PineRefSeq: Towards a Reference Genome for Sugar pine (*Pinus lambertiana*)

Kristian Stevens

Dept. of Evolution and Ecology
University of California, Davis

with

Marc Crepeau, Daniela Puiu, Aleksey Zimin,
Jill Wegrzyn, Maxim Koriabine, Charis Cardeno, Ann Holtz-Morris,
Pieter J. deJong, Steven L Salzberg, James A Yorke,
Chuck Langley and David B Neale

PAG XXIII, 10th January 2015

PineRefSeq: conifer "mega"-genome sequencing

- *"Mega"-genome sequencing* is our strategy for sequencing the **leviatan** genomes of conifers by exploiting the unique haploid characteristic of their mega-gametophytes.
- We recently applied it to the 22Gb genome of Loblolly pine.
- We report on our sequencing and assembly activities towards version V1.0 of the 50% larger sugar pine genome.
- Additional progress on application to the outgroup Douglas fir.

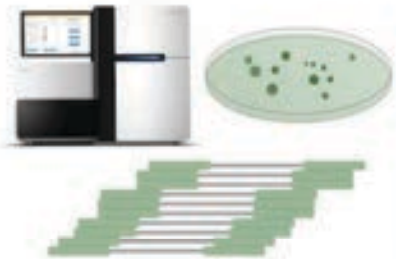
PineRefSeq Project Overview

Loblolly pine
Acc. 20-1010
VA Dept. of Forestry
US Forest Service

Sugar pine
Acc. 5038
US Forest Service

Douglas-fir
Acc. 412-2
Weyerhaeuser Co.

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



Sequencing
UC Davis
- Libaries: short/long inserts
- Sequencing
  - = GAIIX (overlapping reads)
  - = HiSeq (125 bp reads)
- Data curation and fosmid assembly

Genetic Maps
UC Davis
- SNP calls from scaffolds
- Genetic mapping to order scaffolds

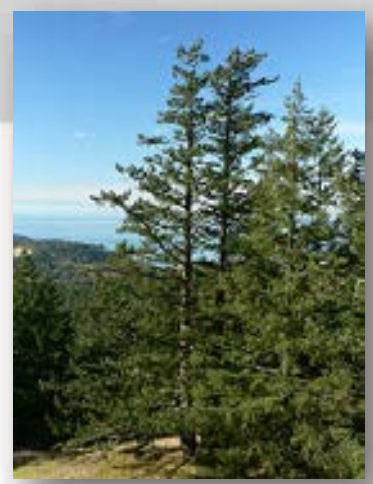
Assembly
Maryland-JHU Genome Assembly Group
- Assembler development (MSR-CA)
- Assembly
  - = WGS
  - = Fosmid
  - = Meta

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

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

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

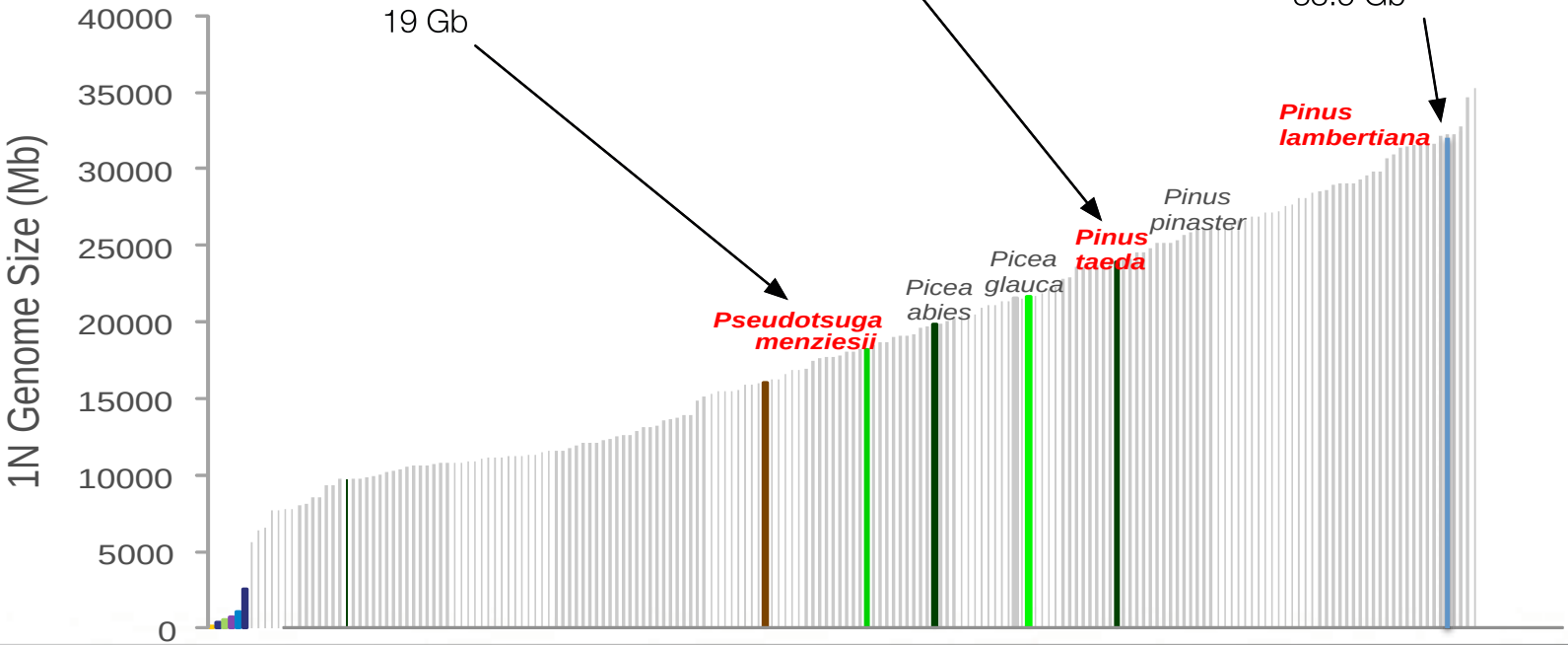
The “Mega”-genomes of Conifers



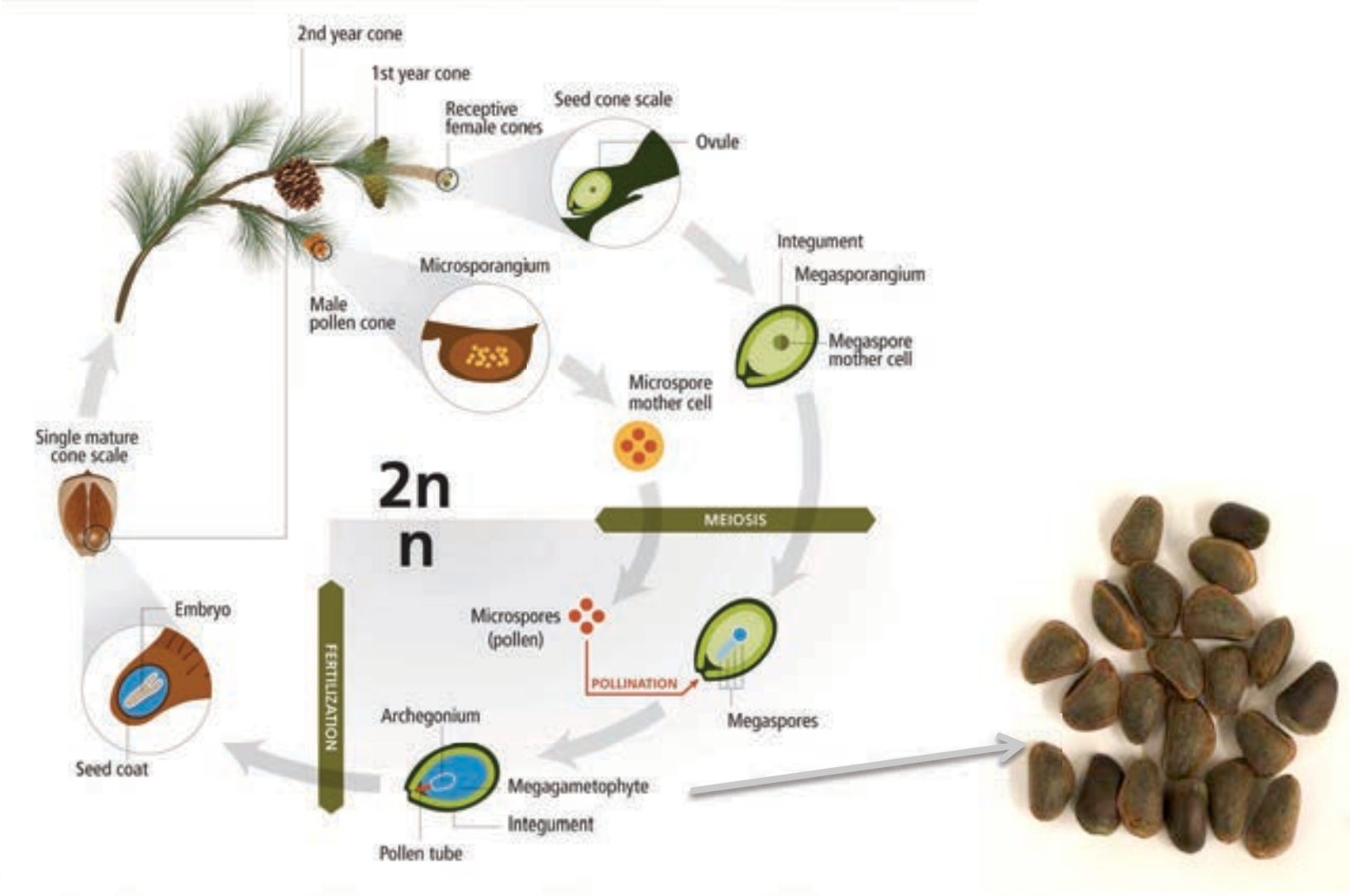
Douglas fir
Pseudotsuga menziesii
19 Gb

loblolly pine
Pinus taeda
22Gb

sugar pine
Pinus lambertiana
33.5 Gb

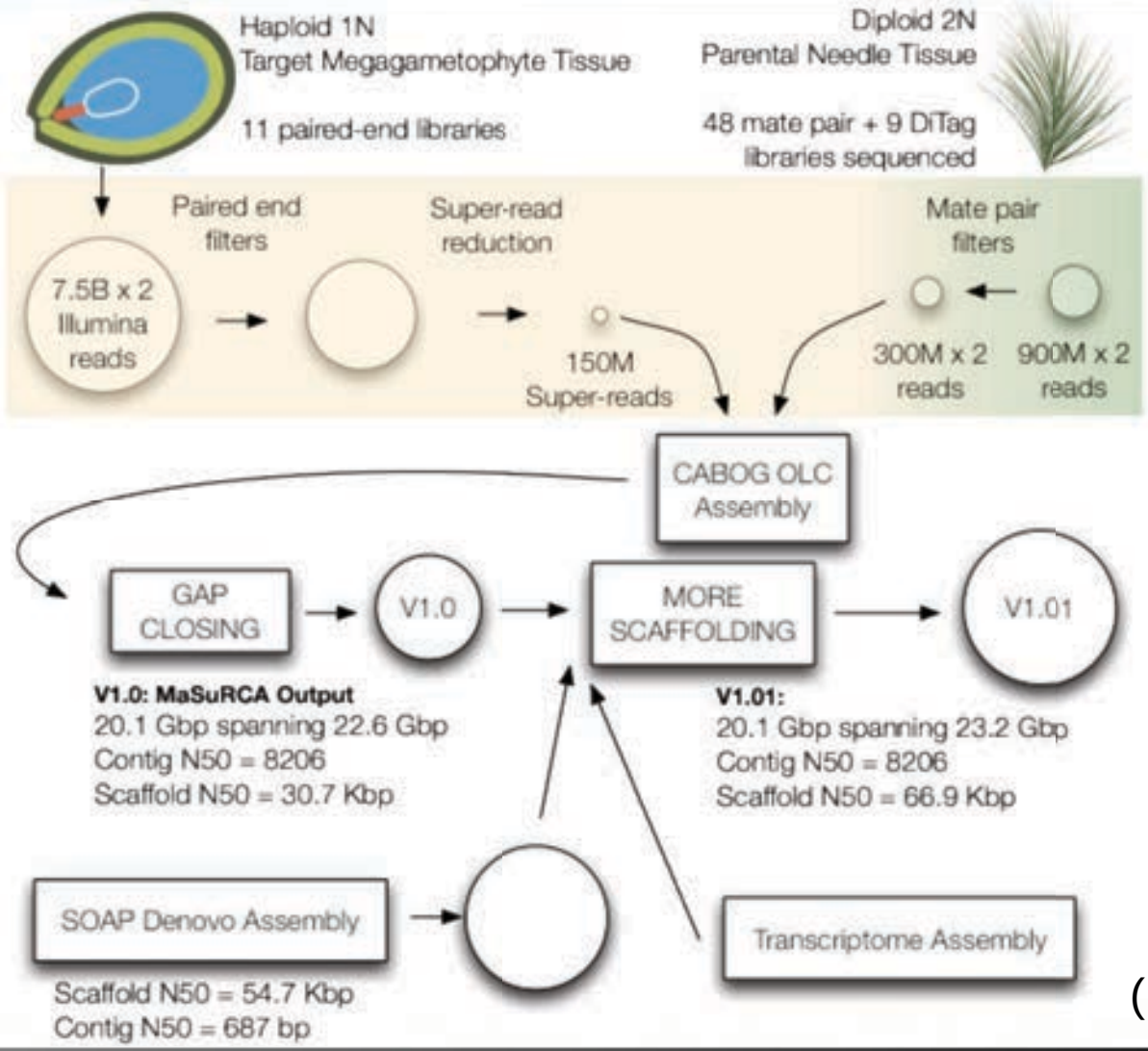


Conifer lifecycle



Sequencing and Assembly Strategy

Loblolly Pine

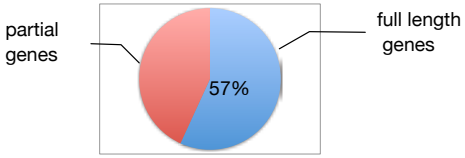


(Neale *et al.* 2014)

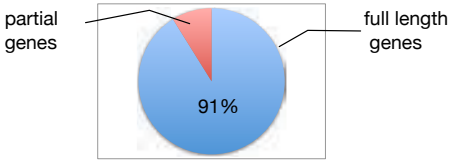
Transcriptome Scaffolding

CEGMA 248 Core Conserved Genes (Parra et al. 2009)

CEGMA: 197 core genes found



CEGMA: 203 core genes found

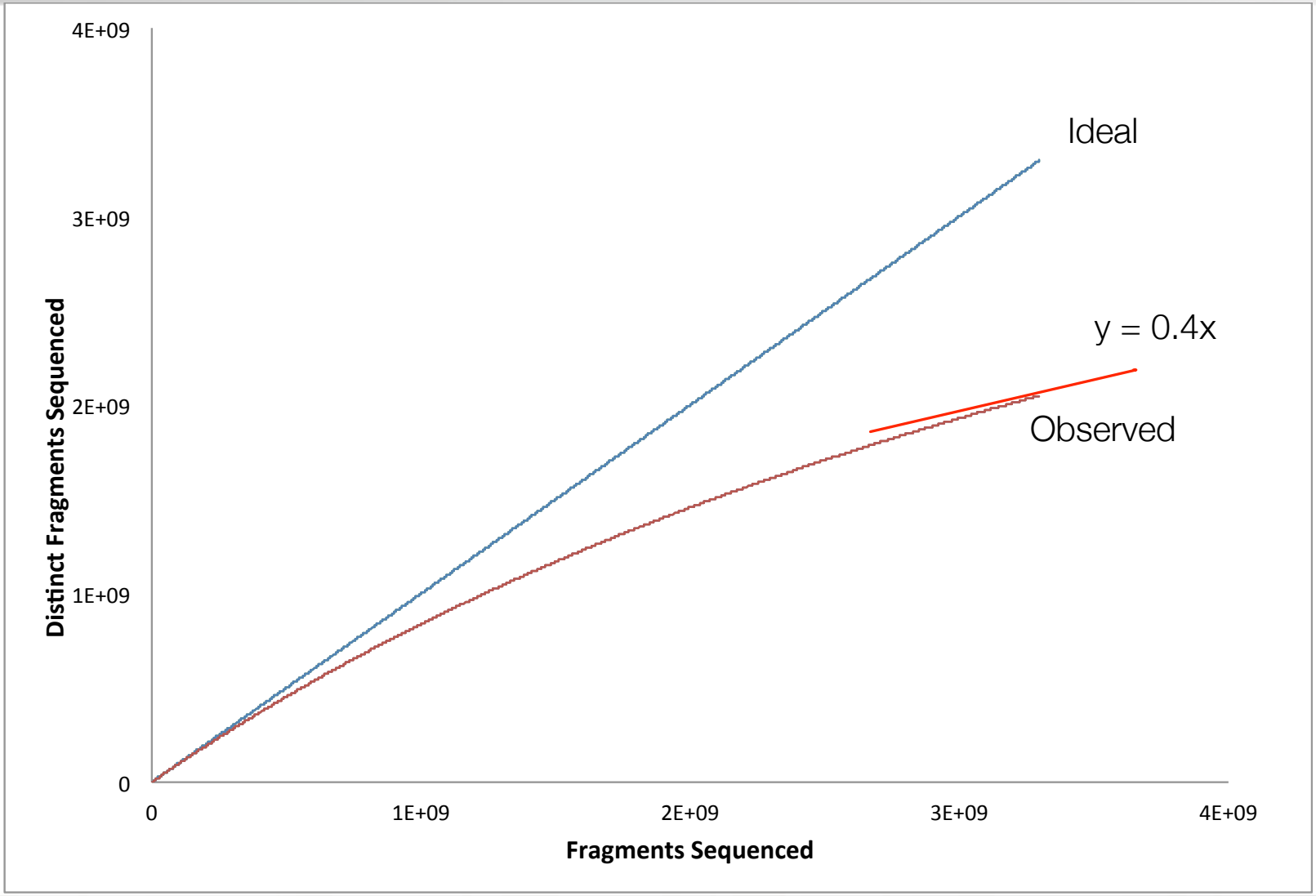


(Neale *et al.* 2014)

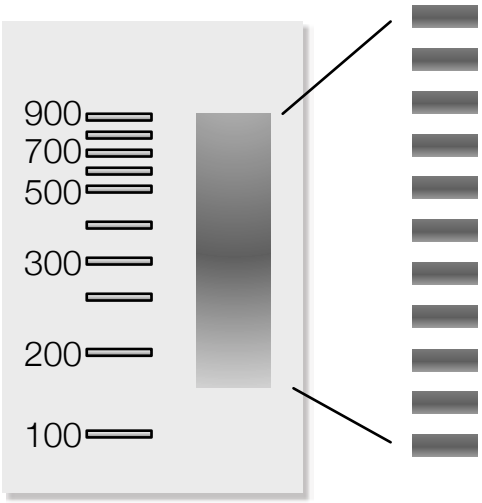
Sugar pine sequencing goals

- Deep (> 50x) paired end sequencing of megagametophyte
  - Increase complexity
  - Reduce bias
  - Greater short scale contiguity
- Deep mate pair sequencing of needle tissue, using longer libraries.
- Enhanced gene model coverage using transcriptome scaffolding
- Very long scale contiguity (35-40kbp) using fosmid end sequences

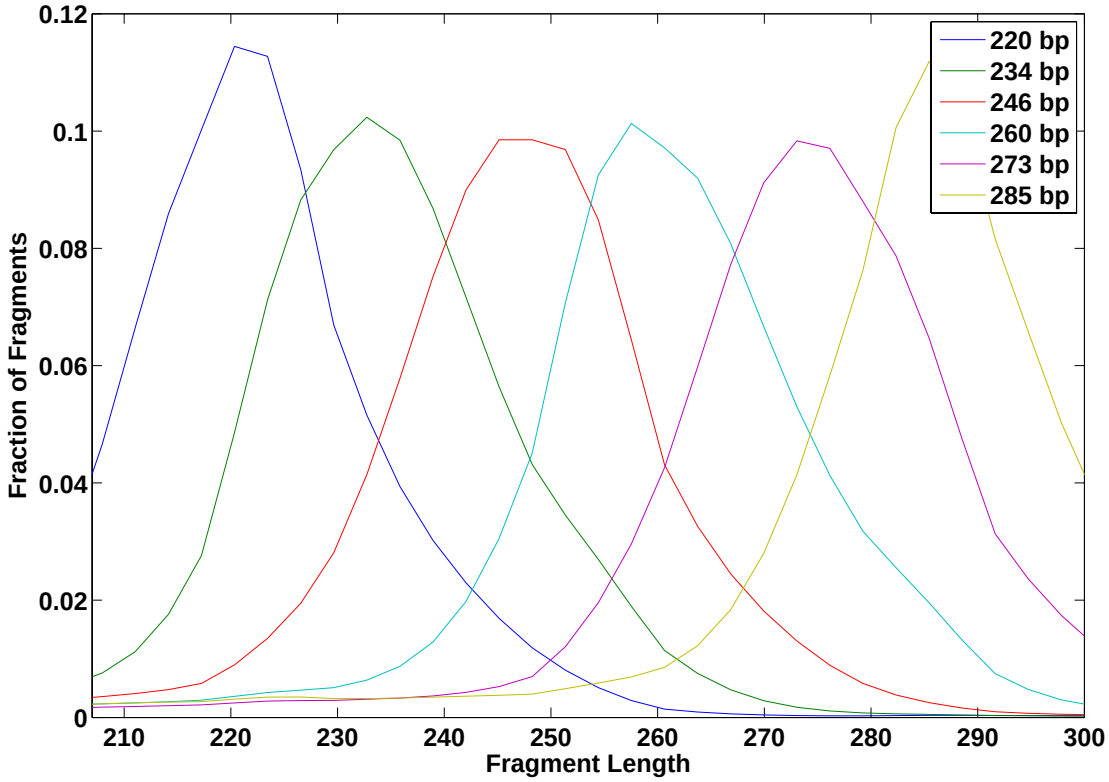
Challenge: Library Complexity
(Daley and Smith 2013)



Mega-gametophyte Partition Libraries



Goal:
Maximize total complexity with
good individual library utility (cv)



First the mega-gametophyte

loblolly pine
Pinus taeda

cytometrically estimated
1N genome size
22Gb

Accession 20-10-10
VA Dept of Forestry
US forest service

megagametophyte dry weight
mean **23.5 mg** (n = 105)

sugar pine
Pinus lambertiana

cytometrically estimated
1N genome size
33.5 Gb

Accession 5038
US forest service

megagametophyte dry weight
mean **224.5 mg** (n = 40)

Douglas fir
Pseudotsuga menziesii

cytometrically estimated
1N genome size
19 Gb

Accession 412-2
Weyerhouser Co.

megagametophyte dry weight
mean **11.1 mg** (n = 105)

Over 2.5 Tbp of sequence from Illumina platforms

HiSeq

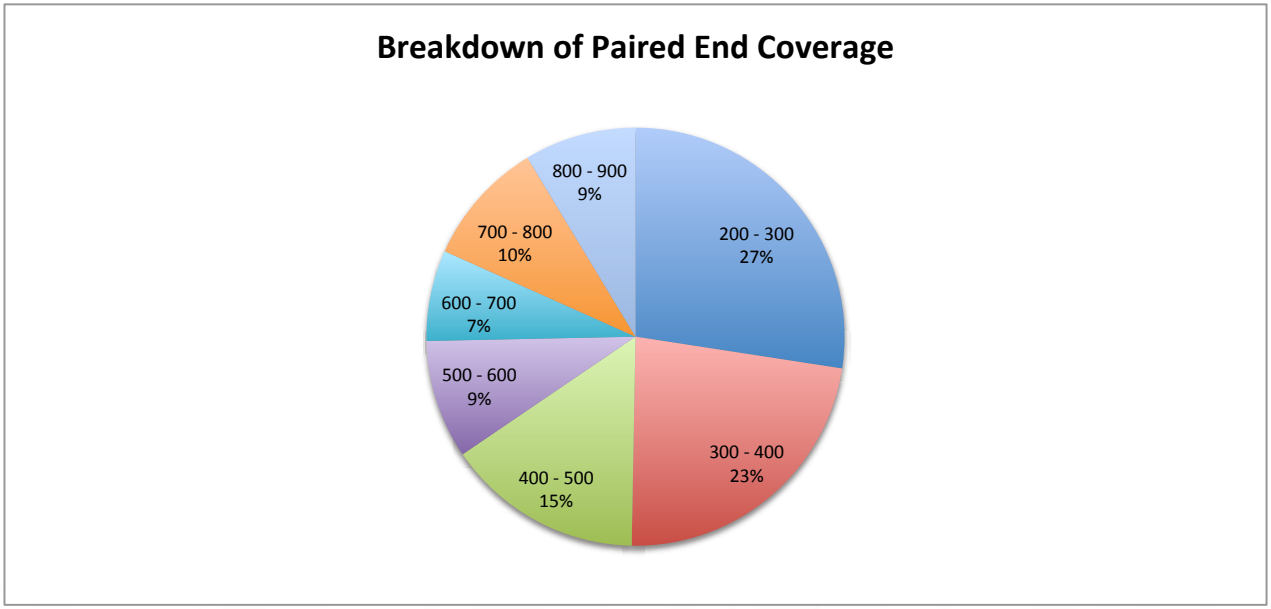
GA II x

MiSeq

Library Type	Instrument	Fragment size	Read length	Coverage
Illumina paired-end	HiSeq	200-657	100f+100r	42x
Illumina paired-end	GAIIx	200-657	160f+156r	22x
Illumina paired-end	MiSeq	350-657	250f+250r	<1x

1.9 Tbp of paired end sequence

Number of Libs	Median Insert	Coverage (bp)	Coverage (Gbp)	Coverage (%)	Coverage (X)
18	200 - 300	5.23778E+11	523.78	27%	16.9
14	300 - 400	4.36456E+11	436.46	23%	14.1
7	400 - 500	2.90146E+11	290.15	15%	9.4
5	500 - 600	1.75373E+11	175.37	9%	5.7
4	600 - 700	1.3422E+11	134.22	7%	4.3
5	700 - 800	1.8416E+11	184.16	10%	5.9
3	800 - 900	1.65167E+11	165.17	9%	5.3
		1.9093E+12	1909.30	100%	61.6



k-mers

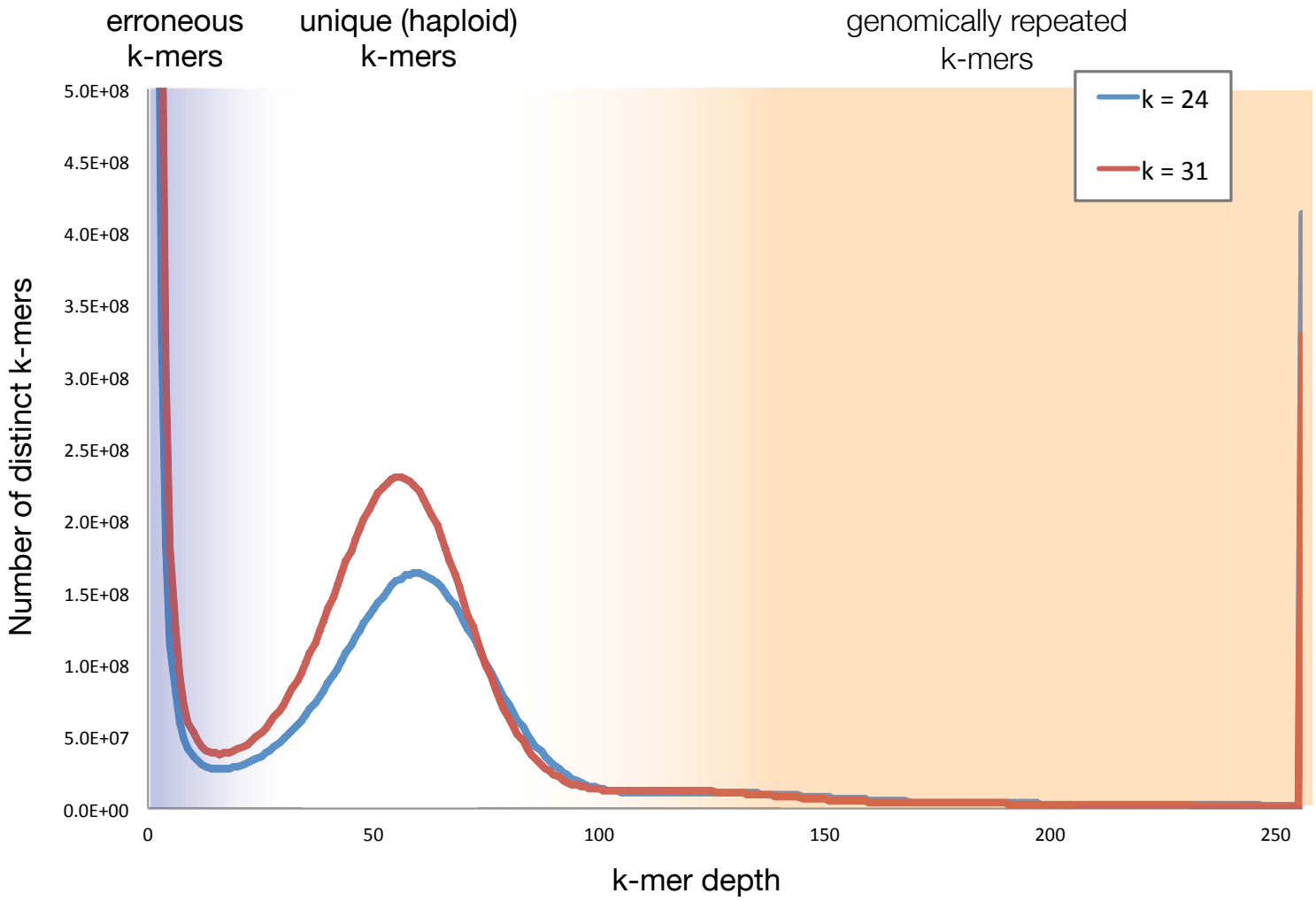
Query: Does a distinct length *k* string occur in the genome?

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

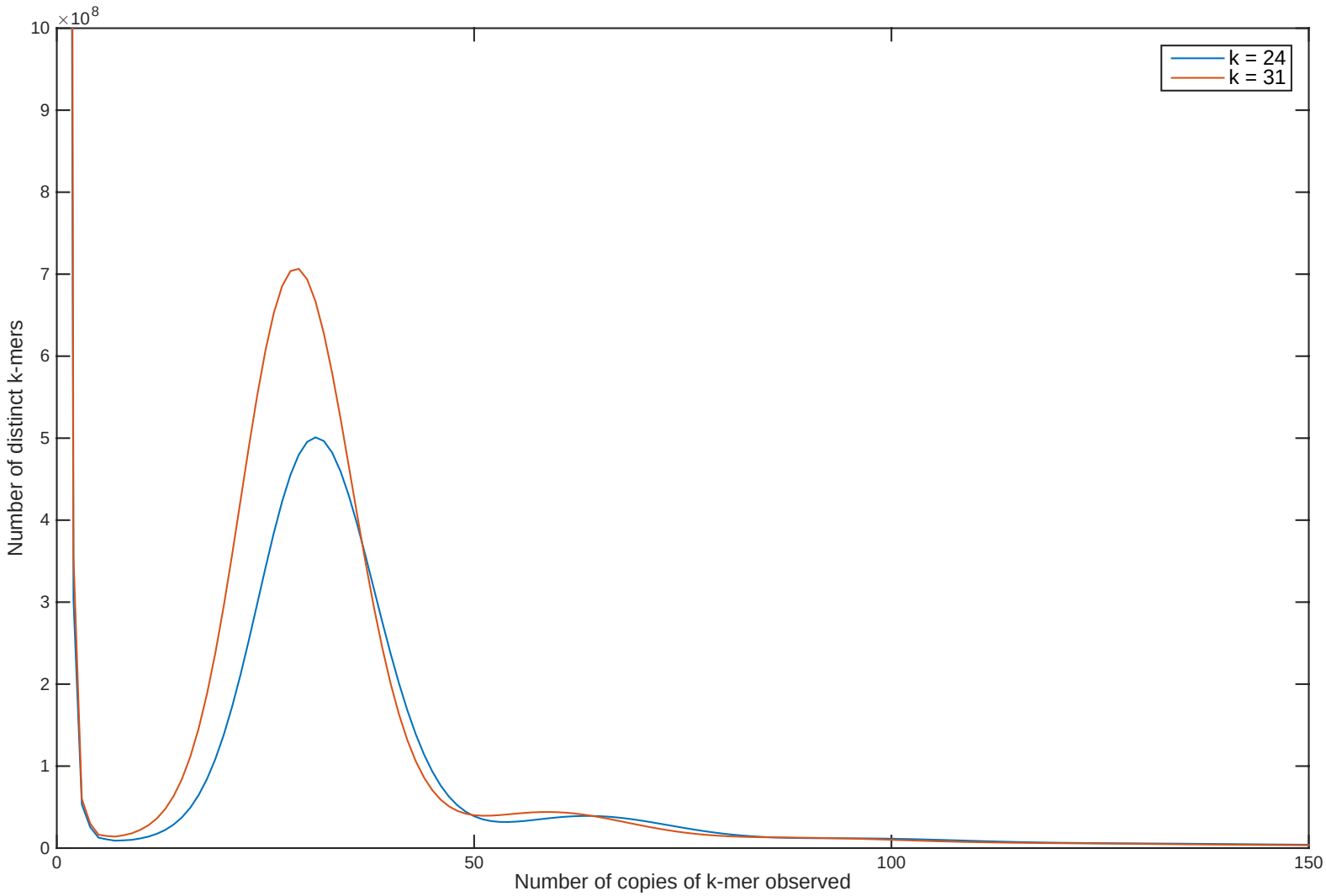
- We choose a *k* that best supports the query above allowing specific locations on the genome to be queried.
- From *k*-mers we estimate genome size, library complexity, correct errors. We can also assemble genomes.

Histograms of 24 and 31-Mers

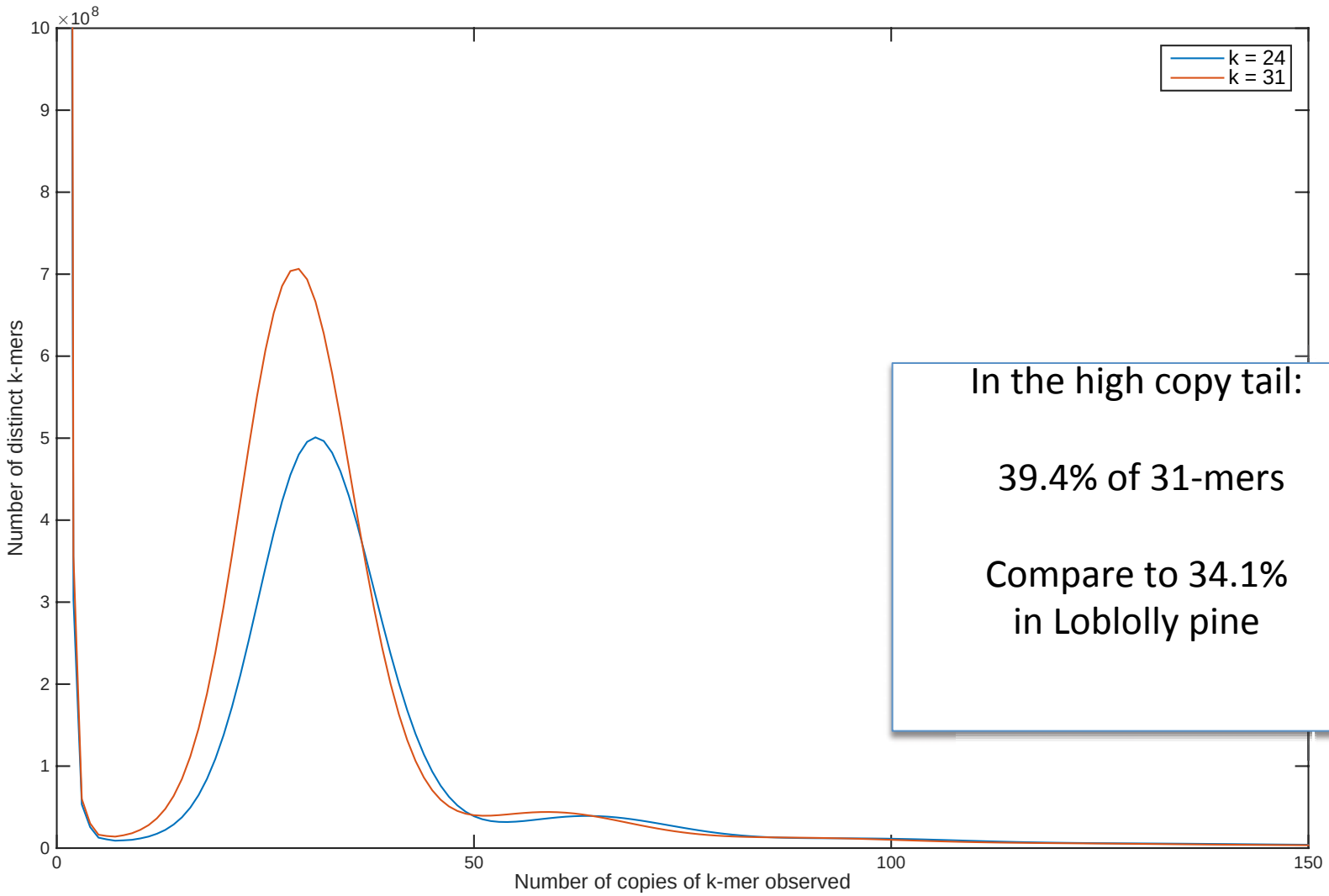
Loblolly pine – *Pinus taeda– HiSeq + MiSeq + GA2x*



62X sugar pine k-mer histograms



62X sugar pine k-mer histograms



In the high copy tail:

39.4% of 31-mers

Compare to 34.1%
in Loblolly pine

Revisiting the sugar pine genome size

How complete are the libraries?

Existing estimates obtained from the

Kew Gymnosperm DNA *C*-values database

Murray BG, Leitch IJ, Bennett MD. 2012.

Gymnosperm DNA *C*-values database (release 5.0)

Genus	Species	Estimation method	1C (Mbp)	Reference
Pinus	lambertiana	RK	10367	Rake and Miksche, 1980
Pinus	lambertiana	Fe	16968	Dhillon, 1980
Pinus	lambertiana	Fe	17213	Dhillon, 1987
Pinus	lambertiana	Fe	28900	Wakamiya et al., 1993
Pinus	lambertiana	FC:PI	29418	Williams et al., 2002
Pinus	lambertiana	FC:PI	31052	Wakamiya et al., 1993
Pinus	lambertiana	FC:PI	33487	Grotkopp et al., 2004
Pinus	lambertiana	Fe	42885	Rake and Miksche, 1980

http://www.kew.org/cvalues/

A k-mer Genome Size Estimate

$$\textit{P. lambertiana genome size} \cong$$

$$\textit{total k-mers in P. lambertiana genome} \cong$$

$$\frac{\textit{total correct k-mers in reads}}{\textit{expected depth of each correct k-mer in reads}}$$

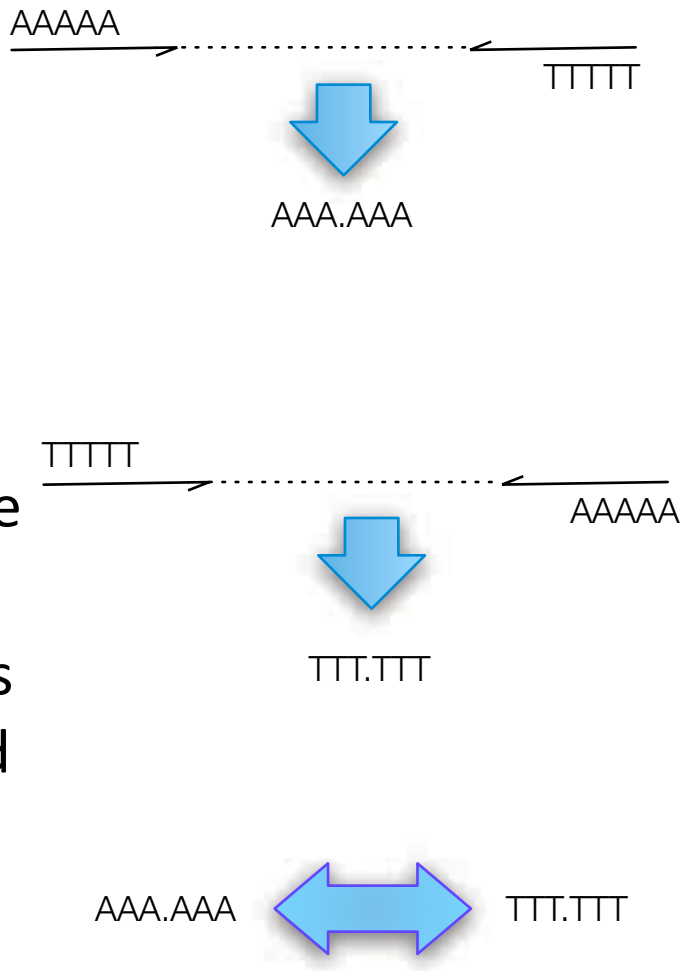
We estimate the expected depth using the
'unique' fraction of the genome
then determine its expected value.

Haploid (1C) Genome Size Estimates

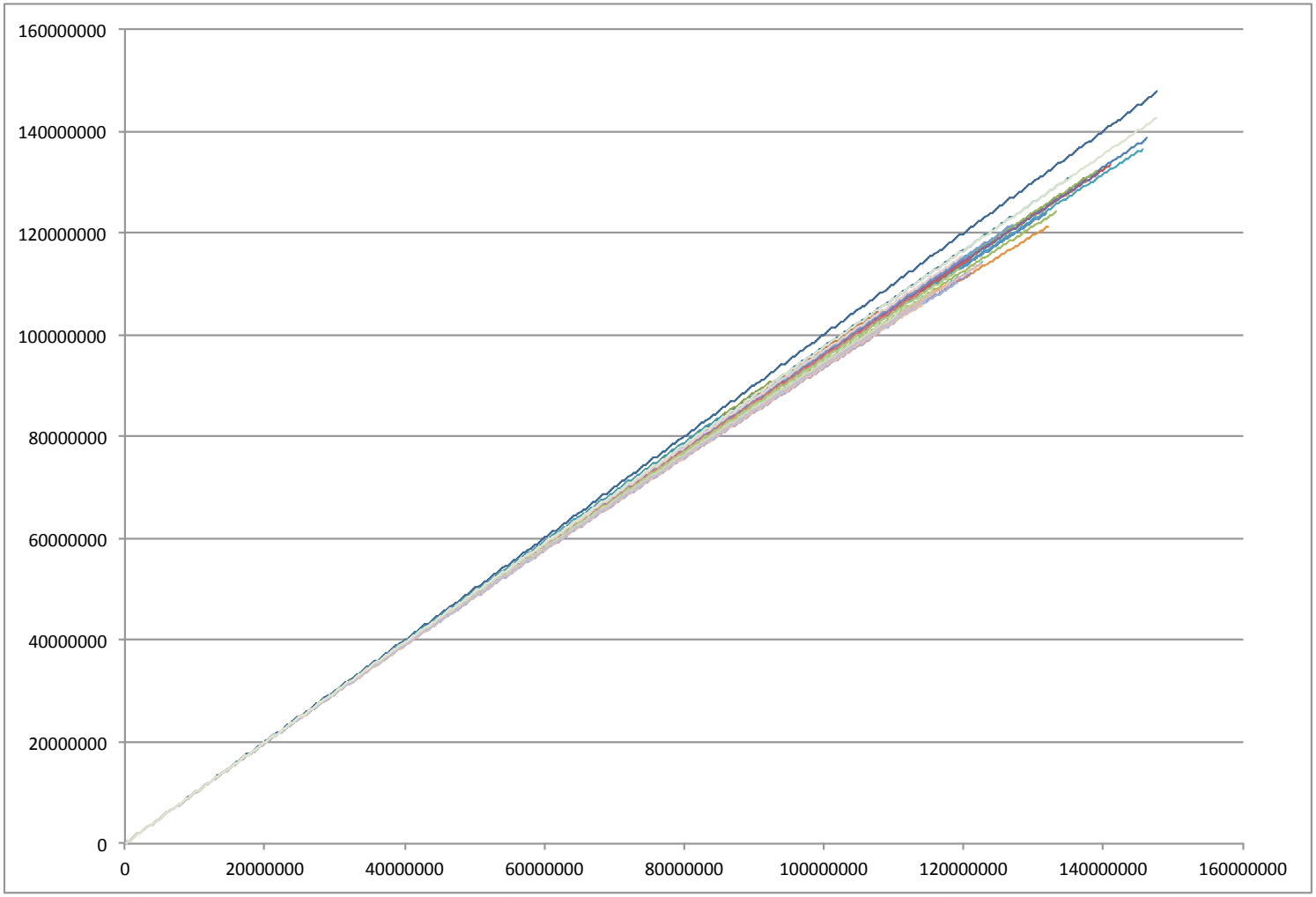
	k-mer length	
	24	31
Total k-mers	1.56E+12	1.47E+12
Erroneous k-mers	1.17E+10	2.19E+10
Total correct k-mers	1.55E+12	1.45E+12
E(unique k-mer depth) mode	49.72	46.77
Genome size	3.11E+10	3.09E+10
E(unique k-mer depth) mean	48.53	46.02
Genome size	3.19E+10	3.14E+10

k-mers for Library Complexity

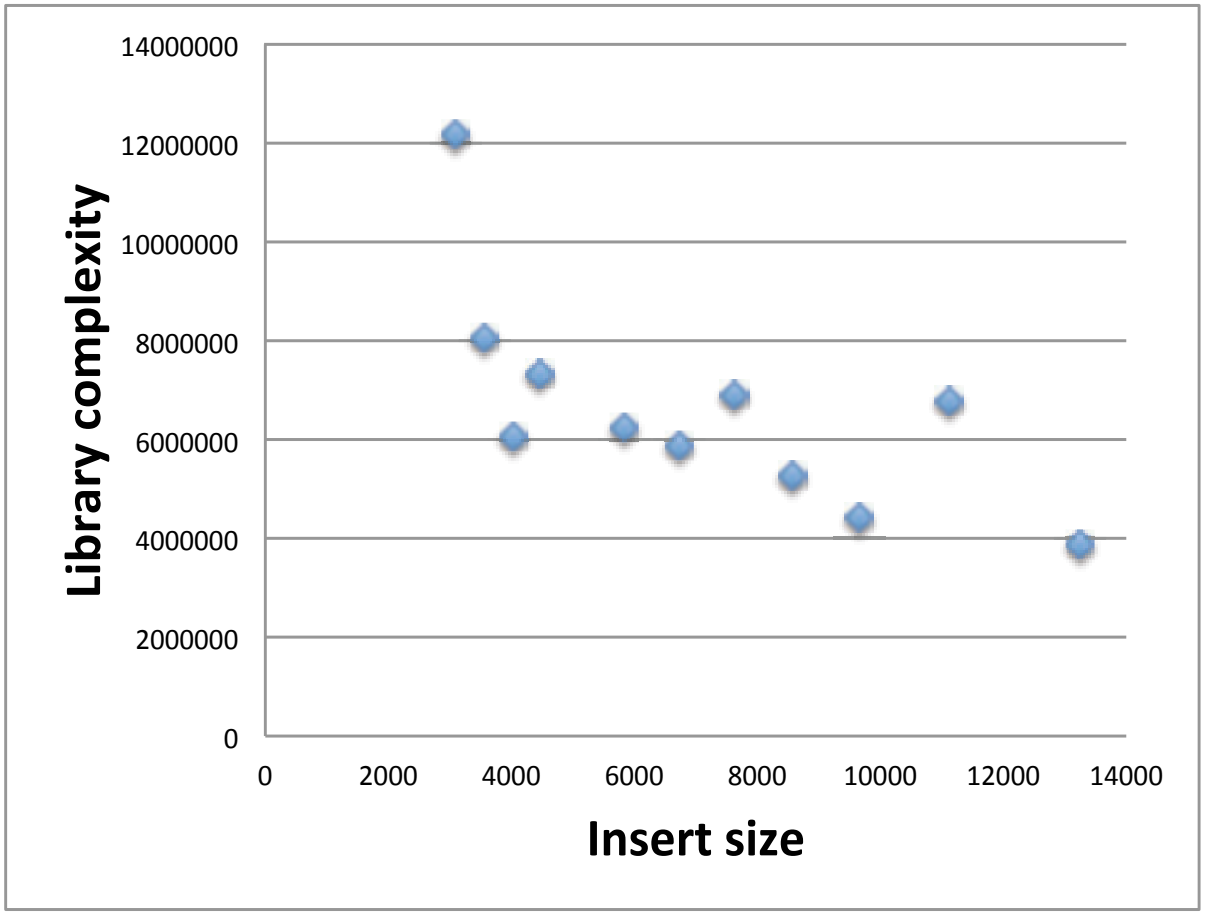
- ***Library complexity*** is the number of distinct DNA fragments in a library.
- We can measure the sequenced library complexity directly.
- We summarize each sequenced fragment using 24-mer suffix prefix tags (right)
- We identify two fragments *u v* as the same (PCR replicated) if *u = v* or *u = rc(v).*
- The histogram of fragment depth provides a full and compact summary of sequenced library complexity.



Sugar pine library complexity curves



Sugar pine mate pair library complexity



Sugar Pine Nextera Mate-Pair

Mate-pair libraries	Insert Size Range	Mbp	Sequence Coverag	Phys. Coverage (Masurca)
3	[3kbp 5kbp)	128756	4.2	21.11
6	[5kbp 10kbp)	78986	2.5	16.51
4	[7.5kbp, 10kbp)	35526	1.1	9.32
11	[10kbp, 15kbp)	133896	4.3	28.64
12	[15kbp, 25kbp)	83158	2.7	13.03
Totals 36		460322	14.8	**88.61**

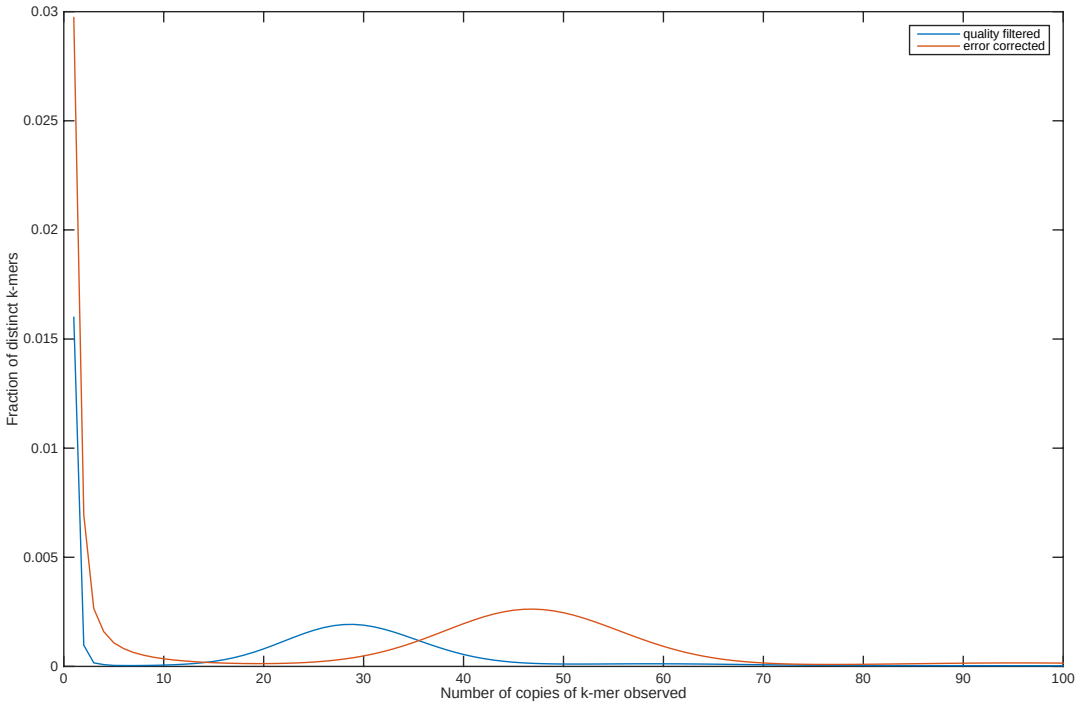
Quorum Error Correction

(Marcais *et al.* 2014)

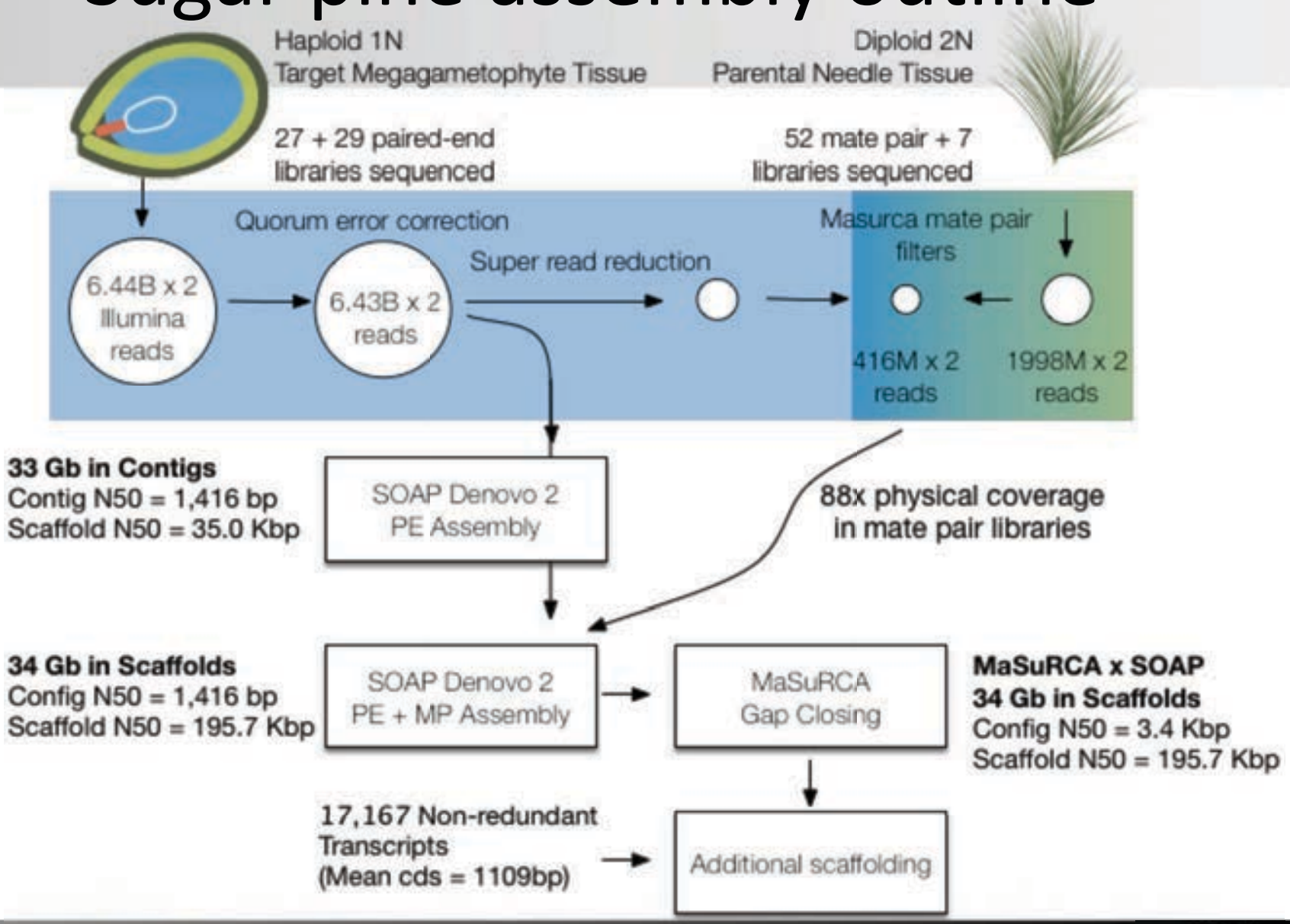
- The statistical property that errors are rare is used to correct reads using a database of classified k-mers (www.genome.umd.edu/quorum.html)

- Input:
  - 12,877,002,750 paired end reads
  - 1,892,820,336,800 bases
  - ??? distinct 99-mers

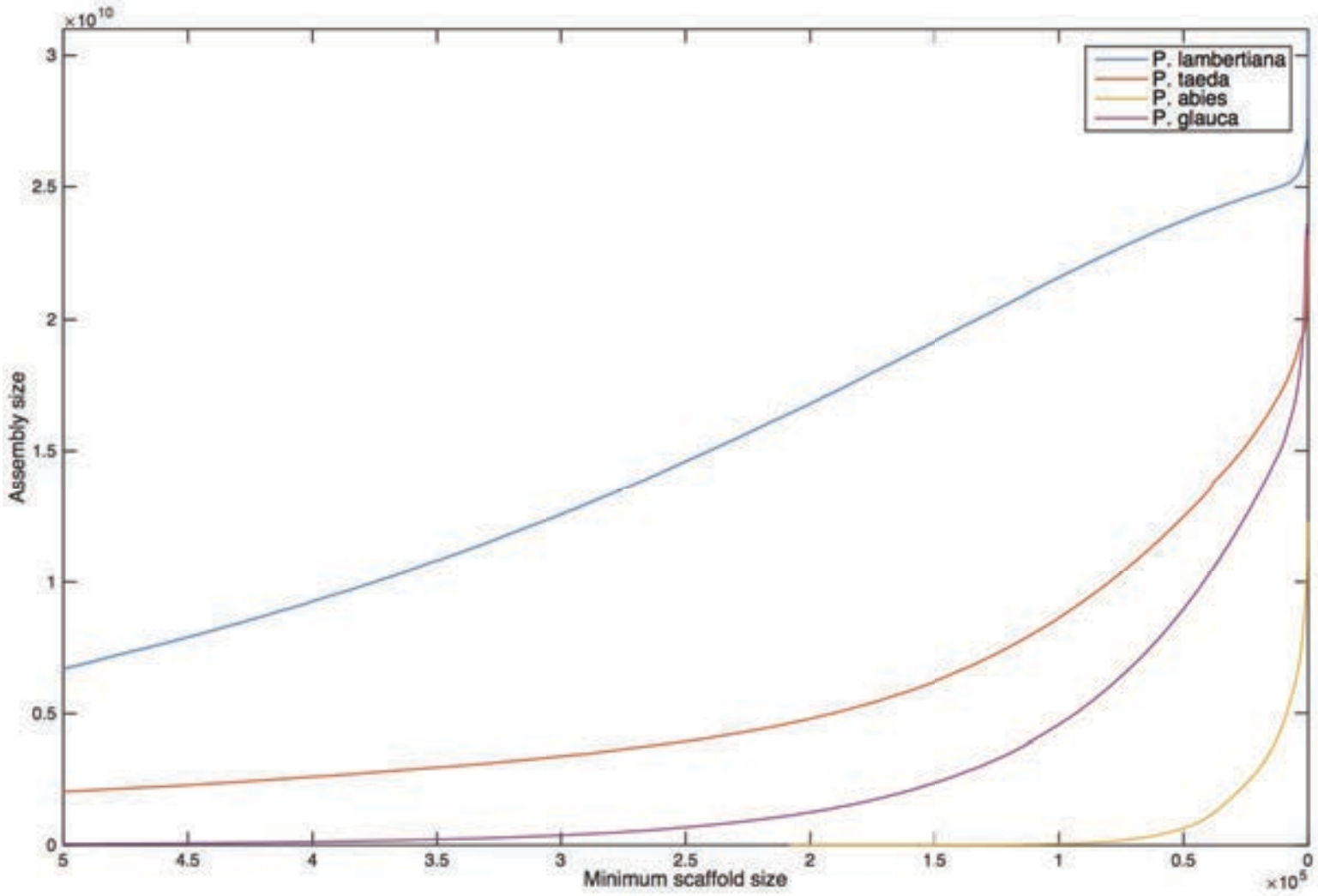
- Output:
  - 12,858,165,085 paired end reads (99.85%)
  - 1,851,750,169,747 bases (97.83%)
  - ??? distict 99-mers



Sugar pine assembly outline

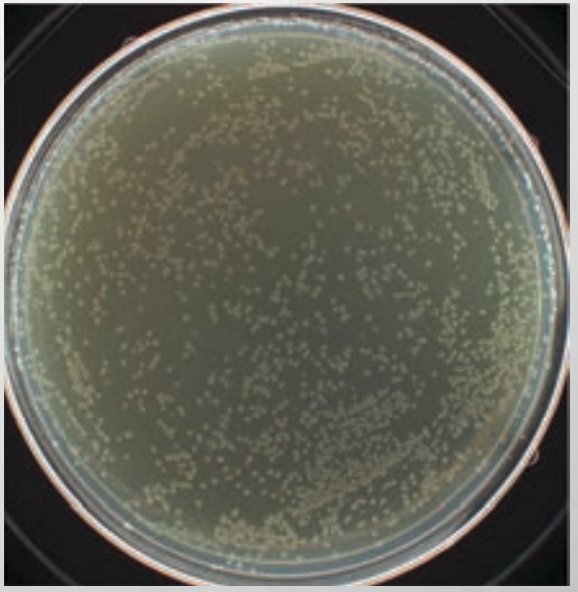


Contemporary Conifer Genomes



Fosmids

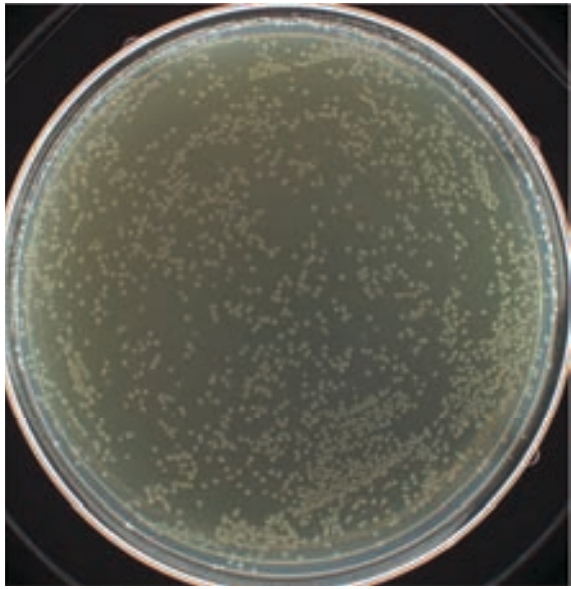
For long range contiguity information, assembly validation, and repeat discovery.



Fosmid colonies

Long Range Scaffolding with Fosmid DiTags

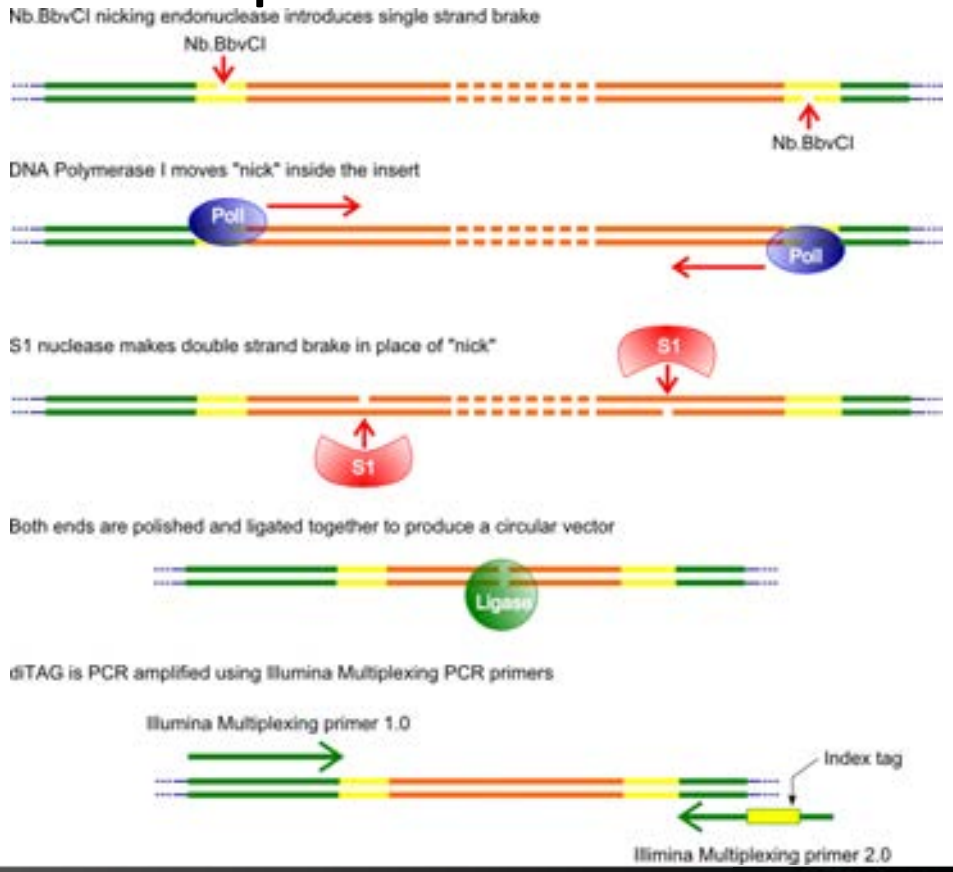
- We start with a fosmid insert (35-40kbp)
- End sequencing protocols developed in the de Jong lab
- A two-pronged strategy:
  - First generation nick translation 'NT' libraries
  - New generation 'TH' libraries



Fosmid infected *E. coli* colonies:

Traditional Fosmid DiTags

- Traditional approaches use *nick translation* to reduce the size of a 35-40kbp insert to a manageable size for paired end illumina library.



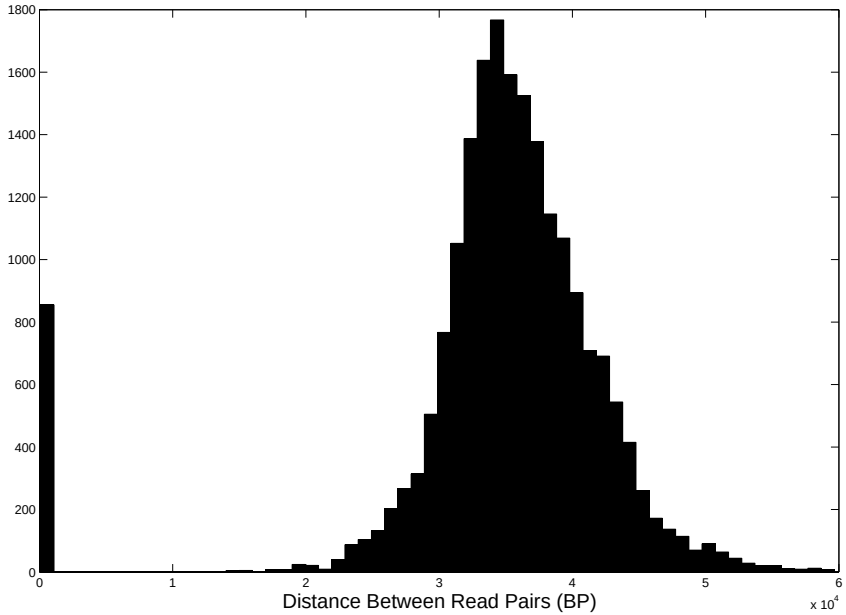
(Gnerre *et al.* 2010)

(Williams *et al.* 2012)

(Zimin *et al.* 2014)

Challenges of NT Fosmid/Fossill DiTags

- Non-junction fragments (2-13%)
- Chimeras (2-8%)
- Limited complexity
- Also see Williams *et al.* 2012

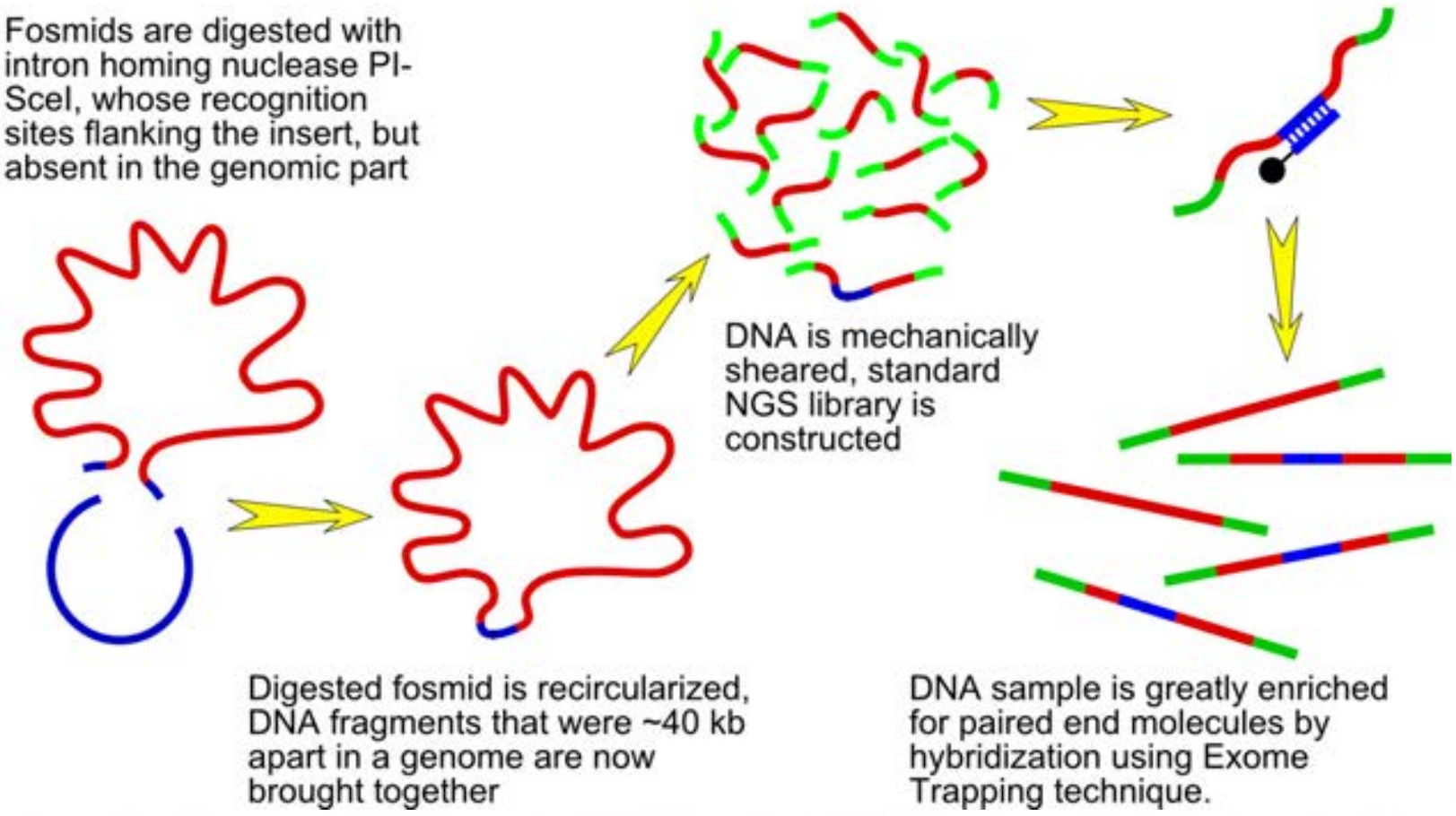


Fosmid size distribution estimated from aligned DiTags.

Median fosmid size estimate is 37.5 Kbp.

An alternative approach to DiTags

Recircularization of the insert and specifically pulling joined end sequences from the mixture of the NGS library

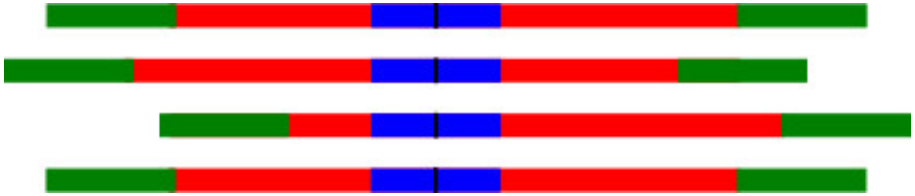


Duplicates vs replicates

Three fosmid replicates, top and bottom are PCR duplicates

Library 1: type “NT”

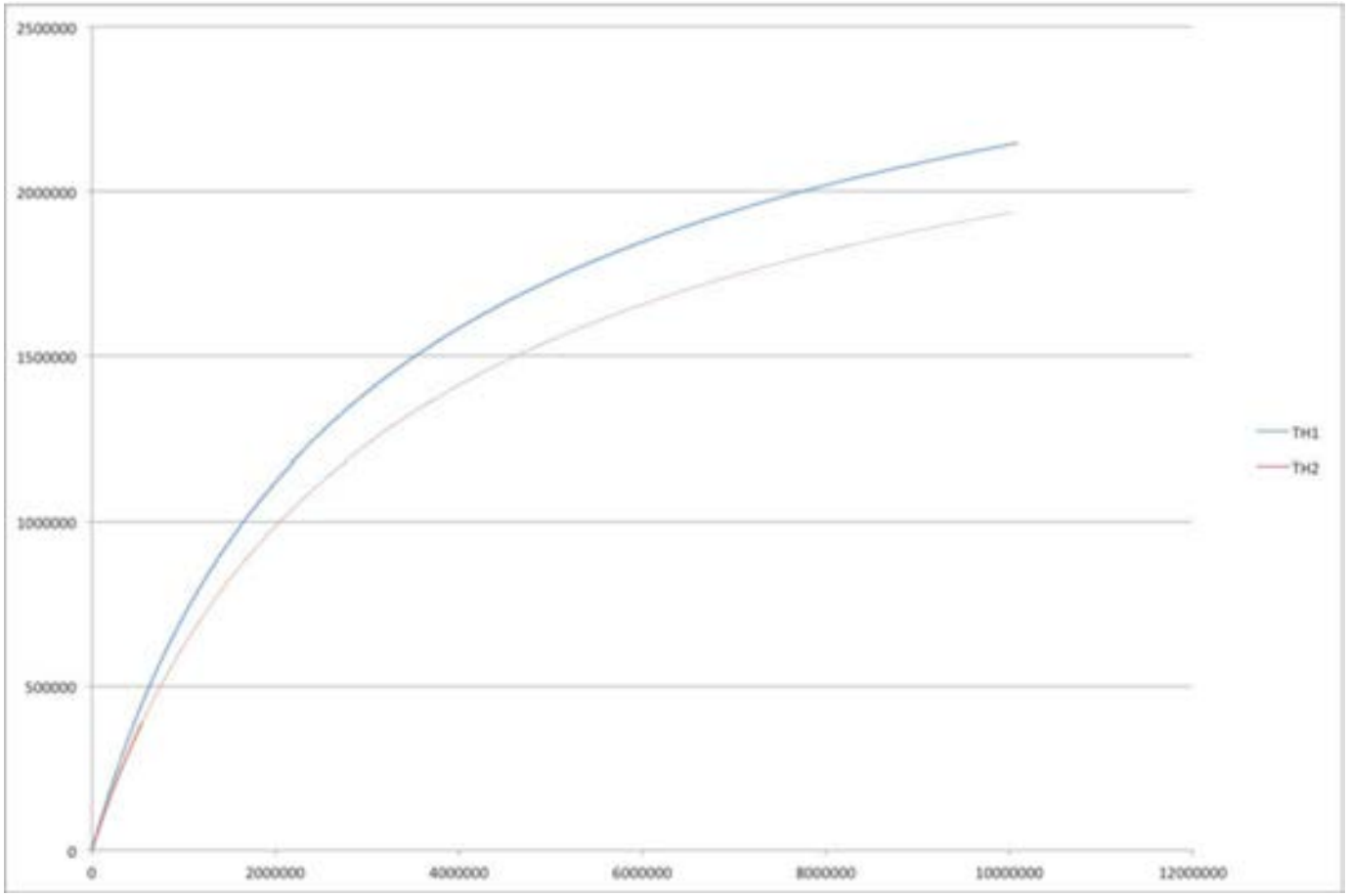
Library 2: type “TH”



DiTag Analysis Pipeline

- Similar to Nextera pipelines (Legett *et al.* 2014; O'Connell *et al.* 2014)
  - Need to distinguish fosmid replication and PCR duplication
  - Fosmid confirmed with 83-mer approximate matching (15 – 50%)
- Classes of 83-mer sequences.
  - Insert is identified as fully sequenced with internal tag (70 – 95%)
  - Tag is found in the first or second read (5 - 30%)
- To reduce chimeras (< 0.3%) we look for two independent fosmid (not PCR) replicates.

Library complexity of type TH



Sugar Pine DiTag Sequencing

Fosmid library set (same pool)	Component Illumina libraries	Construction methodology	Fosmids sequenced	Distinct inserts	Physical coverage
CHORI-3828DT1	4	NT	13035393	1866000	2.26
THSp40mill	2	TH v 1	23140352	2774704	3.36
SP-NA-3H	1	TH v 2	4825255	1456467	1.76
TOTAL	**7**		**41001000**	**6097171**	**7.38**

Sugar pine WGS Summary

- Deep representative haploid coverage can be obtained from a sugar pine mega-gametophyte.
- **62X (1.9Tbp)** high quality **paired end** coverage of the **31Gb sugar pine** genome has been obtained.
- **56 sugar pine mate pair libraries** have been constructed, processed, and resulted in **88X physical coverage**, 40x from libraries 10kbp and larger.
- For longer links**:** Over **7X physical coverage from 35-40Kbp DiTags** from fosmid libraries.
- Gene model coverage to be increased by transcriptome scaffolding (17,184 transcripts)

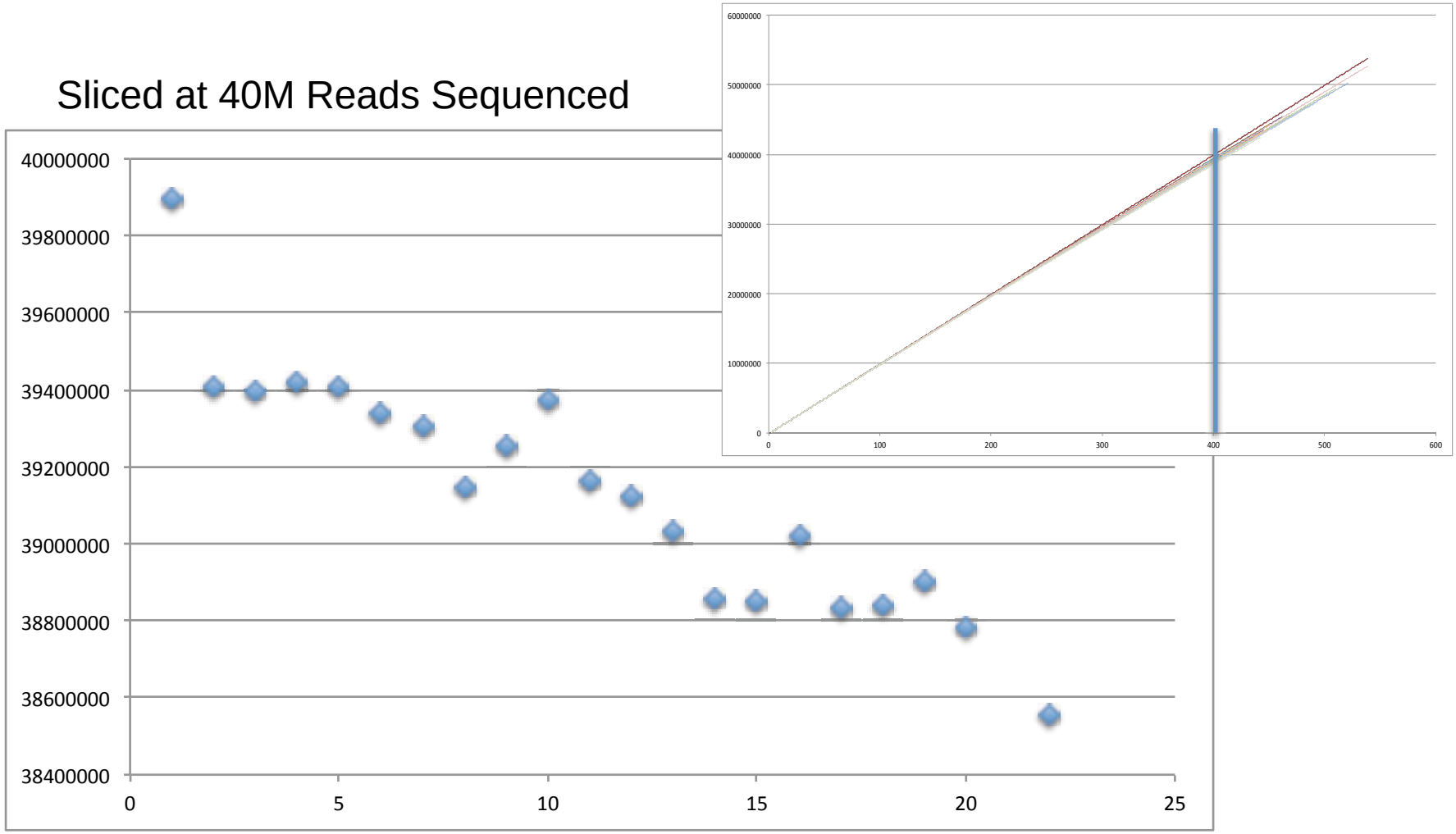
Douglas fir paired end sequencing

The total sequence generated was 1.08Tb representing 58.3x sequence coverage of an 18.6gb (O'Brien 1996) Douglas fir genome.

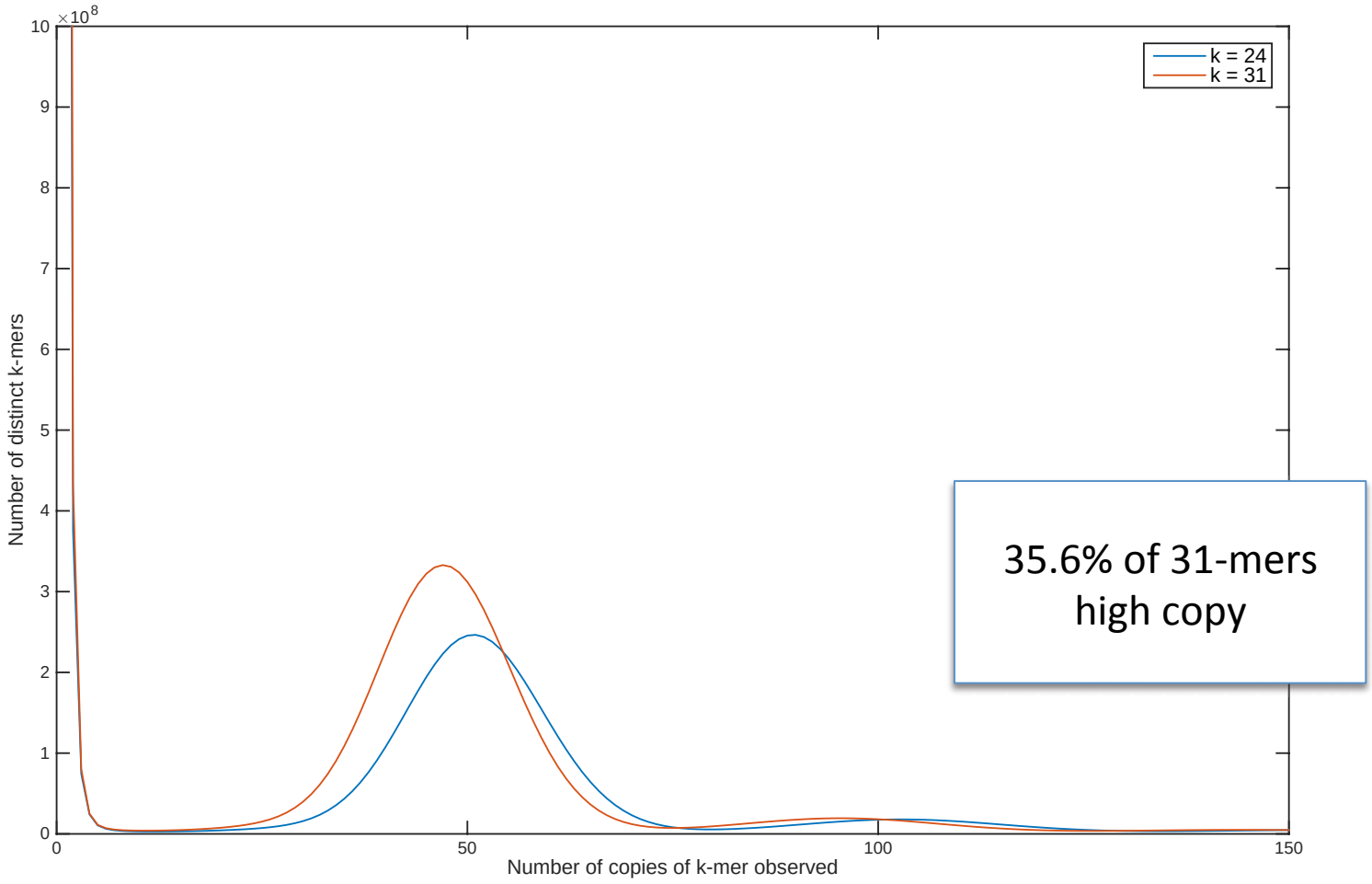
All Hiseq 2500 Rapid run 2x150

	Paired end libraries	Insert Size Range	Mbp	Sequence Coverage
	5	[200bp, 300bp)	210670	11.3
	6	[300bp, 400bp)	244902	13.2
	5	[400bp, 500bp)	207778	11.2
	5	[500bp, 600bp)	218542	11.7
	5	[600bp, 750bp)	201956	10.9
Totals	26		1083848	58.3

Douglas fir library complexity curves



Douglas fir k-mer histograms



Douglas fir Genome Size Estimates

	k-mer length	
	24	31
Total k-mers	8.21E+11	7.63E+11
Erroneous k-mers	5.40E+09	6.92E+09
Total correct k-mers	8.15E+11	7.56E+11
E(unique k-mer depth) mode	50.75	47.07
Genome size	1.61E+10	1.61E+10
E(unique k-mer depth) mean	50.44	46.76
Genome size	1.62E+10	1.62E+10

Compare to 18.6 Gb (O' brian 1996)

Conifer Fosmid Pools

Paired and and mate pair data from one lane of HiSeq 2500 sequencing assembled with SOAP denovo.

Douglas-fir 3832 Fosmids 136.8 Mbp

Scaffolds >= 20k	Min	Q1	Median	Q3	Max	Mean	Total (Mbp)	Coverage
3704	20037	27412	32020	35149	74523	31635.2	117.2	86%
Scaffolds >= 30k								
2340	30000	32347	34280	36784	74523	35034	82.0	60%

Sugar pine 4990 Fosmids 178.1 Mbp

Scaffolds >= 20k	Min	Q1	Median	Q3	Max	Mean	Total (Mbp)	Coverage
4963	20029	27534	32529	35868	92088	31947	158.6	89%
Scaffolds >= 30k								
3214	30006	32714	34819	37445	92088	35557	114.3	64%

Loblolly pine **4600** Fosmids **164.2** Mbp

Scaffolds >= 20k	Min	Q1	Median	Q3	Max	Mean	Total (Mbp)	Coverage
3798	20006	23716	28545	33260	75791	28807	109.4	67%

For more info:

P0988: Paul *et al.*, Repeat Sequence Characterization in Sugar Pine (Pinus lambertiana) and Loblolly Pine (Pinus taeda)

Thank you