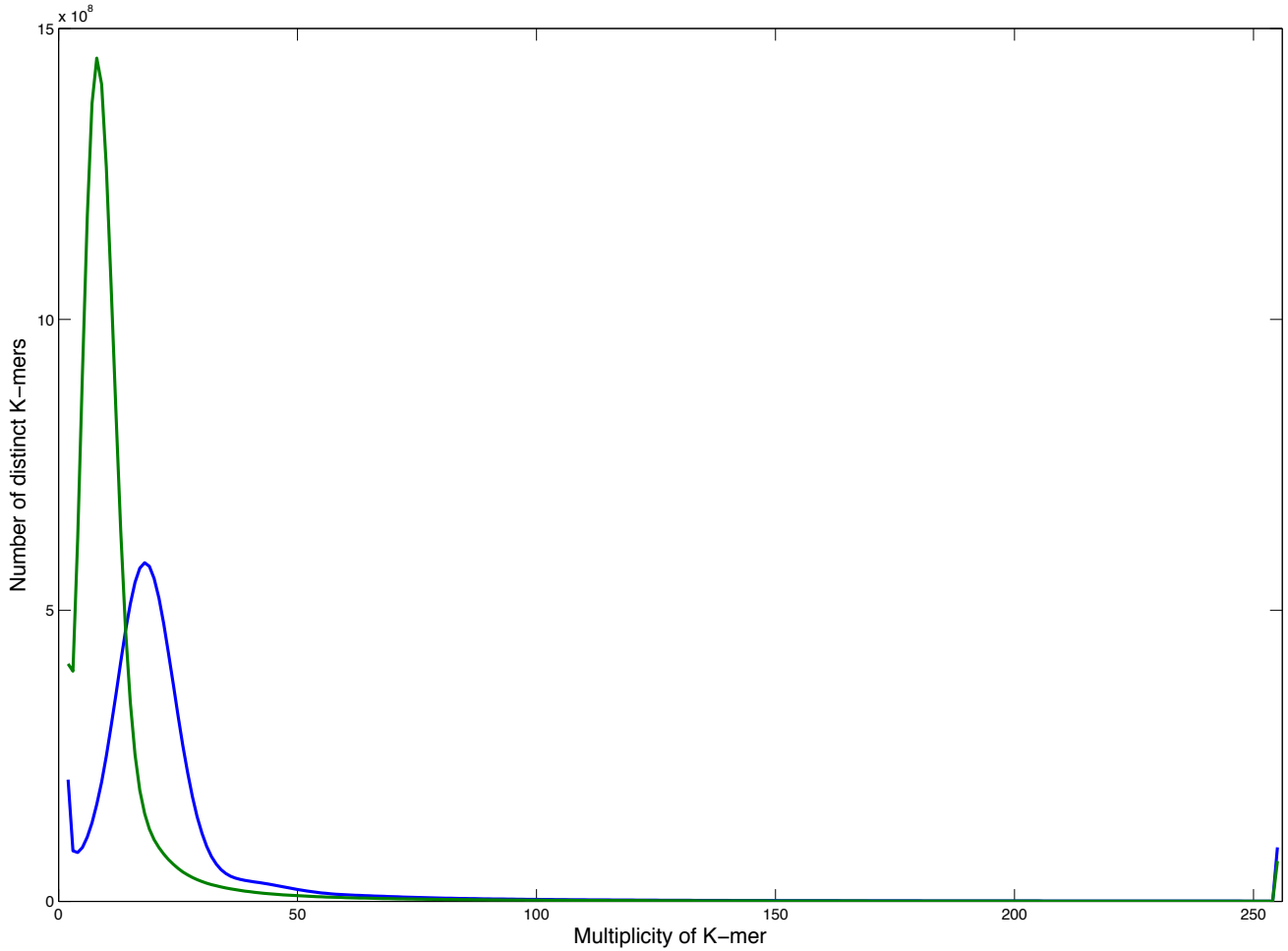
Loblolly pine – *Pinus taeda* – *25x HiSeq*



31-Mer Complexity Analysis

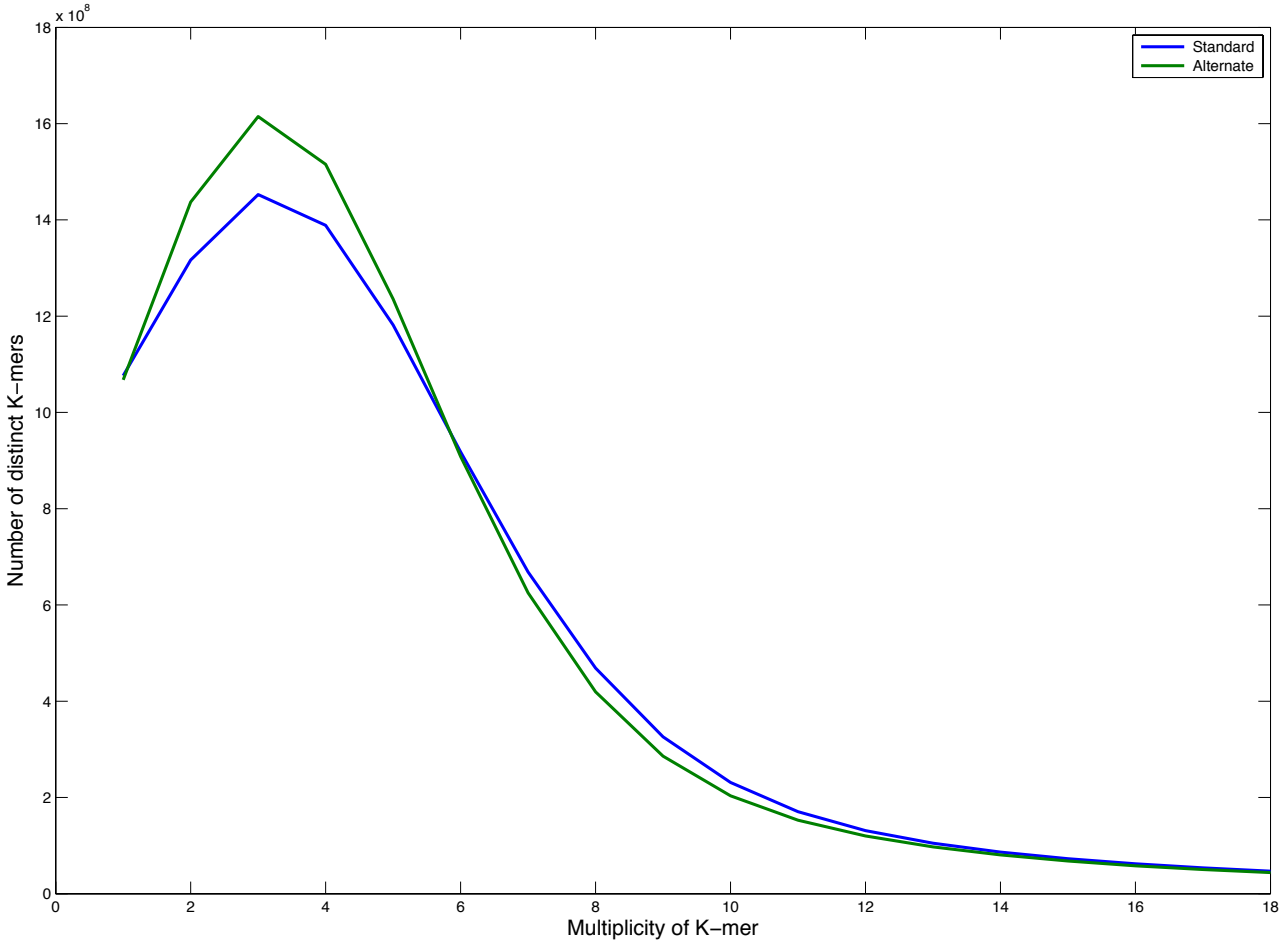
Loblolly pine – *Pinus taeda – 25x HiSeq*

Sugar pine – *Pinus lambertiana – 10x HiSeq + MiSeq + GA2x*



31-Mer Library Construction Analysis

Sugar pine – *Pinus lambertiana – HiSeq + MiSeq + GA2x*



A k-mer Genome Size Estimate

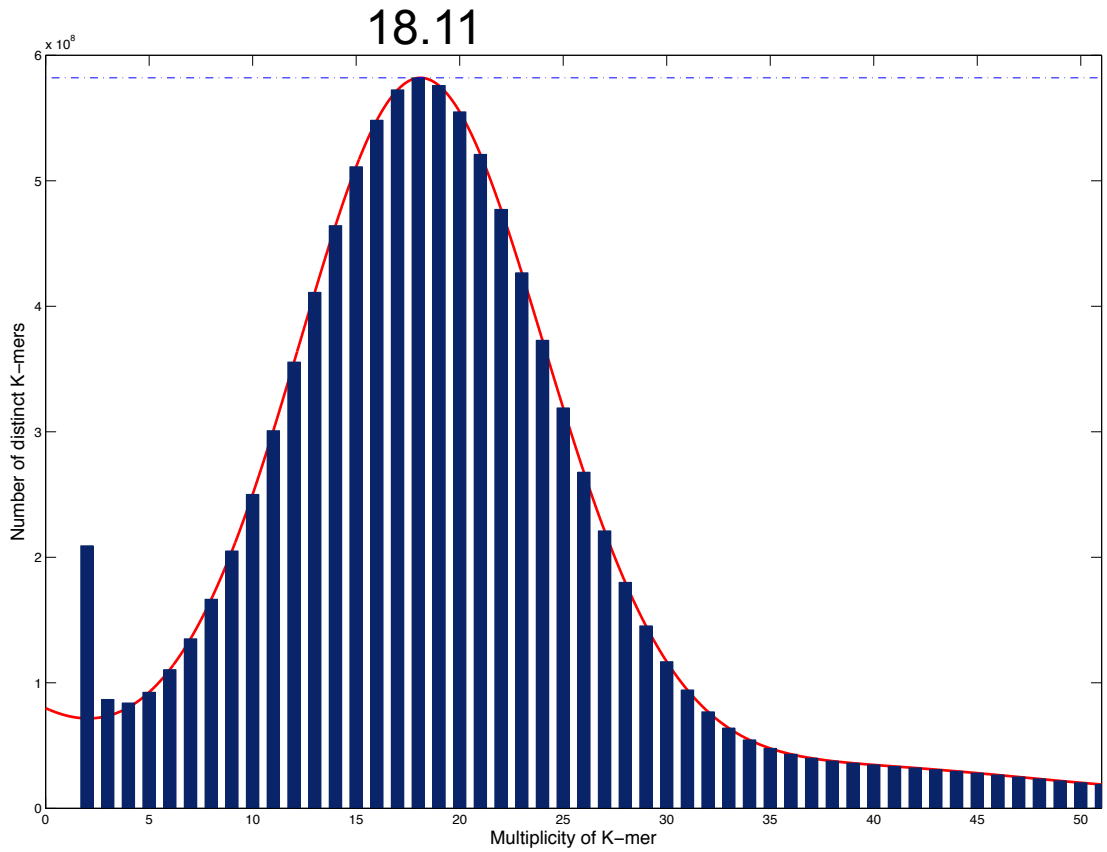
P. taeda genome size ≅ total k-mers in genome

total k-mers in *P. taeda* genome ≅

$$\frac{\text{total k-mers in } \textit{P. taeda} \text{ reads}}{\text{expected number of times a genomically unique k-mer is observed in the reads}}$$

Expected unique k-Mer multiplicity

We will use the mode of a fitted PDF



k-mer Genome Size Estimates

Loblolly pine *Pinus taeda:*

31-mers total: 3.736×10^{11}

Expected k-mer depth: 18.11

Estimated genome size: 20.63 GB

High Copy 31-mers

1.09% of distinct 31-mers

33% of all 31-mers

24-mers total: : 4.092×10^{11}

Expected k-mer depth: 19.79

Estimated genome size: 20.68 GB

Sugar pine *Pinus lambertiana:*

31-mers total: 2.776×10^{11}

Expected k-mer depth: 8.12

Estimated genome size: 34.19 GB

High Copy 31-mers

0.35% of distinct 31-mers

33% of all 31-mers

24-mers total: 3.031×10^{11}

Expected k-mer depth: 8.89

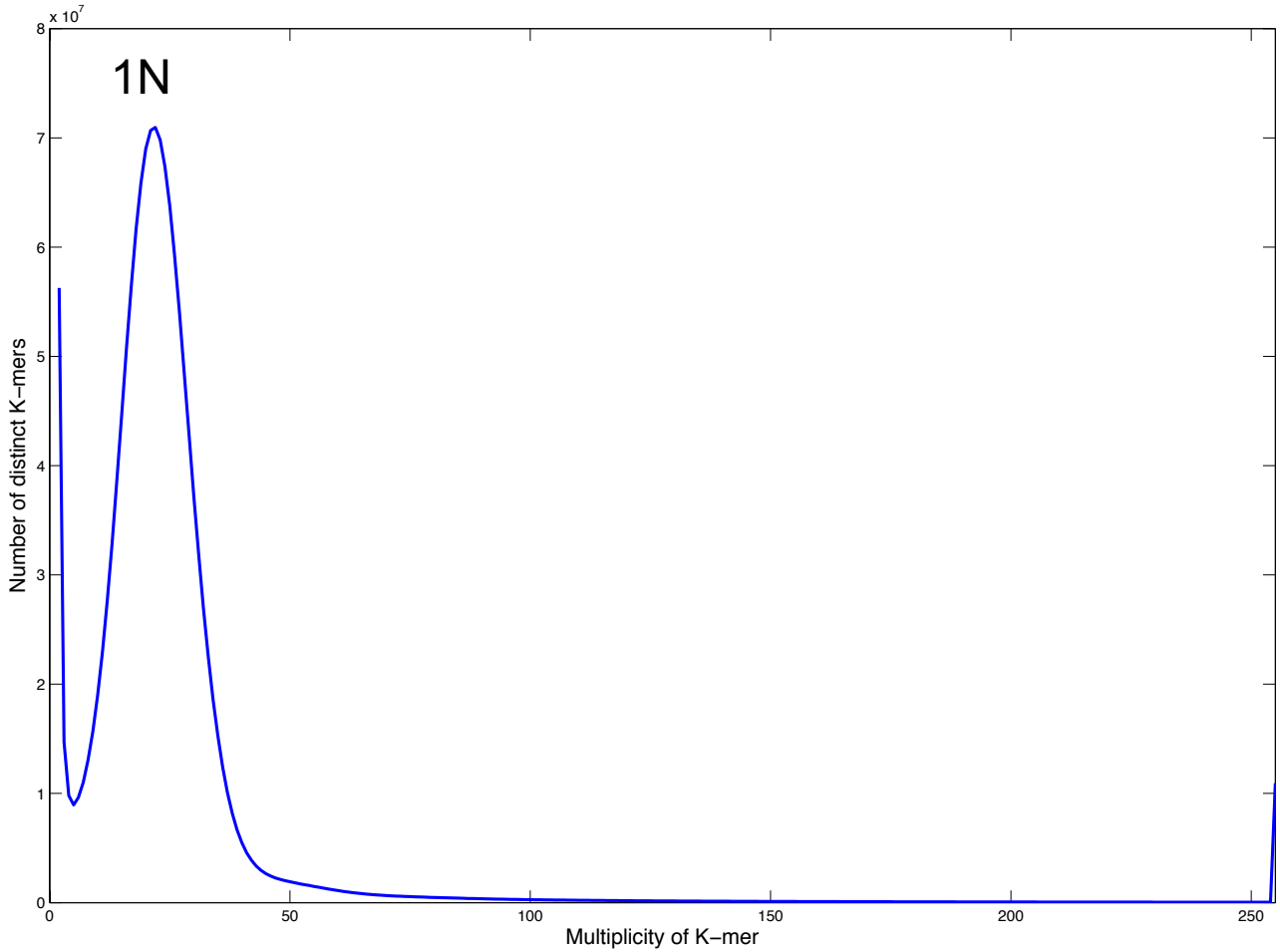
Estimated genome size: 33.98 GB

P. taeda Version 0.8 Library Statistics

- Haploid short insert libraries
  - 10 short insert libraries 200 - 640bp
  - 1.4Tbp raw total sequence
  - 60 fold coverage
- Diploid jumping libraries
  - 47 jumping libraries 1300 – 5500bp
  - 280Gbp raw total sequence
  - 12 fold coverage
- 13 Fosmid DiTag Libraries 38,500bp

79-mer Complexity Analysis

Loblolly pine – *Pinus taeda* – (62x *HiSeq + MiSeq + GA2x* coverage)



Early Access *P. taeda WGS V0.6*

- Approximately 35X coverage
- Total Sequence: 18,321,727,393 bp
- Total contig sequence: 14,606,783,345 bp
- N50 1199bp (9.16 Gbp is contained in contigs of 1199 bp or longer)
- Total scaffold sequence (with imputed gaps): 18,428,460,141bp
- N50 1230bp (9.21 Gbp is contained in scaffolds of 1230 bp or longer)
- Degenerate contig sequence 3.8Gb

Fosmid Pools

A Molecular Approach to Complexity Reduction

Reduced Complexity from E.coli clone pools

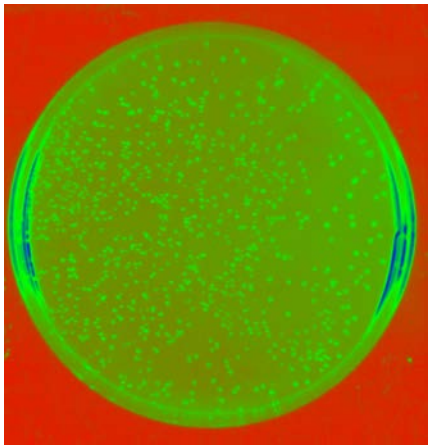
- Traditionally clone tiling paths were used to create good assemblies.
- Recently clone pools have been used to derive long-haplotype information.
- Ideally: larger clones (BACs) are more desirable, more likely to span repeats
- Economically: much cheaper to create large fosmid libraries using sheared DNA

Fosmid Pooling:
Genome partitioning for reduced assembly complexity

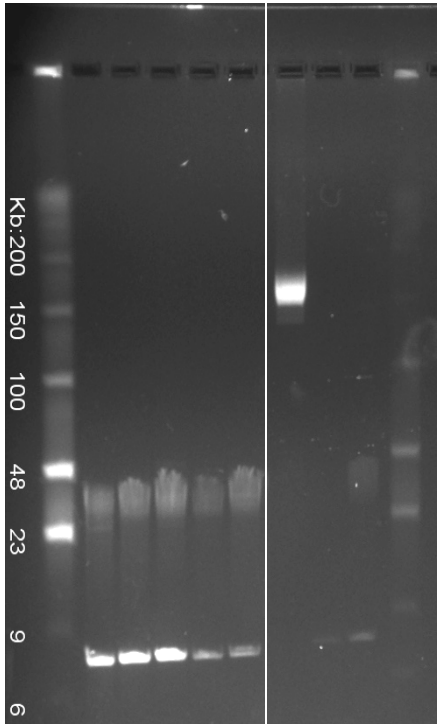
- The immense and complex diploid pine genome can be *economically* and *efficiently* partitioned into smaller, functionally haploid, pieces using pools of fosmid clones.
- Fosmids in a pool should have a combined insert size far less than a haploid genome size; This assures a haploid genome representation.
- The sequence data obtained from a single fosmid pool may be up to 80 X deep.
- The sequence data obtained from a pool must be screened for vector and E. coli contamination

Fosmid Pool DNA Preparation

- Genomic DNA extraction
- Shearing (to average ~40 kbp)
- Size-purification (pulsed-field gel electrophoresis)
- Ligation to excess vector (dephosphorylated ends)
- Packaging (extracts from mcrA, B, C strains)
- Determine titer of the "particle library"
- Create E.coli colonies
- Isolate DNA from colonies
- Quality Assessment: Quantitation & Q-PCR (TAQMAN assays) to determine E.coli & fosmid vector contamination



Fosmid colonies: Documented & counted prior to pooling



Fosmid pools: DNA digested (PI-SceI); Lane 6: Undigested pool

Fosmid Pool Sequencing Pipeline

- Fosmid pool DNA received from CHORI
- Quantify DNA

Short Insert Library

- Aliquot 5 ul DNA
- Sonicate to fragment DNA
- End repair fragments
- A-tail fragments
- Ligate Illumina multiplex adapter
- Size select adapter-ligated fragments by agarose gel electrophoresis
- QC and quantitate using Agilent Bioanalyzer
- Enrich 10 ng size selected fragments using 10 cycles PCR
- QC and quantitate enriched library using Agilent Bioanalyzer
- Sequence

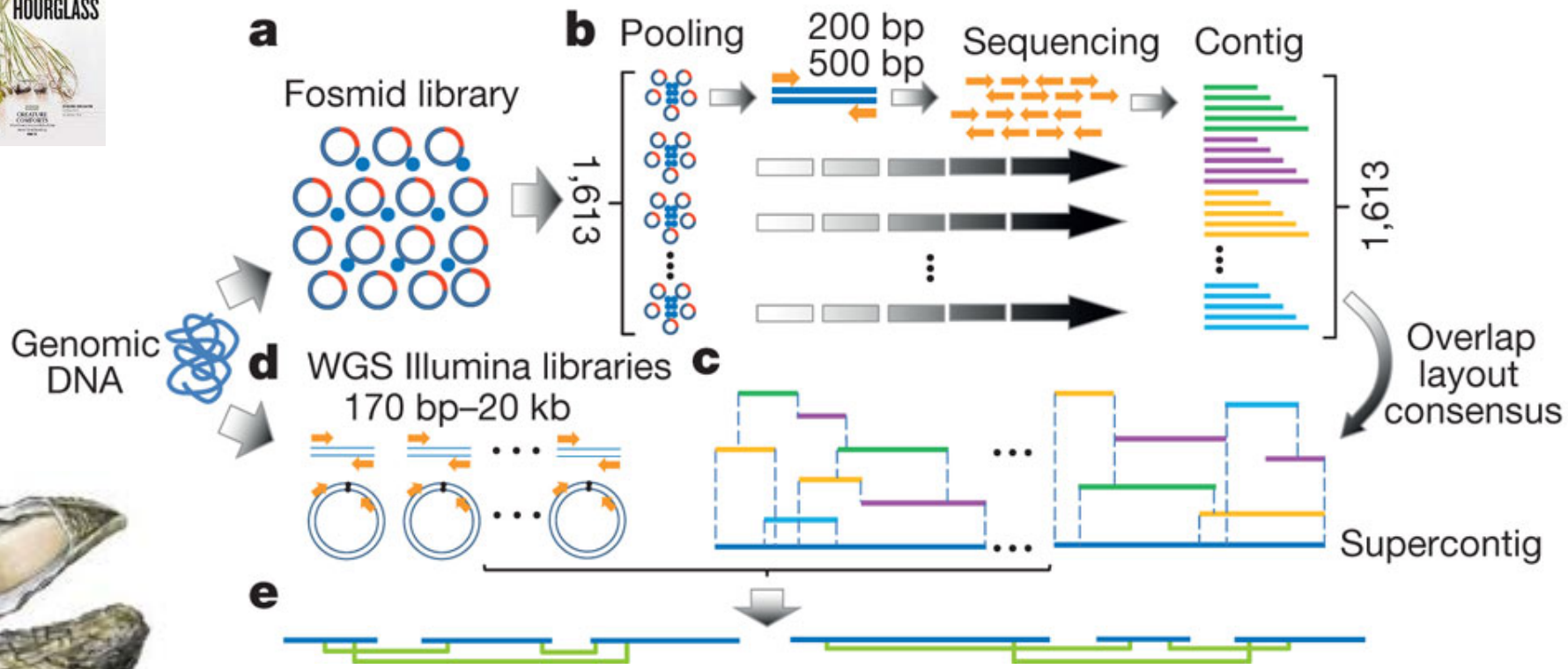
Long Insert Library

- Pool DNAs to create pool of pools
- Aliquot 10 ul DNA
- Fragment DNA using HydroShear
- End repair and biotinylate fragments
- Size select by agarose gel electrophoresis
- QC and quantitate using Agilent Bioanalyzer
- Circularize size selected fragments
- Digest un-circularized DNA
- Sonicate to fragment circularized DNA
- Bind biotinylated fragments to streptavidin beads
- End repair fragments
- A-tail fragments
- Ligate Illumina multiplex adapter
- Enrich adapter-ligated fragments using 18 cycles PCR
- Remove enriched library from beads
- Size select enriched library by agarose gel electrophoresis
- Sequence

Fosmid Pools

Pacific oyster - *Crassostrea gigas*

(Zhang et al. Oct 2012)

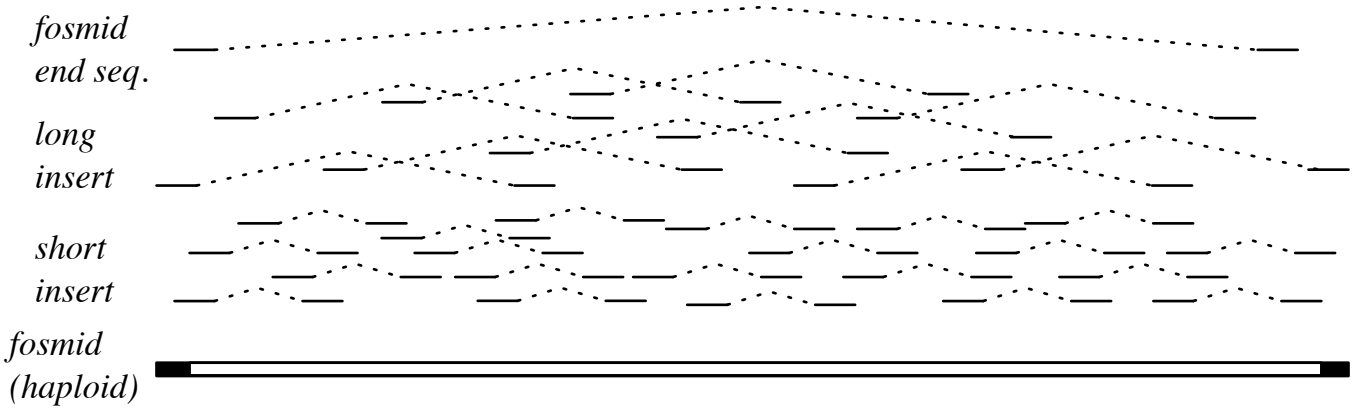


Fosmid Pools

Strategy made possible by cheap sequencing.

- Pacific Oyster 0.56 Gbp Genome
  - Fosmid Pool Size: 90
  - Number of Pools: 1,613
  - Pool Coverage: 10x
  - Sequence Coverage per Pool 60x
  - Total fragment library coverage: 360Gbp (600x)
- Loblolly Pine 24Gbp Genome
  - Number of pools with 90 fosmids per pool: 69,000
  - Total fragment library coverage: 14,400Gbp

Fosmid Sequence Components



- Haploid fosmids with vector tagged ends
- Primary coverage from short insert libraries
- Additional coverage from long insert libraries from equi-molar pool of pools.
- Fosmid end sequences (diTags) link ends of the assembly and count fosmids in a pool

Fosmid Pools

Determining the Best Assembler for the Job

Assembly results for a relatively large pool of approximately 500 *P. taeda* fosmids

			quartiles				
Assembler	Stat	Count	Q1	Q2	Q3	N50	Sum
Allpaths-LG	scf	987	2499	7781	30271	26298	14×10^6
	ctg	1524	2355	6031	12509	10324	14×10^6
	scf30K+	248	33595	35682	38361	30114	9×10^6
MSR-CA	scf	2162	506	1375	9224	14753	15×10^6
	ctg	3519	503	1339	5000	6826	14×10^6
	scf30K+	136	32603	35087	38119	30147	5×10^6
SOAP	scf	3251	123	185	495	33389	15×10^6
	ctg	23873	76	175	348	1515	15×10^6
	scf30K+	322	33907	35766	38683	33389	12×10^6

Fosmid Pools

Increasing the Pool Size

- Five nested fosmid pools
  - 1, 2, 4, 6, and 8 unit pools of approx 600 fosmids
- Assembled with SOAP denovo
- Scaffolds larger than 30kbp are reported below

estimated pool size	elem	min	max	mean	sum	end verified	coverage
600	567	30030	60754	35513	20135877	343	87.2%
1200	1094	30029	66421	35567	38910427	694	84.2%
2400	1465	30001	78520	35093.85	51412495	740	55.6%
3600	2598	30003	74253	35330.12	91787645	1343	66.2%
4800	3305	30015	73986	35309.59	116698186	1676	63.1%

Fosmid Pools

Increasing the Pool Size

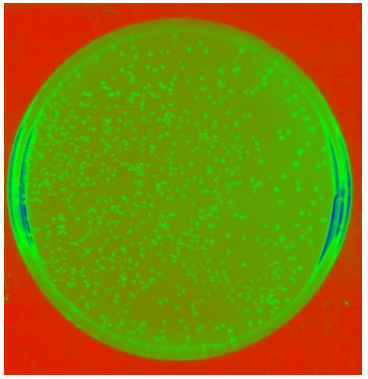
- 4800 fosmid pool
- Assembled with SOAP denovo
- Iterative gap closing applied

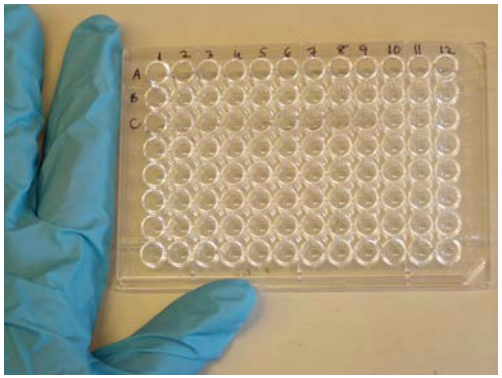
	elem	min	max	mean	sum	end verified	coverage
scf20K+	5323	20013	75791	32284.95	171852787	2001	93%
scf30K+	3719	30010	75791	35426.96	131752852	1911	71%
scf40K+	338	40000	75791	43007.42	14536509	181	8%

- 20K+ scaffolds for largest pool exceeds target coverage
- This assembly appears in later

Components of a 96 Well Production Format

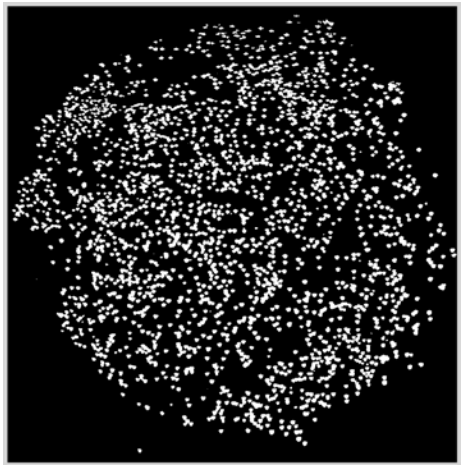
- Harvest 96 bacterial pools of colonies with one fosmid per colony.
- Amplify the harvested colonies as “pools”
- Purify fosmid pool DNA
- Digest DNA with the homing endonuclease PI-SceI
- *Capture fosmid inserts, using a biotinylated DNA Triple Helix motif, washing away the fosmid vector backbone.*
- Convert to 96-well format
- Shear the individual pools for short insert libraries in a 96-well micro-tube plate.
- Illumina’s 96-well Tru-Seq protocol.
- MiSeq QC step for more efficient use of HiSeq
- Superpooling to create long insert libraries

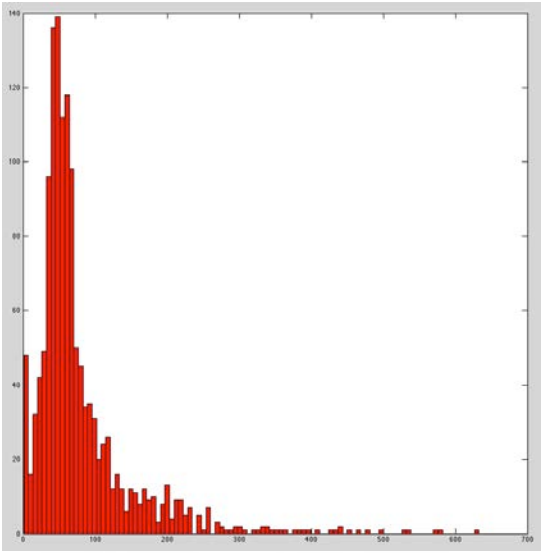




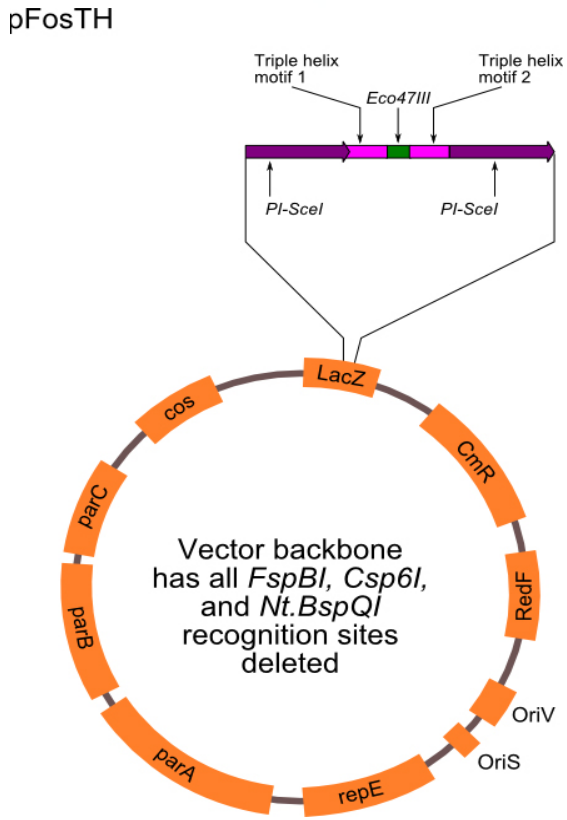
Fosmid Pool Size Estimation

- Threshold image using minimum intra-class variance
- Compute connected components *C* based on adjacency.
- Estimate the expected component size *s*.
- Estimated pool size = $|C|\ s^{-1}$





Triple_Helix Purification of fosmid inserts*



- Removal of vector & E.coli contamination
- Scalable
- Simple Purification

*** Rapid isolation of cosmid insert DNA by triple-helix-mediated affinity capture.**
Ji H, Smith LM, Guilfoyle RA. Genet Anal Tech Appl. 1994;11(2):43-7.

*** Rapid restriction mapping of cosmids by sequence-specific triple-helix-mediated affinity capture.** Ji H, Francisco T, Smith LM, Guilfoyle RA. Genomics. 1996 Jan 15;31(2):185-92.

Results from Triple Helix Purification

Comparison of fosmid pools with and without TH

Scale
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0

Fosmid Pool Library	% E. coli + Vector
LFP_500_336L	17.8
LFP_1000_336L	17.8
LFP_2000_336L	16.3
LFP_3K_3	17.7
LFP_4K_3	17.6
LFP_5500 (jumping)	15.9
Median	17.6

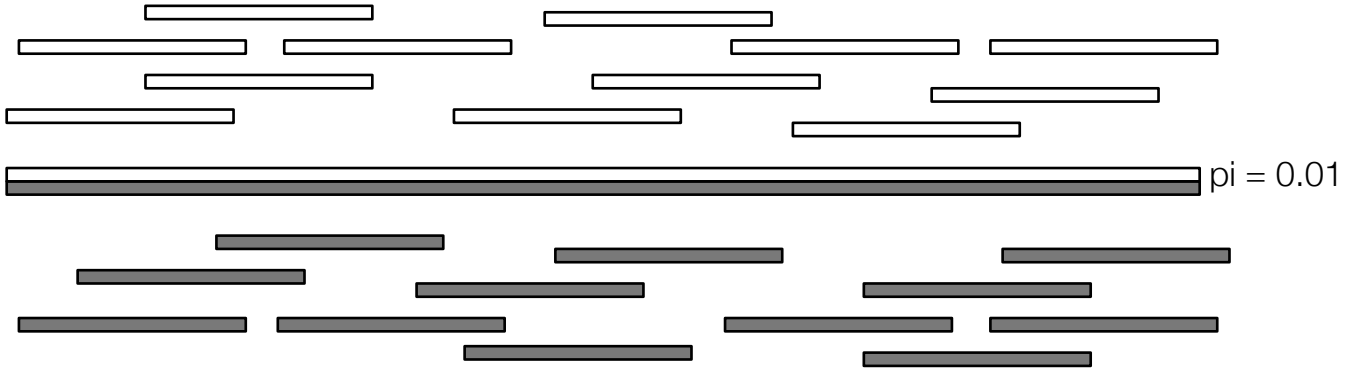
94.7% Reduction in E. coli + Vector

16.3% Reduction in sequencing costs

CHORI2820_LP1 % E. coli + Vector

	1	2	3	4	5	6	7	8	9	10	11	12
A	4.0	2.2	1.6	1.1	1.5	2.4	2.5	1.8	2.1	2.2	1.8	6.5
B	1.0	1.3	2.3	1.1	1.5	1.7	1.3	2.2	1.4	1.4	1.3	1.6
C	0.0	0.0	1.8	1.4	1.1	1.2	1.3	1.2	2.6	2.1	1.1	1.8
D	1.8	0.8	0.5	1.8	1.2	1.1	1.7	1.9	1.2	1.8	1.3	1.8
E	0.4	0.6	0.8	1.3	1.0	0.9	1.1	1.1	1.1	1.0	1.0	1.2
F	1.9	1.4	1.1	2.4	1.0	1.4	2.3	1.5	2.1	1.0	1.6	1.0
G	1.3	1.1	1.6	1.2	1.2	1.1	2.5	1.2	1.1	1.1	1.0	0.9
H	1.6	1.7	1.4	4.8	1.2	1.1	2.1	1.1	2.4	1.1	1.3	0.9
Median			1.3									

A Strategy for Fosmid Based Assembly



- Approx 4000 fosmids per pool (approx 150 Mbp)
- 2.5 – 3.8x fosmid clone coverage (4 - 6 plates)
  - pilot plate + 3 additional plates
- Assemble fosmid scaffolds into larger scaffolds
  - OLC Approach (e.g. MSR-CA, CABOG)
  - Use existing WGS scaffolds with MUMmer/nucmer

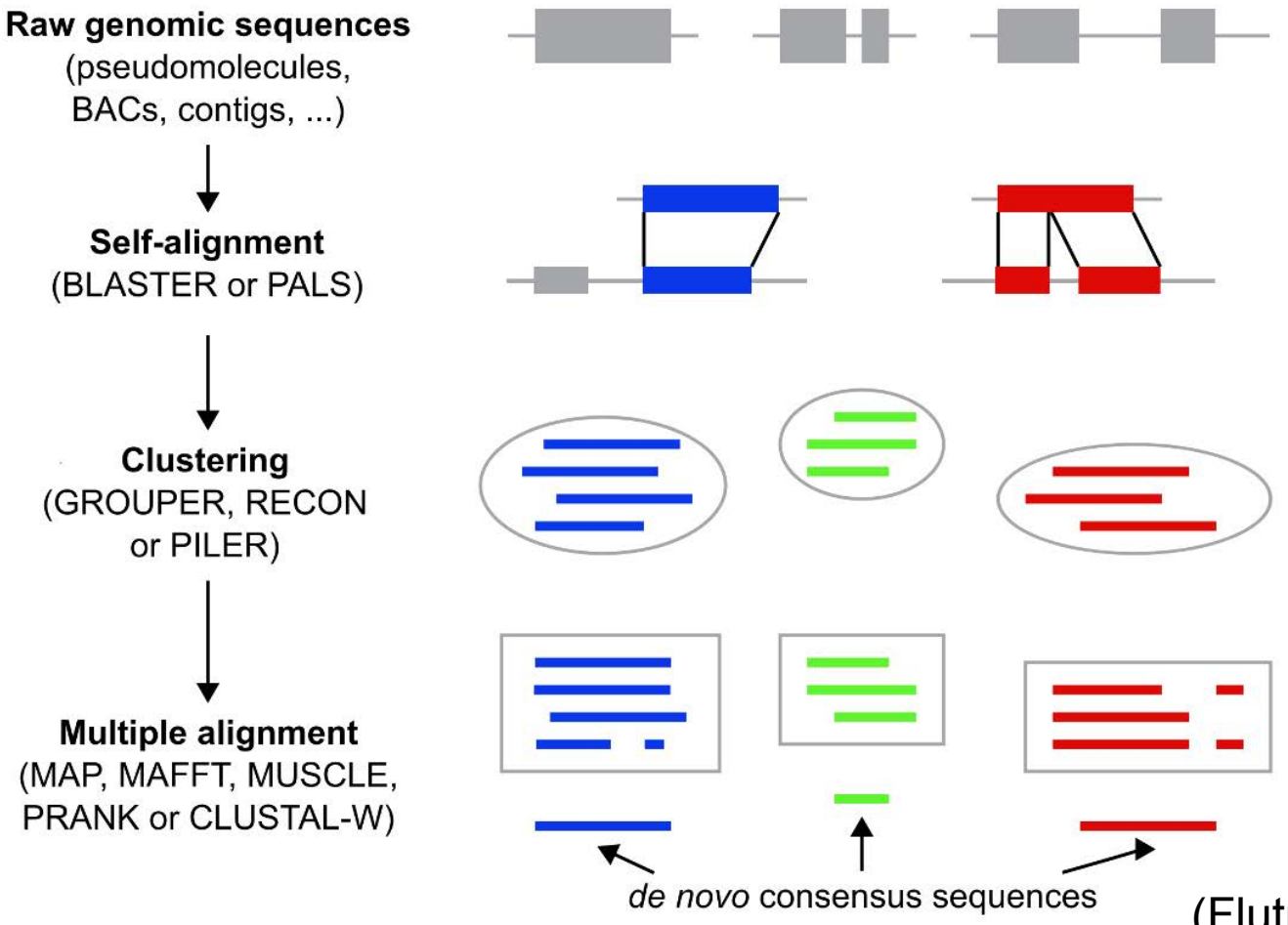
Additional Use Cases for Fosmid Pools

- Assembler Evaluation
- Repeat Library Construction
- SNP Discovery

Repeat Discovery in Fosmid Pools

- For similar repeats methods utilizing k-mers, suffix trees, suffix arrays, or De Bruijn graphs work well.
- Very long contigs allow us to use a “top-down” strategy to find more divergent repeats.
- We used the REPET pipeline (Flutre et. al. 2011) developed at URGI - Unité de Recherche Génomique Info.
  Authors are well known for TE annotation.

REPET *TEdenovo* Pipeline



(Flutre et al, 2011)

SNP Discovery in 20-1010 Aligned Fosmid Consensus Results

- Contigs from the 4800 fosmid pool were aligned to genome using MUMmer/nucmer package developed in the Salzberg lab.
- SNPs per contig:

(bi-modal)

Number of contigs	Min	Max
11	50	60
58	40	49
190	30	39
221	20	29
384	10	19
391	6	10
2,543	0	5

- Estimated heterozygosity of 1.5%

SNP Discovery in 20-1010

- Established paradigm for NGS sequence
  - Align reads to reference sequence
    - MAQ
    - Bowtie, Bowtie2
    - BWA
    - SOAP
  - Determine alleles
    - MAQ
    - Samtools
    - GATK

Only BWA works with genomes larger than 2^{32}

20-1010 SNPs: Aligned Fosmid Read Results

- Reads from 4800 fosmid pool were aligned to the genome with bowtie2 from the Salzberg lab.
- Genomic target recruited using reciprocal best hit
- Use samtools for allele calling.
- 262,290 snps were identified as high quality in 66,685 contigs
- Alignments span approximately 67 Mbp
- SNP Rate of 0.4%
- Estimated heterozygosity of 0.8%
- This implies approx 70-140 million SNPs from 20-1010

Thank You

Up Next:

The Loblolly Pine WGS *V0.8*