

Genome Assembly at JGI

Alicia Clum
Genomic Technologies Workshop
JGI User Meeting
March 22, 2016

Outline

- Overview
- Improving assemblies with long read technology
- Future improvements

- **Overview**
- Improving assemblies with long read technology
- Future improvements

Genome assembly review

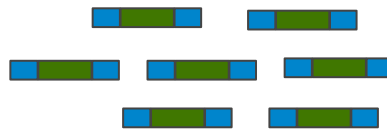
Genomic
DNA



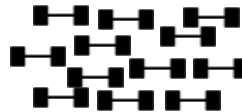
fragmentation



Library creation



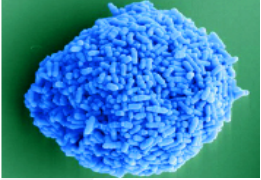
Sequencing



Assemble reads



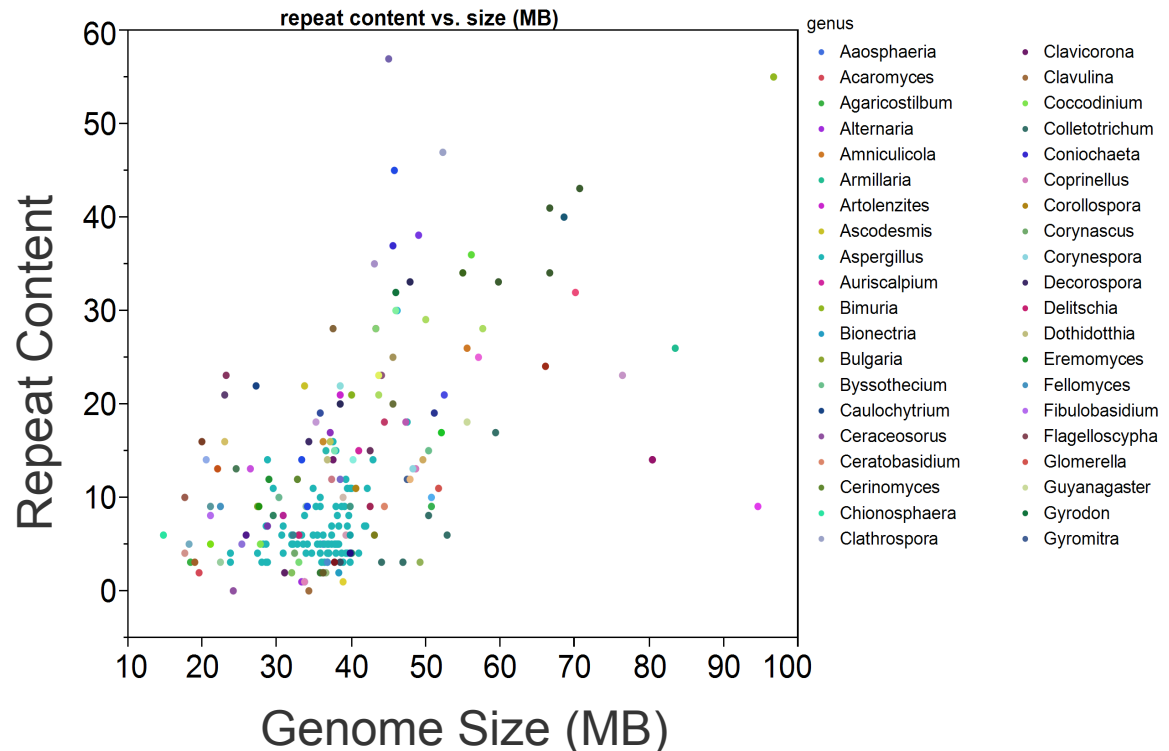
Overview of assembly at JGI

	Program	Size (MB)	Libraries	Assembler	Target assemblies / year
	Microbe	5	1	SPAdes/ HGAP	1,330
	Fungi	10's	1	ALLPATHS-LG/ Falcon	160
	Plant	100-10 000	3+	Arachne/ ALLPATHS-LG/ Falcon	20
	Metagenome	10-100 00	1	MEGAHIT	825

Challenges in genome assembly

- Repeat content
- Genome size
- GC content
- DNA quality and quantity
- Ploidy

Fungal Repeat Content vs Genome Size (MB)



- 37 MB median genome size
- 9% median repeat content

Making assemblies better

experimental

Optimizing Illumina library / sequence types
(short insert, long insert, jumping, cre-lox, inv-cre-lox, LFPE, CLRS, etc.)

computational

Sequence pre-processing
(Quality filtering, error-correction, normalization, dereplication, etc.)

computational

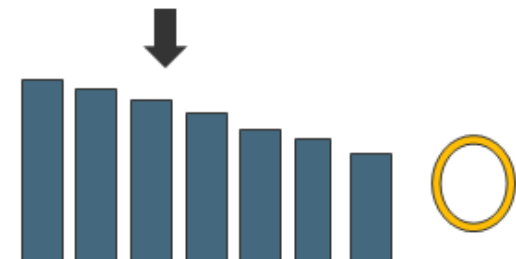
Sequence assembly
(Assembler, parameter choices)

computational

Post-assembly processing
(Scaffolding, gap closing, etc.)

experimental + computational

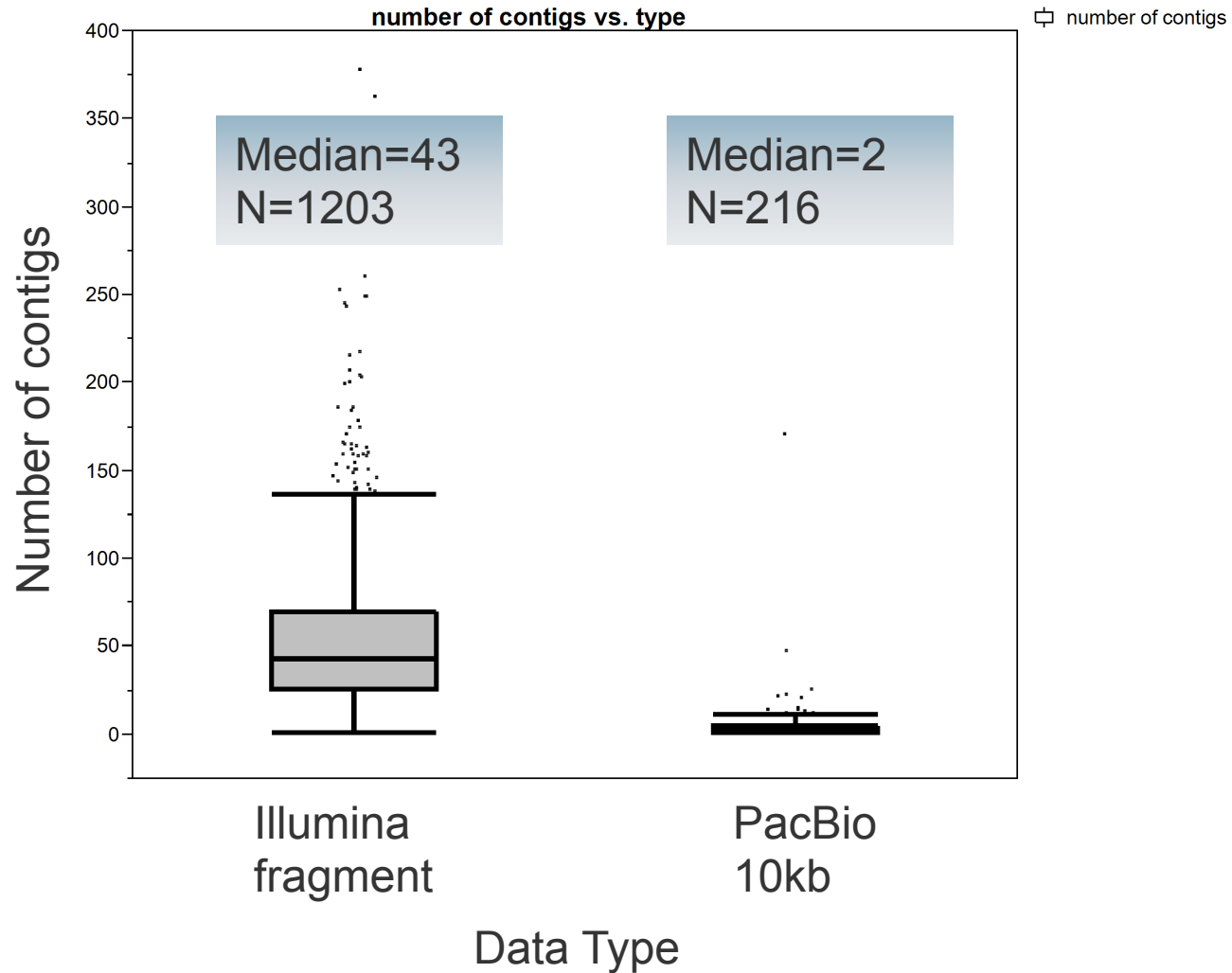
Improvement
(Gap closing, error correction)



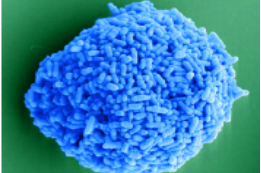
Outline

- Overview
- **Improving assemblies with long read technology**
- Future improvements

Microbial drafts- number of contigs by data type

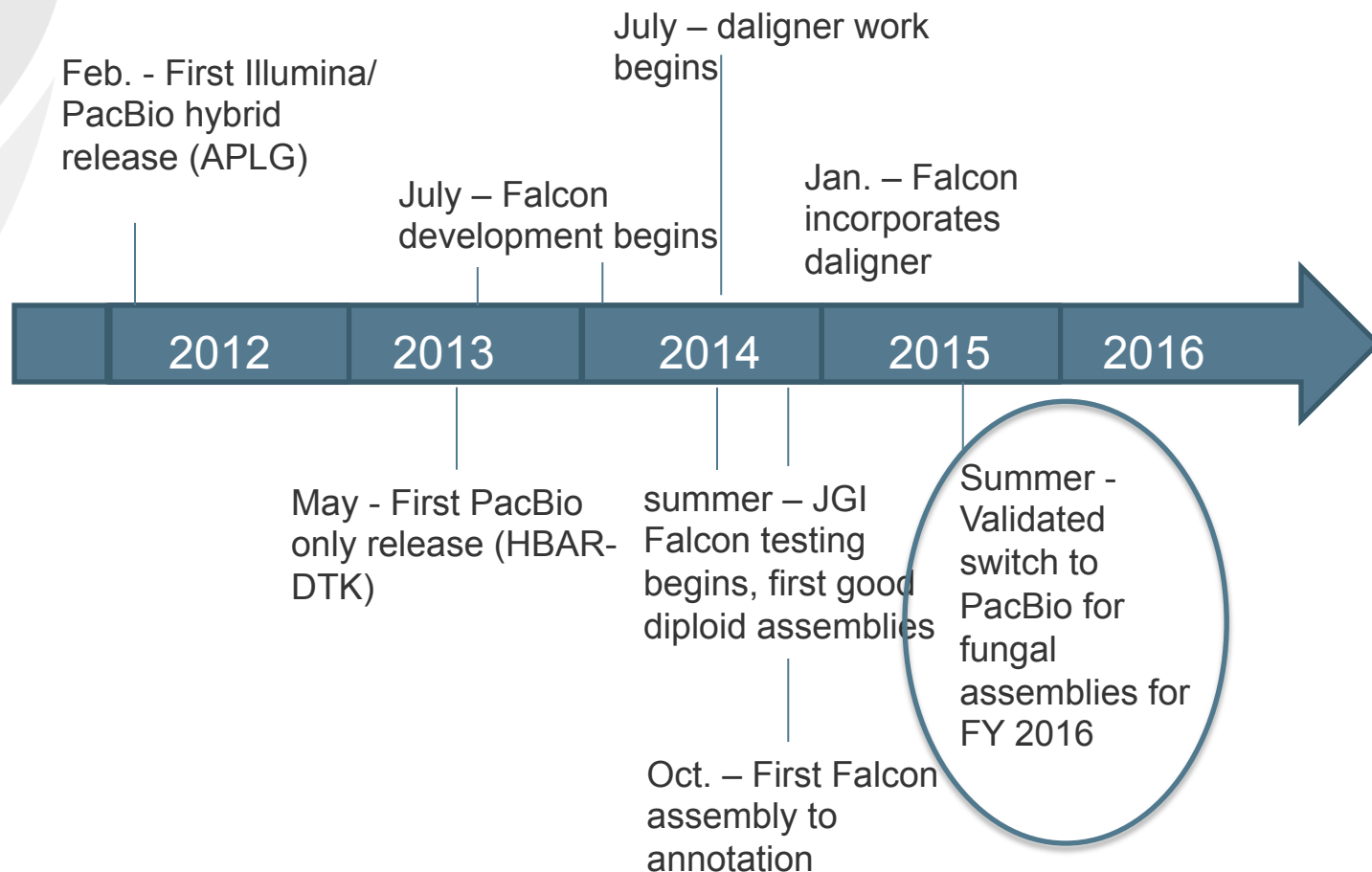


Overview of Assembly at JGI

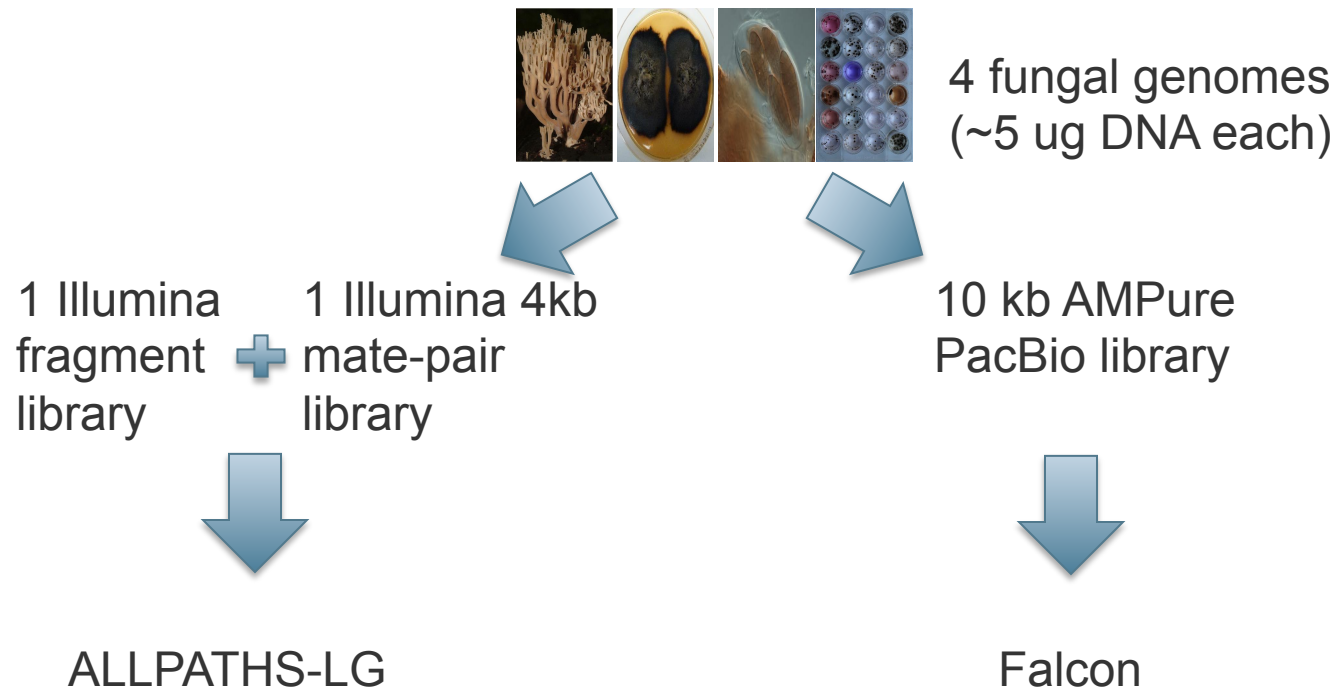
	Program	Size (MB)	Libraries	Assembler	Target genomes / year
	Microbe	5	1	SPAdes/ HGAP	1,330
	Fungi	10's	1	ALLPATHS-LG/ Falcon	160
	Plant	100-10 000	3+	Arachne/ ALLPATHS-LG/ Falcon	20
	Metagenome	10-100 00	1	MEGAHIT	825



Timeline - PacBio for fungal genomes

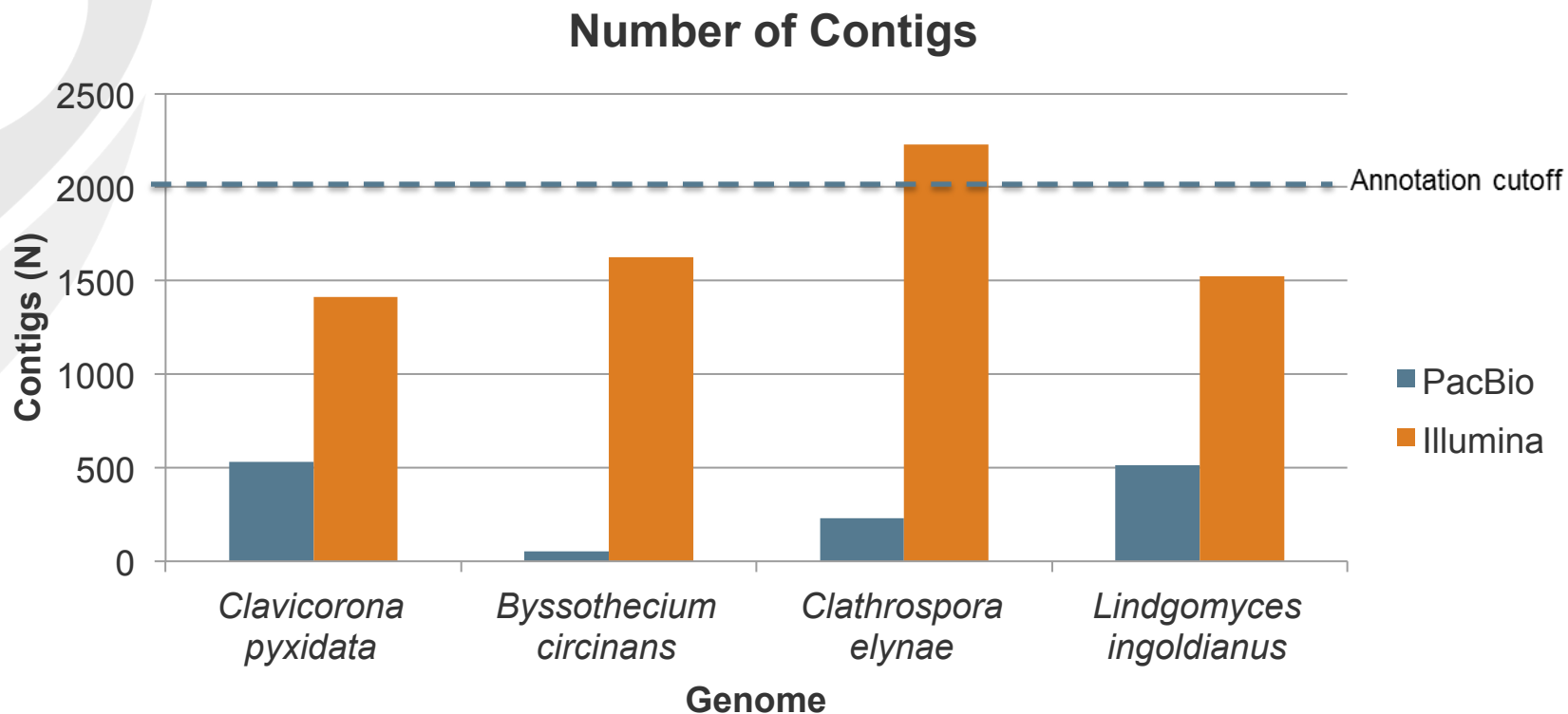


Can a single PacBio library approach produce better fungal assemblies?

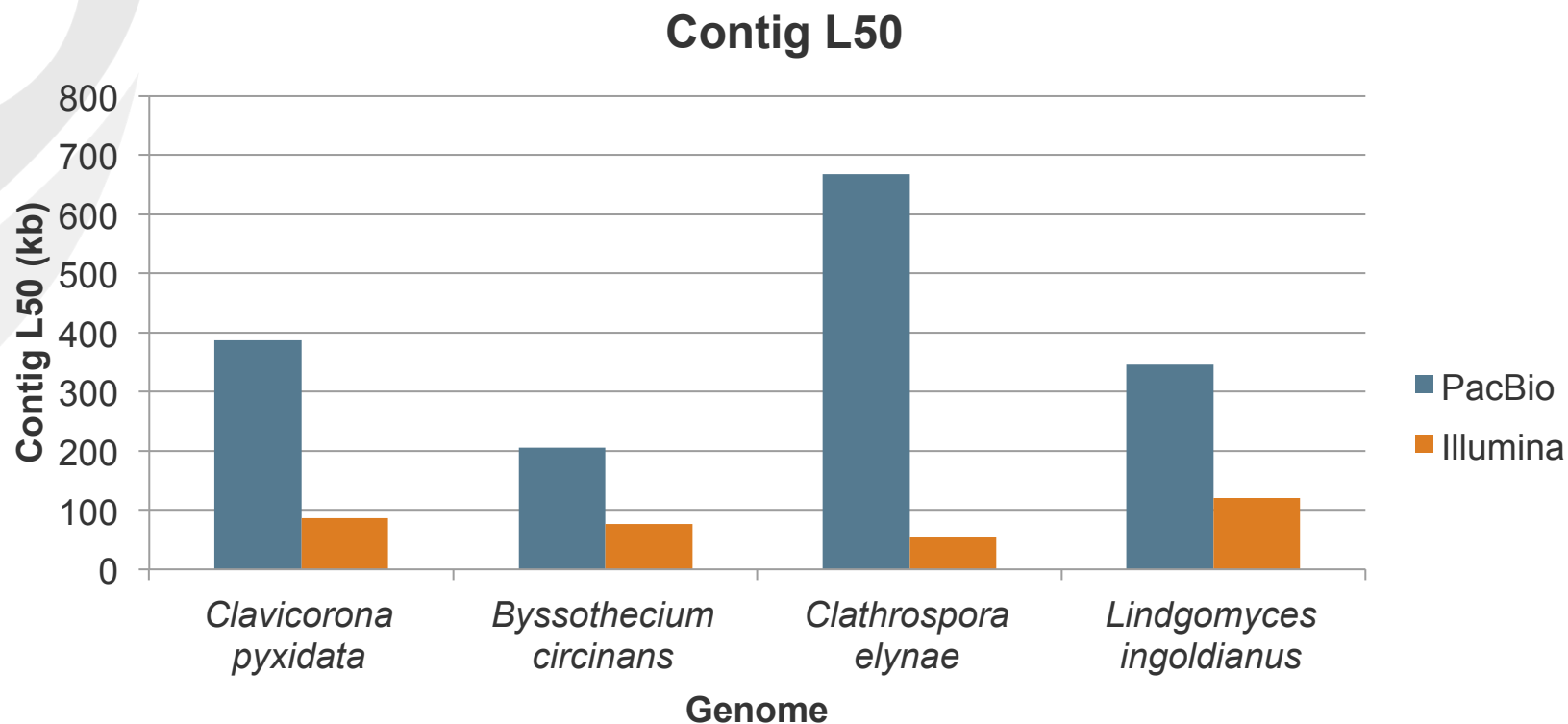


Genome	Size (MB)	Repeat Content (%)	Ploidy
<i>Clavicornia pyxidata</i>	43	14	diploid
<i>Byssothecium circinans</i>	48	15	haploid
<i>Clathrospora elynae</i>	45	47	haploid
<i>Lindgomyces ingoldianus</i>	66	20	diploid

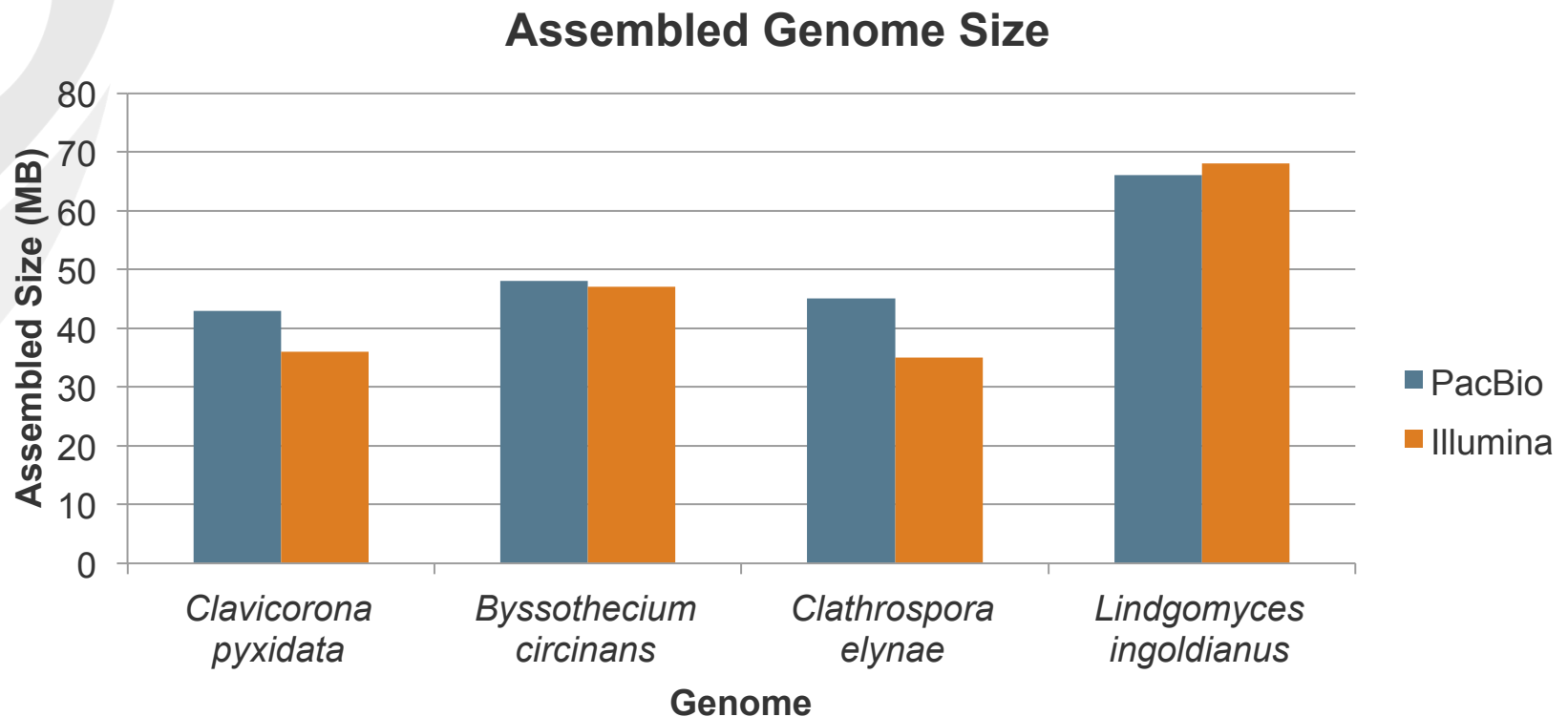
PacBio assemblies have fewer contigs



PacBio assemblies produce longer contigs

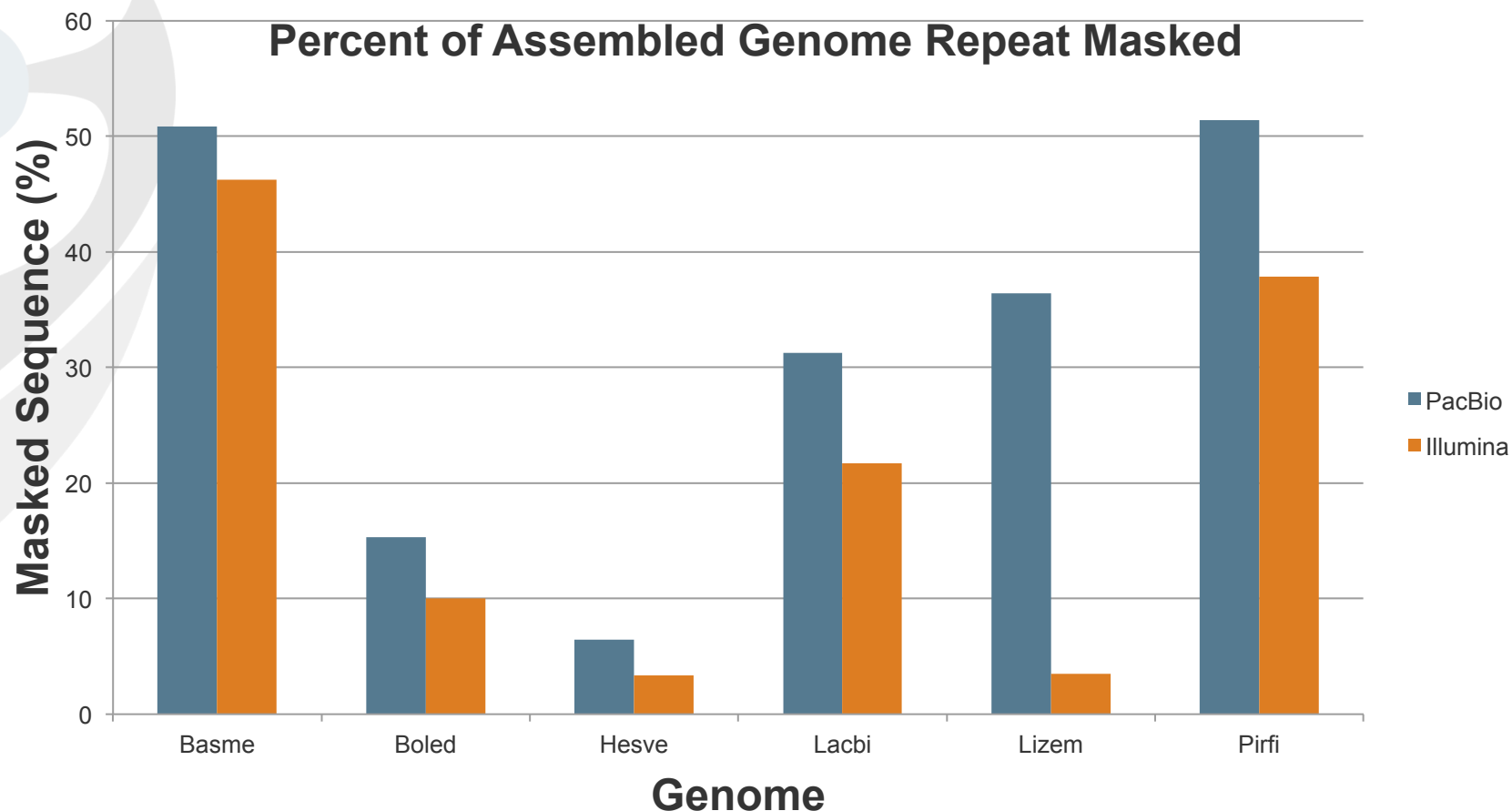


PacBio assemblies are larger



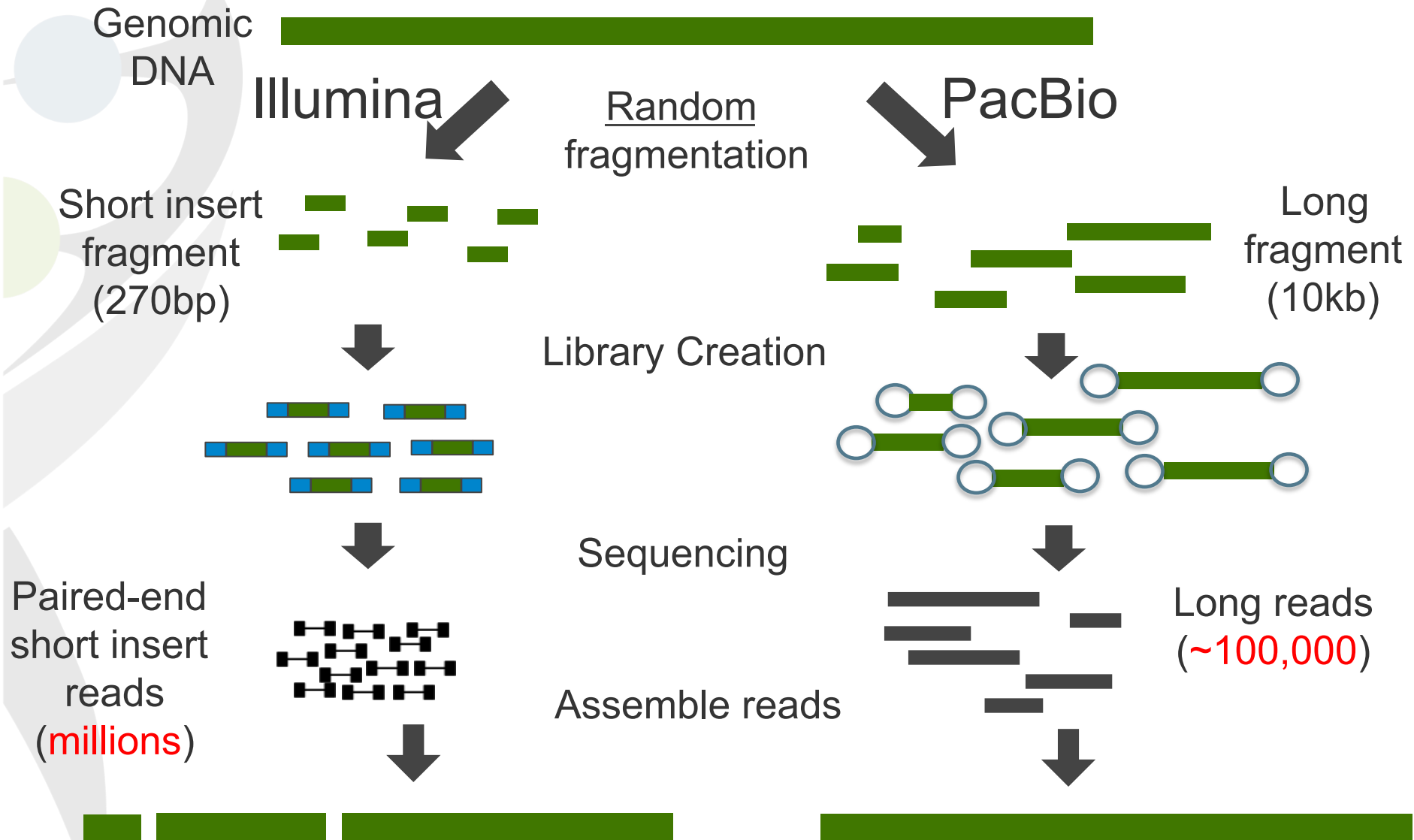
- larger assembled genome sizes representing assembled repeat content

PacBio assembles more repeat content



Median difference of 7 % between how much sequence is masked in Illumina vs. PacBio

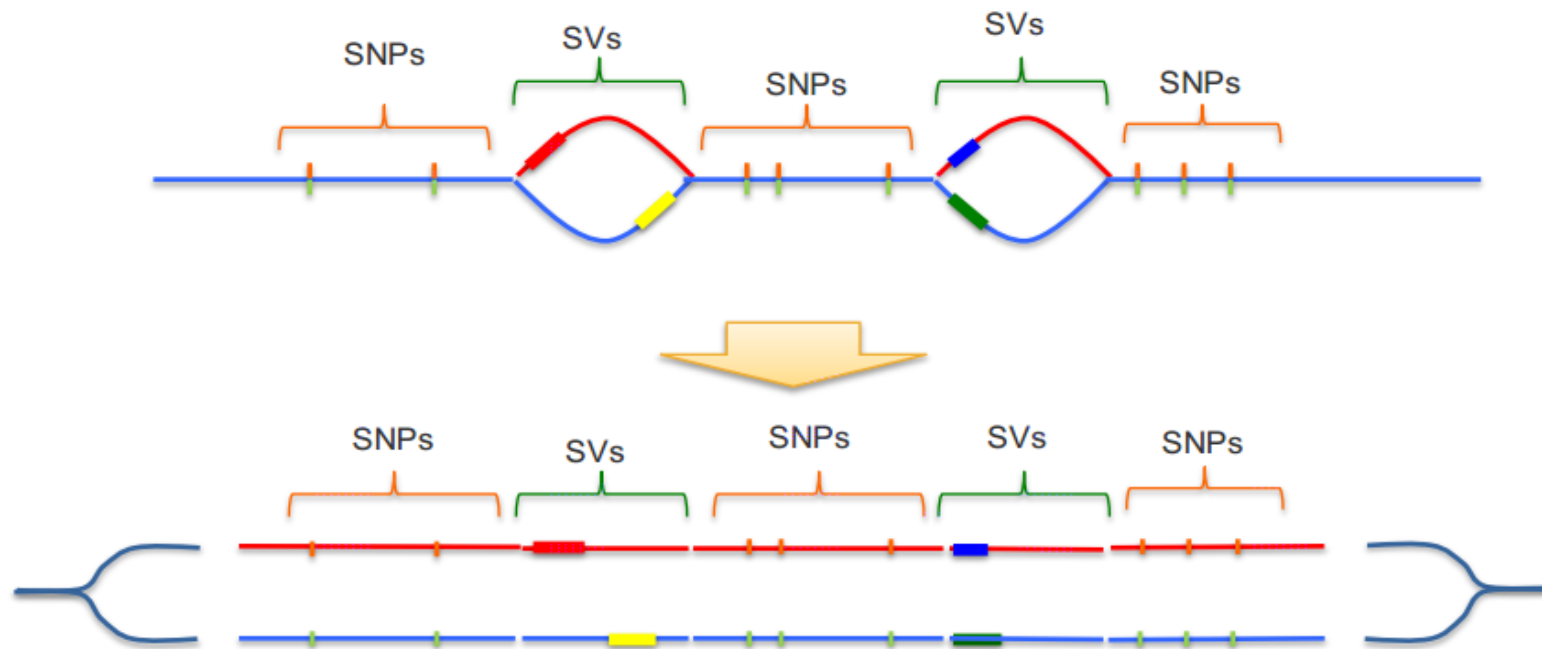
PacBio only assembly now implemented for fungal assembly



Outline

- Overview
- Improving assemblies with long read technology
- **Future improvements**

“Unzipping Collapsed Paths” With SNP Information



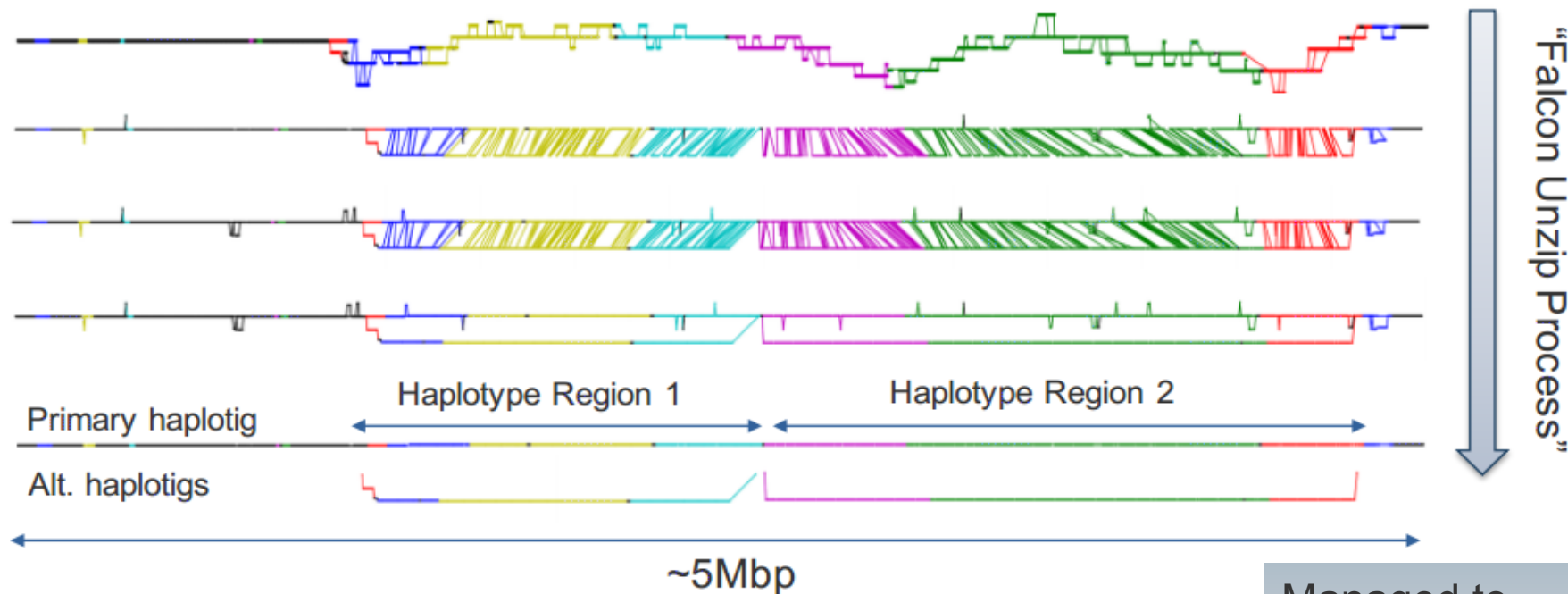
Chaining together all kinds of variants to assemble haplotigs for a diploid genome

Courtesy: Jason Chin

13

Unzip a Wild Type Diploid Fungal Genome

(*Clavicornia pyxidata* HHB10654)



Primary Haplotig Stats#
(not including contig < 25kb)

#Seqs	69
Max	5,163,978
Total	41,871,923
n50	1,915,982
n90	276,011
n95	126,158

Alt. Haplotig Stats

#Seqs	256
Max	2,085,149
Total	24,355,024
n50	891,525
n90	29,084
n95	15,599

Managed to phase >50% of the genome. JGI data with current Falcon is at < 25%.

Courtesy: Jason Chin

Conclusions

- **Assembly pipelines vary by program and input data**
- **Long read technology and assembly algorithm development have improved assembly results**
- **Continued efforts for further improvements**

Acknowledgments

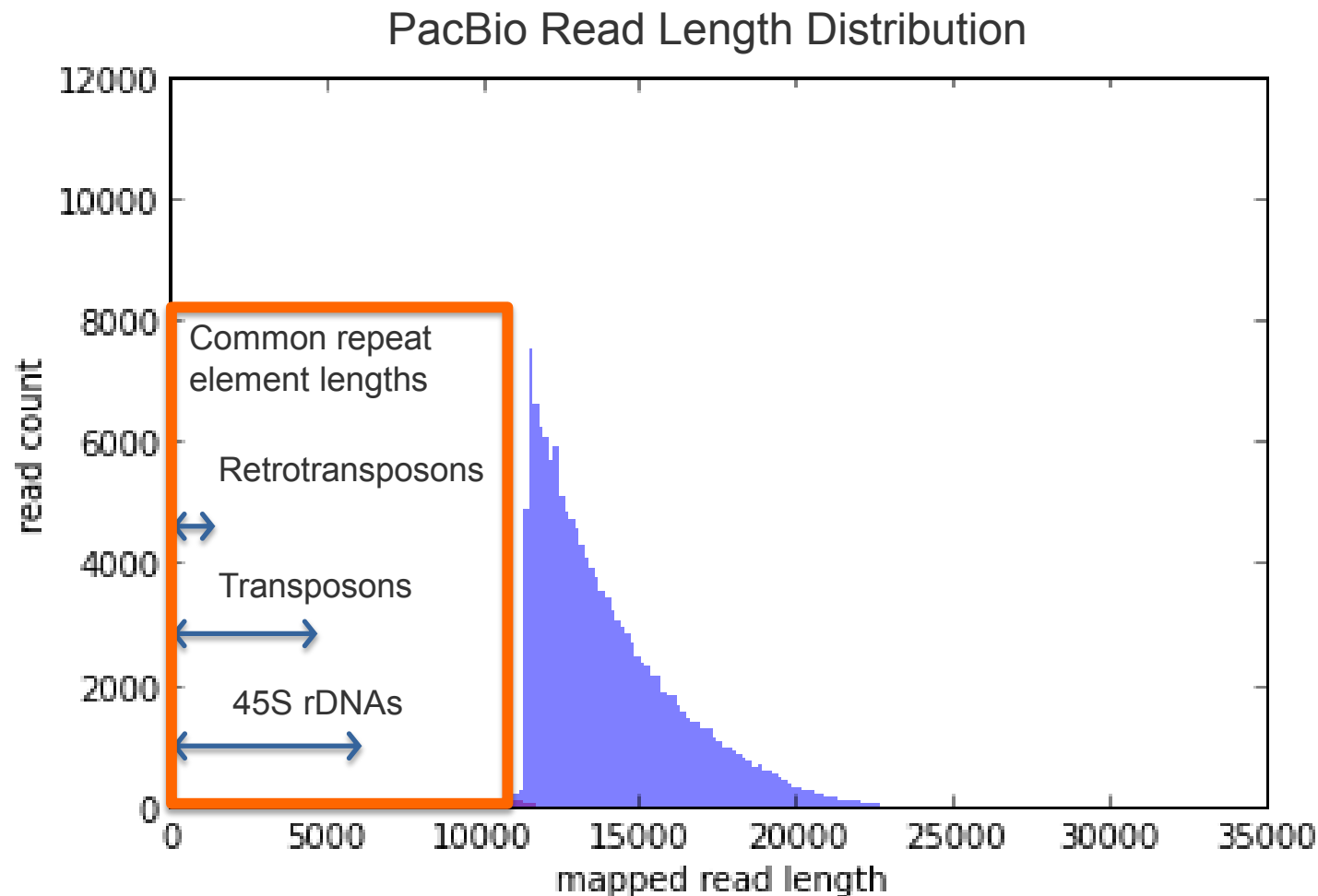


JGI
Alex Copeland
Igor Grigoriev & Fungal Annotation Group
Chris Daum & Sequencing Technologies Group
Genome Assembly & QA/QC Groups

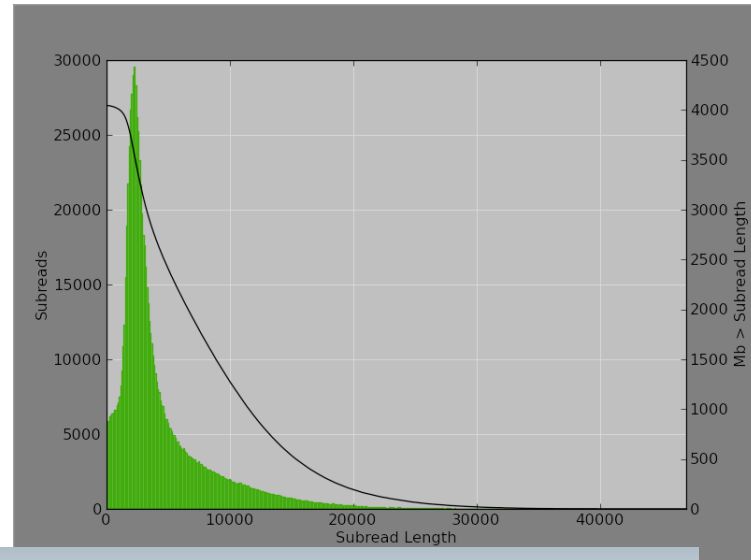
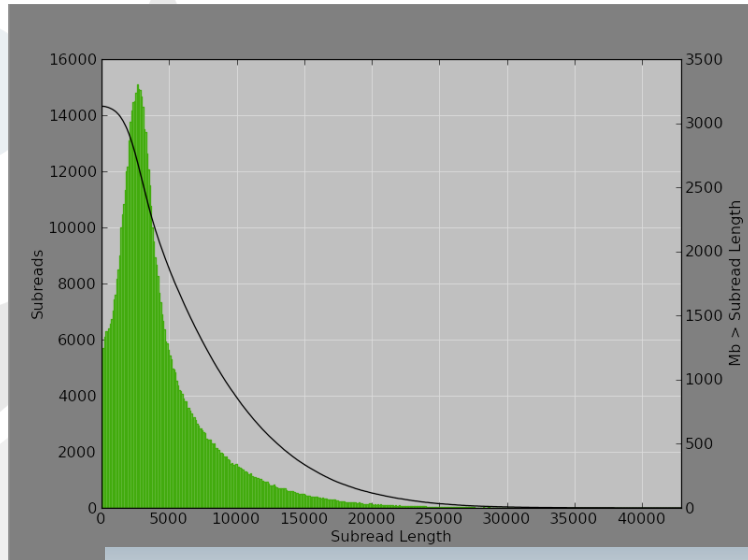
Pacific Biosciences
Jason Chin
Paul Peluso
David Rank
Kristi Spittle

Supplement

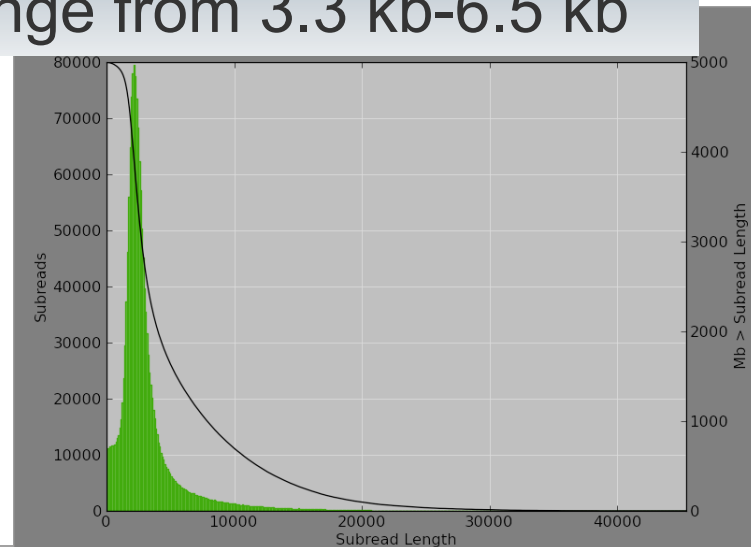
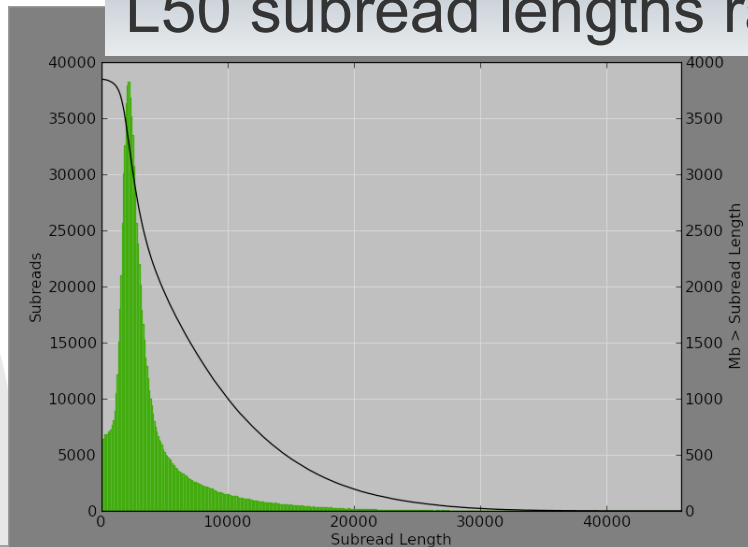
Long Reads Span Common Repetitive Elements



>10kb AMPure Subread Lengths



L50 subread lengths range from 3.3 kb-6.5 kb



Evaluating Assemblers

nucleotid.es



an assembler catalogue

continuous, objective and reproducible evaluation of genome
assemblers using docker containers