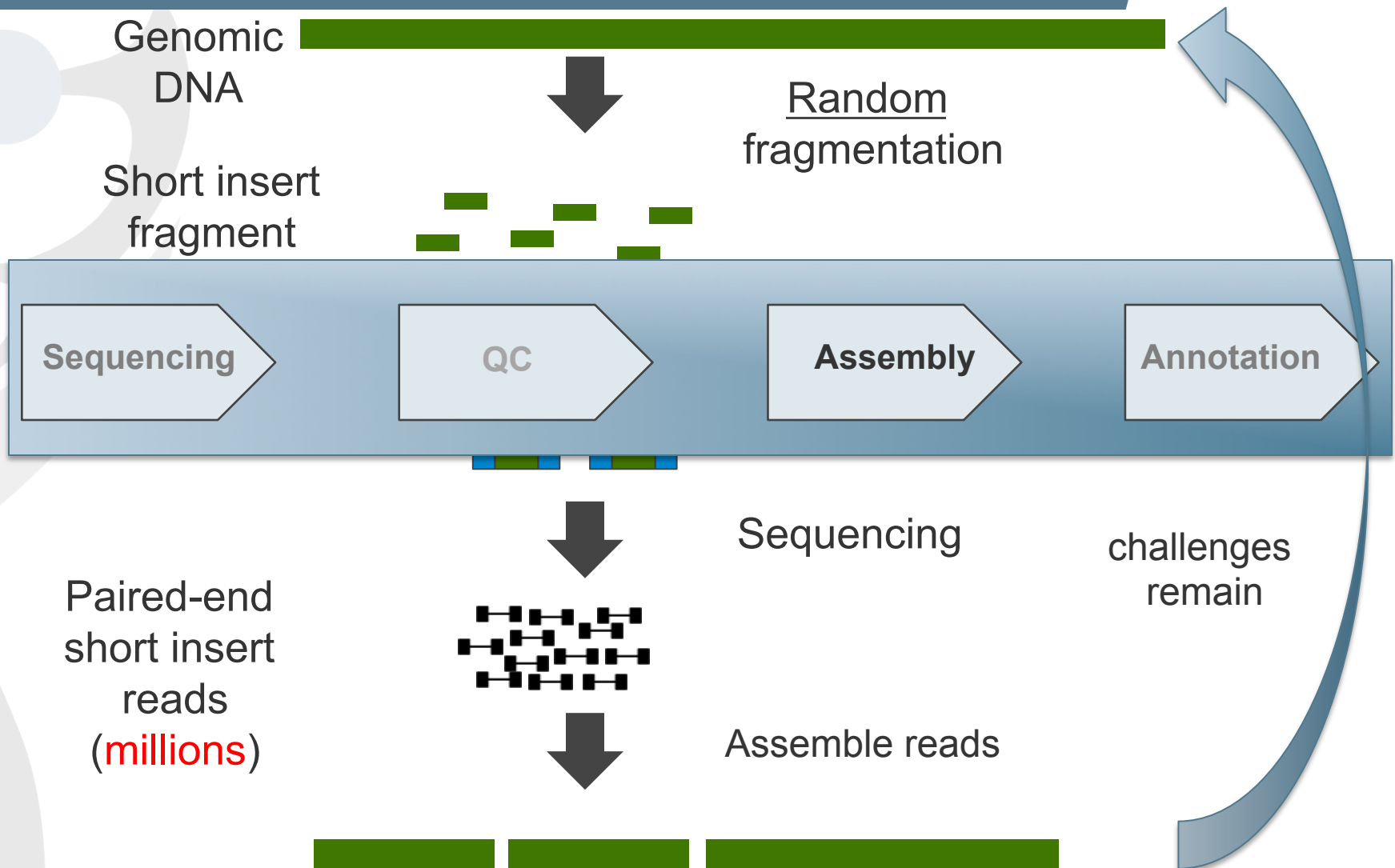


Genome Assembly at JGI

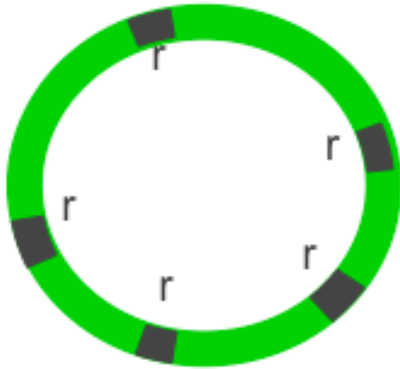
Alicia Clum
Genomic Technologies Workshop
JGI User Meeting
March 24, 2015

Genome Assembly Review



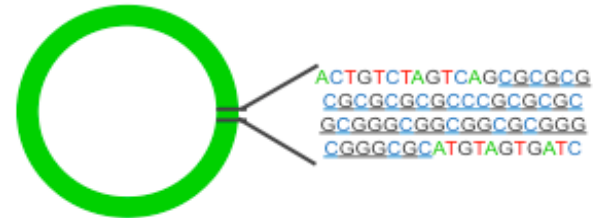
What Causes Gaps in Assemblies?

High repeat content

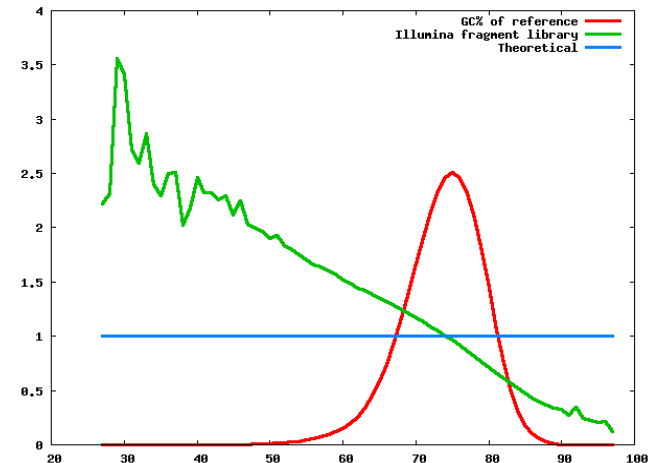


RESULT:
misassemblies / collapsed assemblies

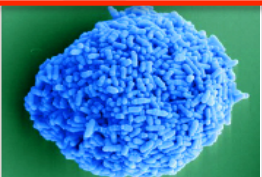
Biased sequence composition



RESULT:
incomplete / fragmented assembly
GC bias in reads:
↑%GC, ↓ Read Representation



Overview of Assembly at JGI

	Program	Size (MB)	Libraries	Assembler	Target
	Microbe	5	1	ALLPATHS-LG/ HGAP	1,000
	Fungi	10's	1-3	ALLPATHS-LG/ HGAP/Falcon	110
	Plant	100-10 000	3+	Arachne/ ALLPATHS-LG	20
	Metagenome	10-100 00	1	MEGAHIT	580

Long Read Sequencing Improves Microbial Genome Assembly

PacBio RS



Genomic DNA

Random
fragmentation

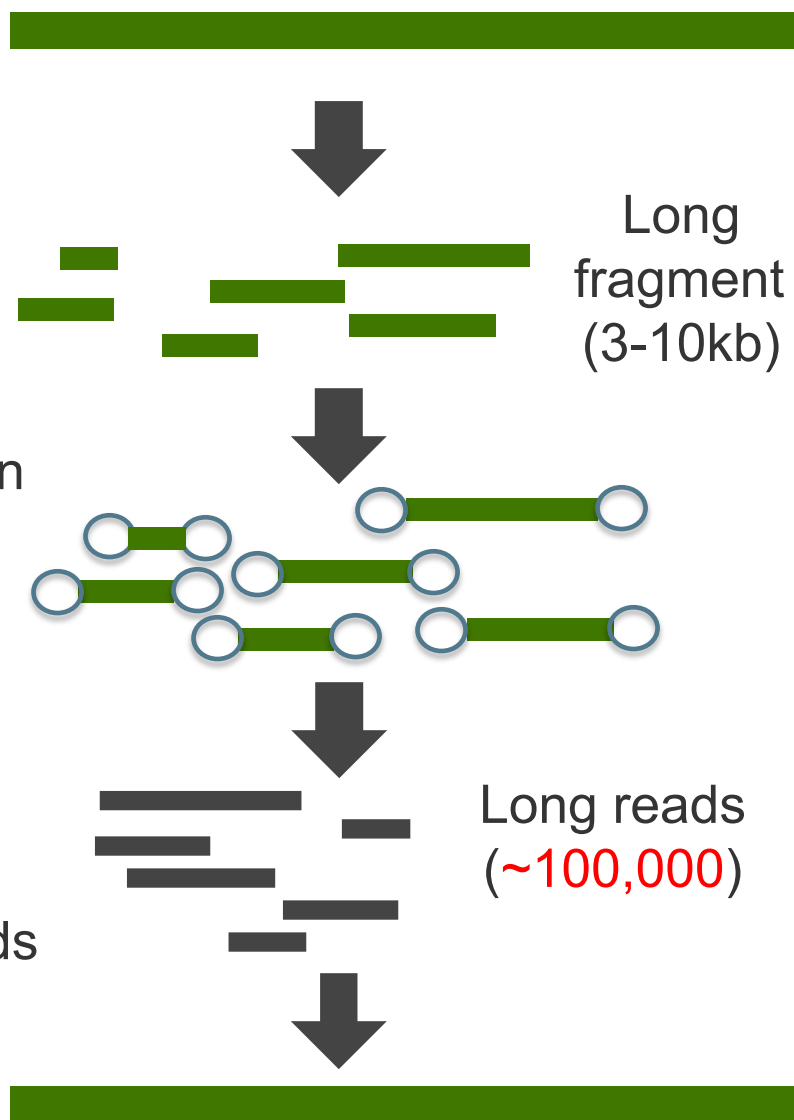
Library Creation

Sequencing

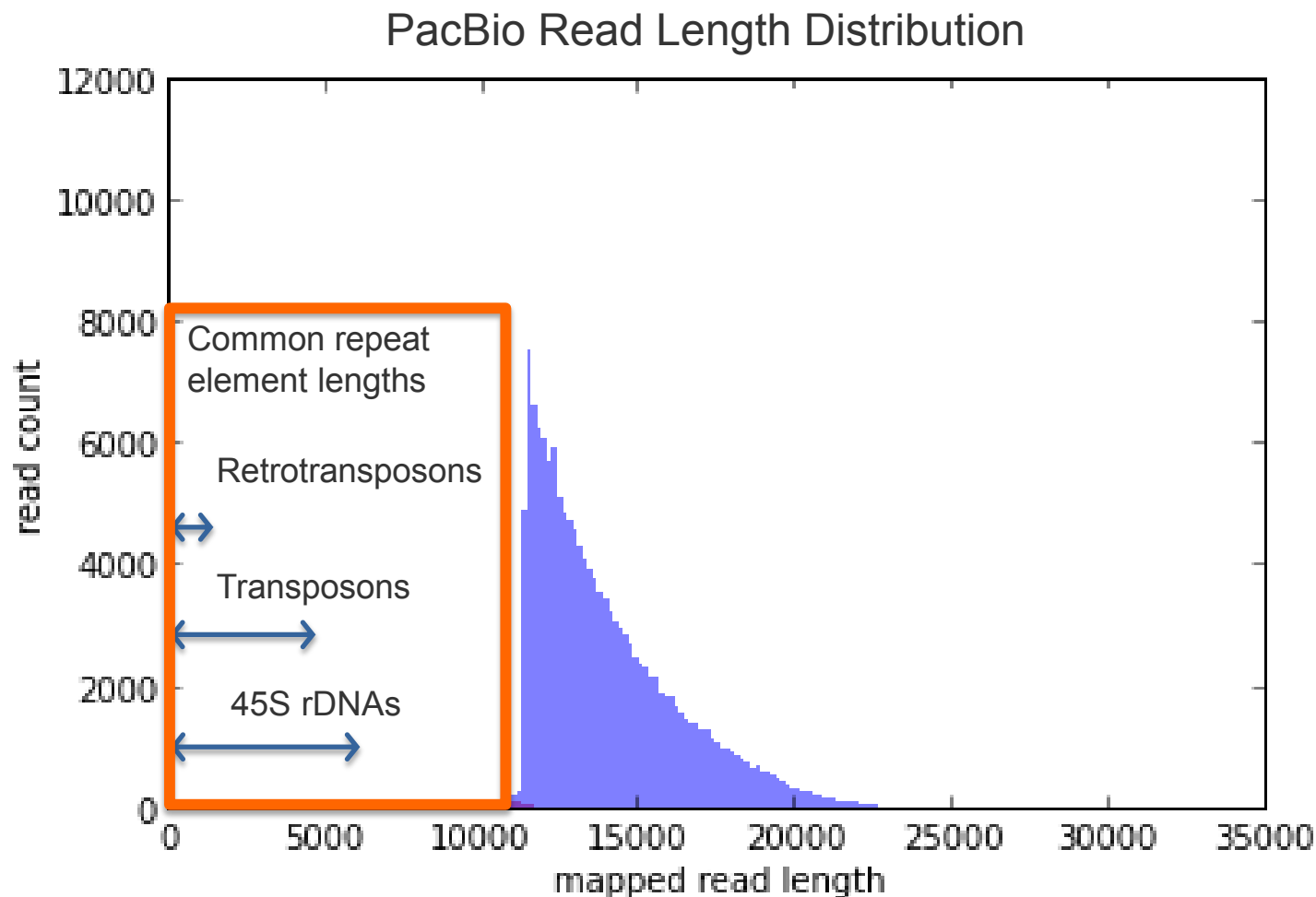
Assemble reads

Long
fragment
(3-10kb)

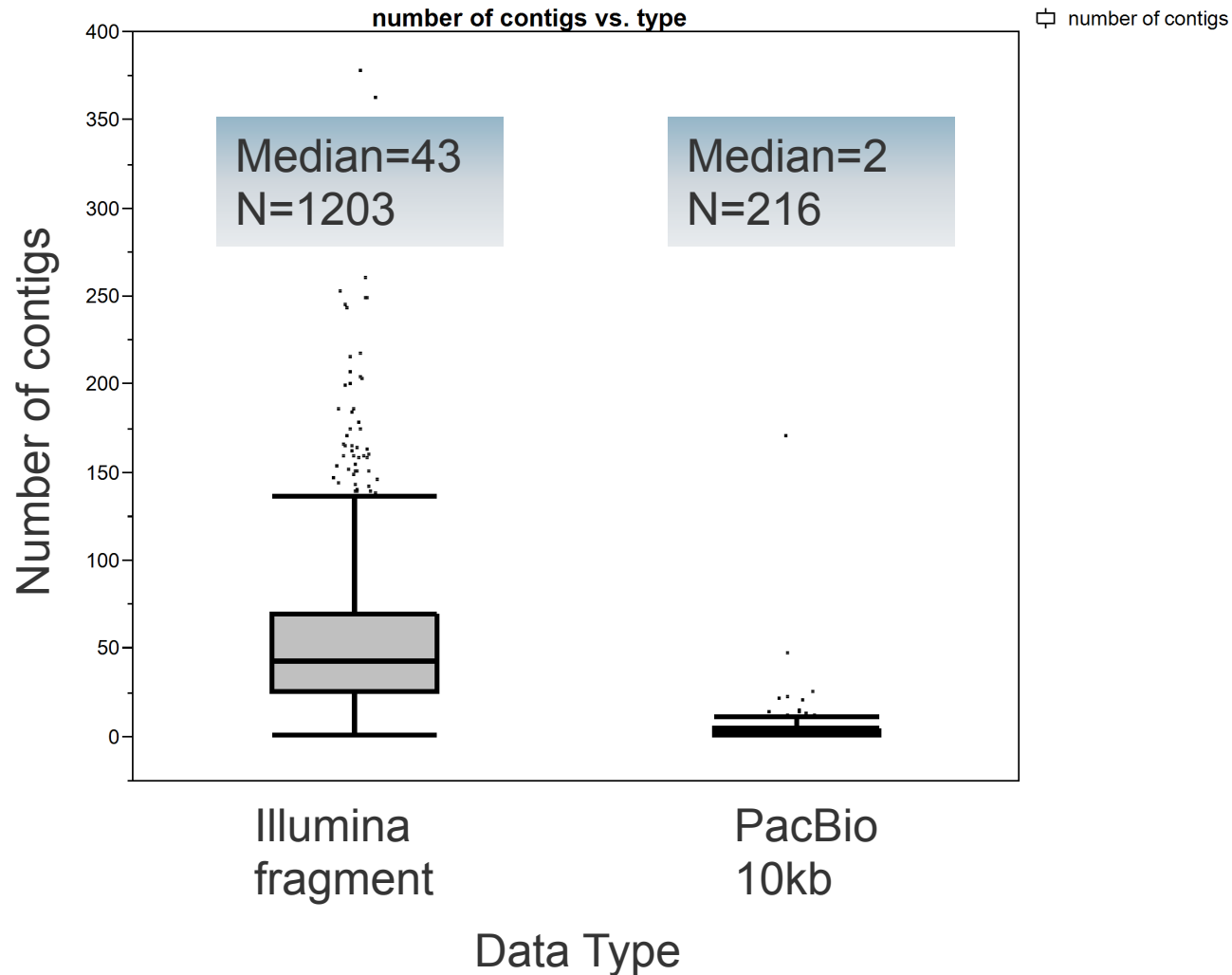
Long reads
(~100,000)



Long Reads Span Common Repetitive Elements in Bacteria

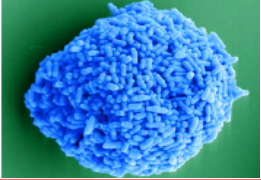


Microbial Drafts- Number of Contigs by Data Type



42% of microbial genomes assemble into a single contig with PacBio!

Overview of Assembly at JGI

	Program	Size (MB)	Libraries	Assembler	Target
	Microbe	5	1	ALLPATHS-LG/ HGAP	1,000
	Fungi	10's	1-3	ALLPATHS-LG/ HGAP/Falcon	110
	Plant	100-10 000	3+	Arachne/ ALLPATHS-LG	20
	Me	Challenges of eukaryotic genomes: - Longer repeats - polyploidy			580

New technologies enable recovery of very long reads from eukaryotic samples

- **P6/C4 chemistry and 4 hour movies**
 - Average read lengths 10,000-15,000 bp
 - Throughput ~ 550 MB/SMRT cell
- **BluePippin size selected 20kb libraries**



Genome	Size selection	Mean subread length (bp)	Subread L50 (bp)
<i>E.coli</i>	AMPure	3,100	4,300
<i>E.coli</i>	BluePippin	10,000	15,500

*Data on AMPure and BluePippin size selection provided by PacBio

PacBio Computational Improvements: Falcon



PacificBiosciences / **FALCON**

Watch 33 Star 30 Fork 15

branch: master **FALCON** / doc / falcon_manual.md

pb-jchin 12 days ago fix a typo for the build script

3 contributors

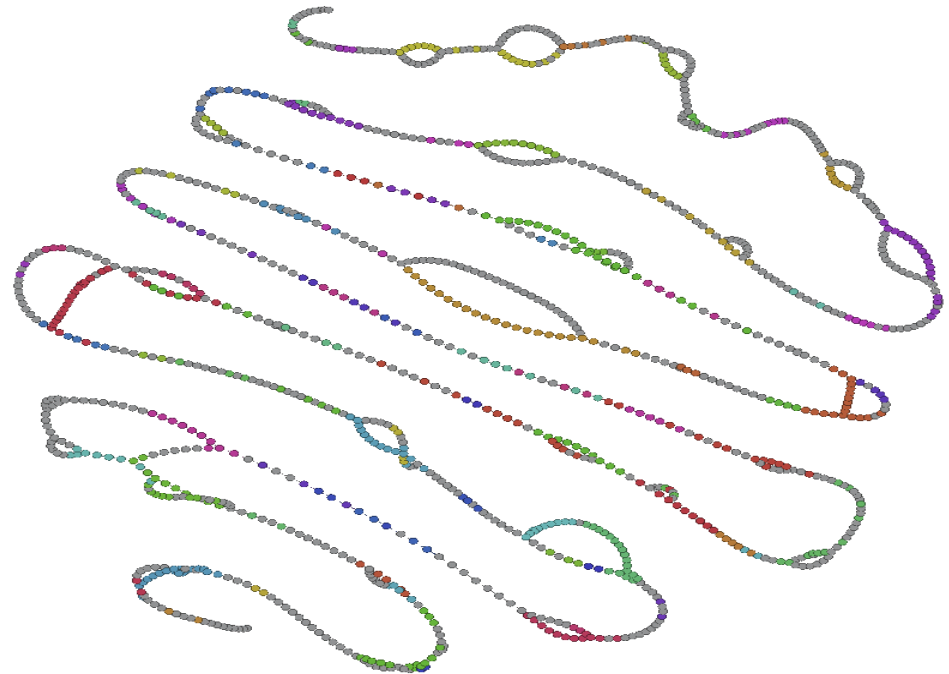
692 lines (536 sloc) | 32.118 kb



Falcon Genome Assembly Tool Kit Manual for v0.2 Branch

Author: Jason Chin Date: Dec 4, 2014 Date: Feb 2, 2015

Minor Update: Dec 17, 2014



Falcon Results:

Melampsora allii-populina 12AY07

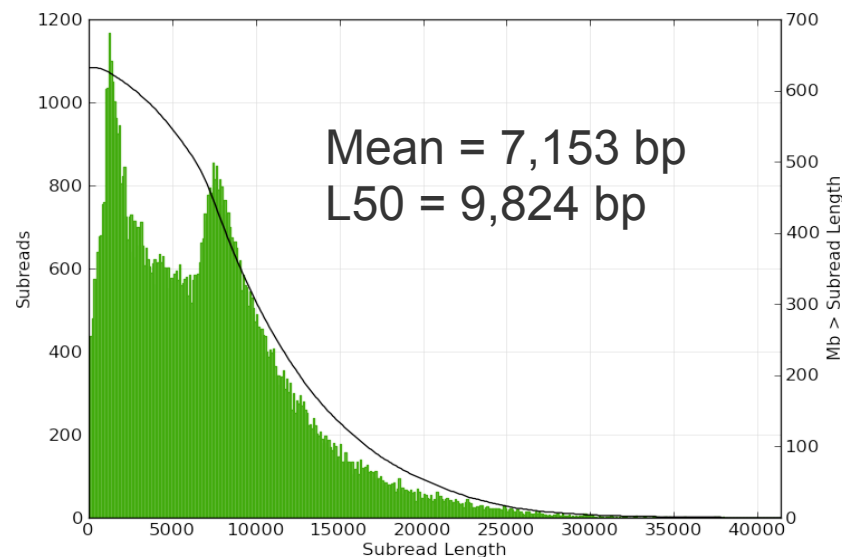
Genome properties:

Diploid

Large (158-347 MB)

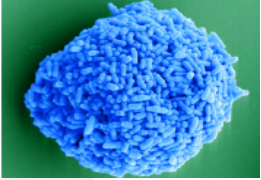
53% repeat

40% GC



	ALLPATHS-LG Illumina (frag+LMP)	Falcon 20kb PacBio
Scaffolds	7,013	3,679
Contigs	27,307	3,679
Scaffold L50 (kb)	49	210
Contig L50 (kb)	10	210
Scaffold (Mbp)	158	293
Contig (Mbp)	123	293
Gap (%)	21	0
EST capture (%)	91.12	97.34

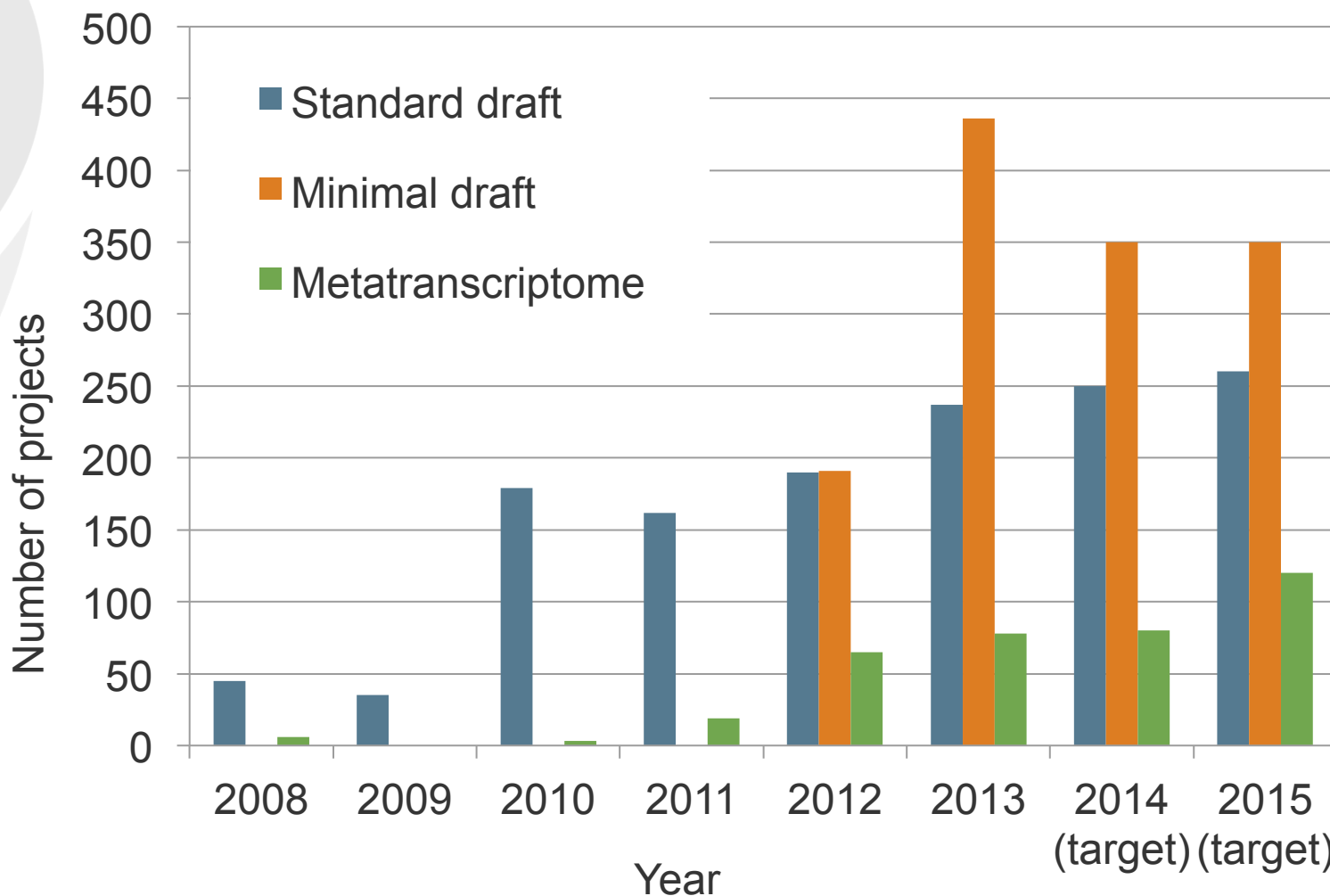
Overview of Assembly at JGI

	Program	Size (MB)	Libraries	Assembler	Target
	Microbe	5	1	ALLPATHS-LG/ HGAP	1,000
	Fun			ALLPATHS-LG/	110
	Plan				20
	Metagenome	10-100 00	1	MEGAHIT	580

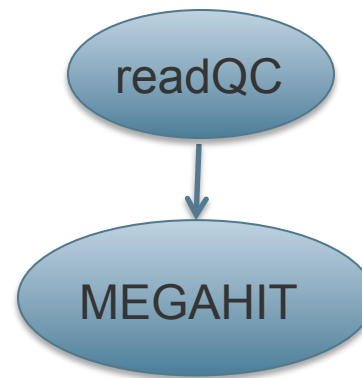
Challenges of metagenomes:

- All of above
- Complexity (10,000s genomes)
- Massive data sizes

Number of Metagenome Projects by Year



Metagenome Assembly Pipeline



Advantages of MEGAHIT

- **Comparing results for 23 projects**
 - Fewer points of failure
 - Completed 6 hours faster
 - Scheduled 15 hours faster
 - More reads mapped back to the assembly
 - Larger assembled genome size
 - No loss in contiguity

nucleotid.es



an assembler catalogue

- Easy to compile and run software
- Easy to set up benchmarks
- Can be extended to other bioinformatics tools

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

For more information about nucleotid.es visit Michael Barton's poster Wednesday evening

Conclusions

- **Assembly pipelines vary by program and input data**
- **Laboratory and computational improvements in long read technology have improved assembly results**
- **Evaluating assemblers is important**
 - Including compute time and resources

Acknowledgements



Libraries

Chew Yee Ngan & libraries team

QC and Assembly

Kecia Duffy

Alex Spunde

Brian Foster

Alex Copeland

Michael Barton

PacBio (JGI)

Chris Daum

Jennifer Chiniquy

PacBio

Jason Chin

Nicole Rapicavoli