Incorporating New Sequencing Technologies Into Finishing Strategy

Sean M. Sykes

2007 Finishing In The Future
Santa Fe, NM
06/19/2007

Overview

- Problems With Conventional Methods

- Roche/454 Brings Something New To The Table: Both Good And Bad

- What Can We Do About 454 Limitations?

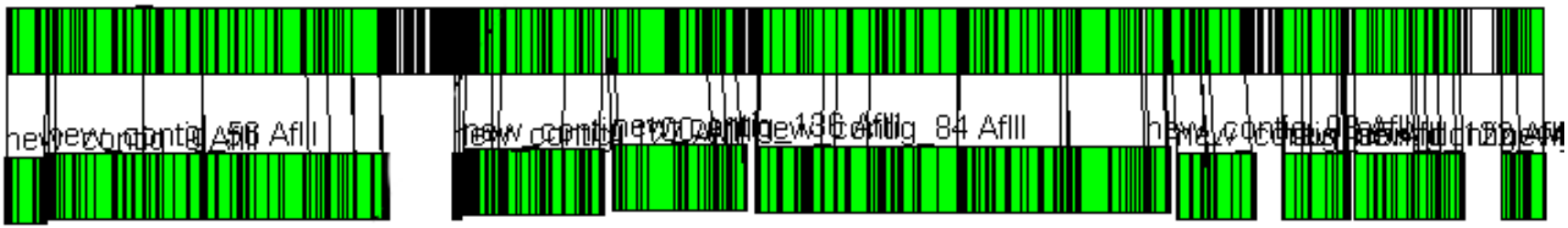
Neurospora Finishing

39.2 Mb Finished Quality Sequence

- > 20x Sequence Coverage
- > 100x Physical Coverage
  - 8 library types

41 Mb In Optical Map

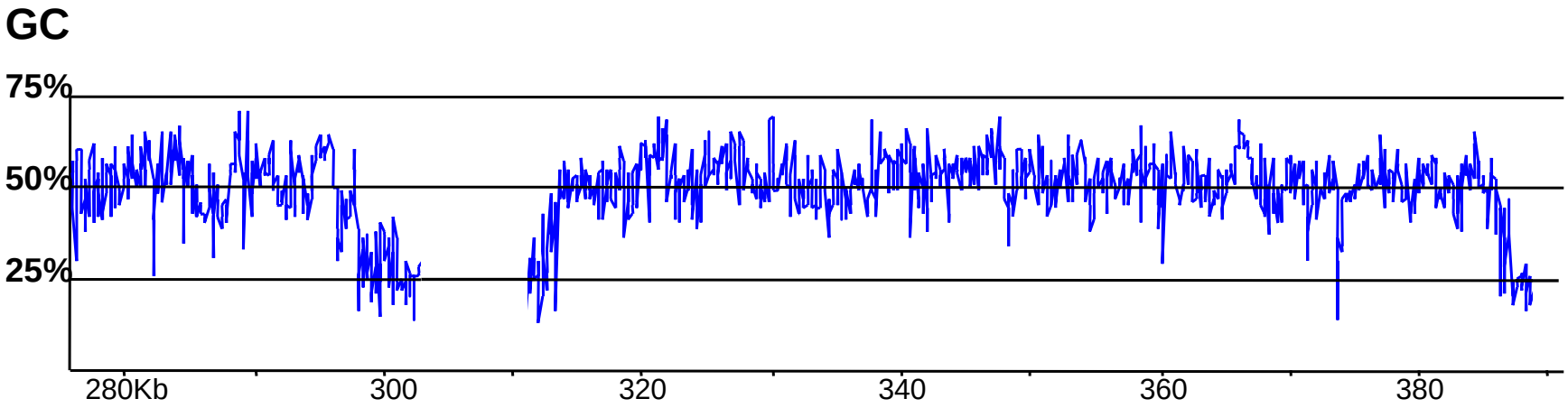
Neurospora Chromosome 2



Neurospora Gaps In “Finished” Assembly

~180 Gaps Remain

- Unclonable
- Low % GC



GC Distribution (120 kb)

What Does 454 Provide?

<u>24x 454 Assembly</u>

41.5 Mb Total Length

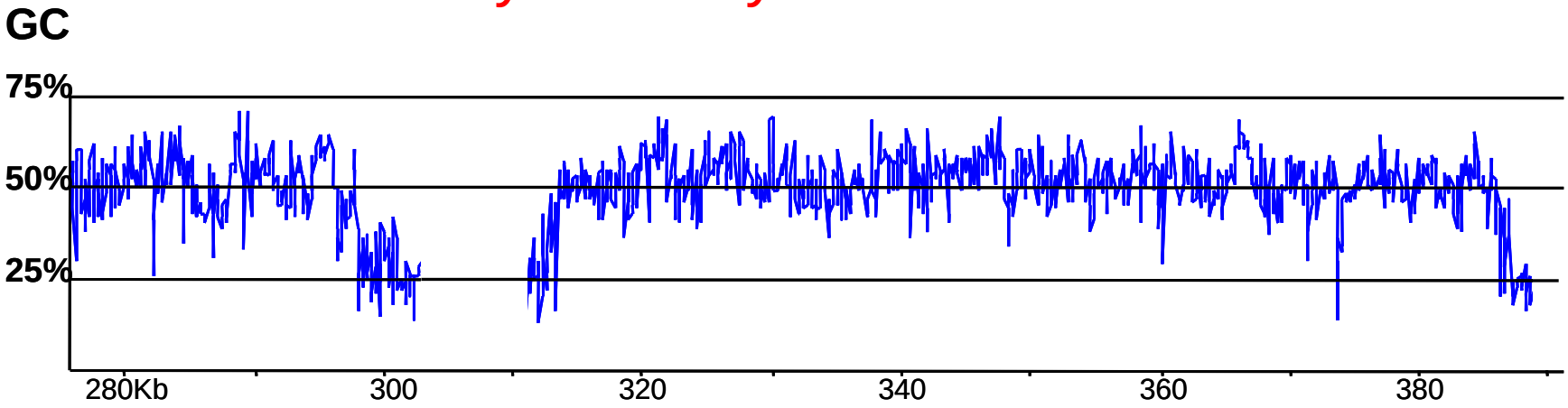
5017 Contigs

Adds ~2 Mb of new sequence

Does 454 Improve *Neurospora* Assembly?

41.5 Mb 24x 454 Assembly

- Finished Territory: 44% GC
- 454 Only Territory: 27% GC



- Spans 36 gaps
- Extends 200 additional ends

Who Else Needs Help?

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Listeria monocytogenes

2.9 Mb Genome

	ABI
Coverage	8X
% Genome	80%
Contigs	315

- Typical 8x draft assembly covers >= 95%
- Significant cloning bias here

454 Gets Stuff ABI Misses

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Yes indeed!

	ABI	454
Coverage	8X	23X
% Genome	80%	95%
Contigs	315	51

Listeria story like *Neurospora*?

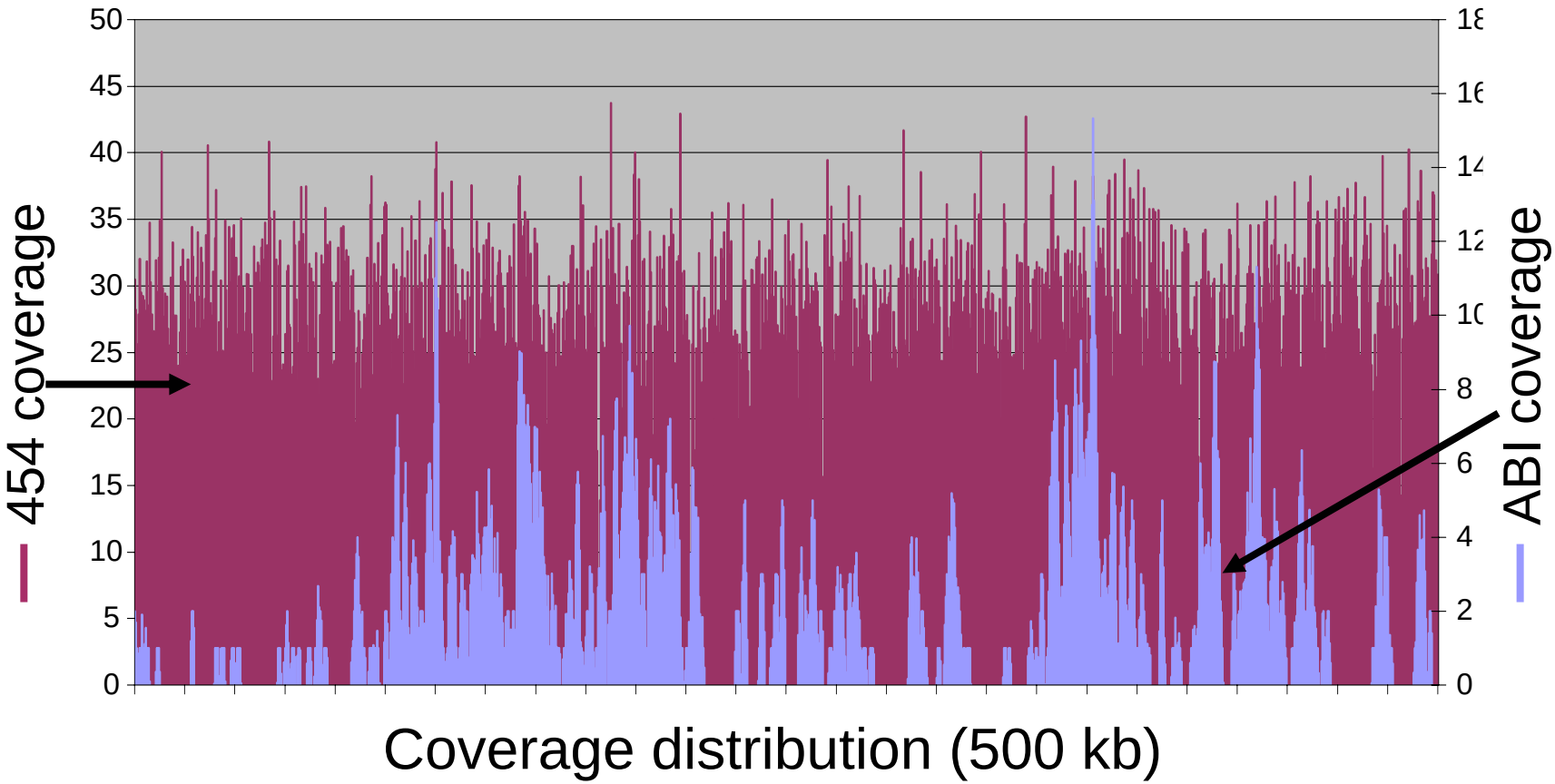
Yes

- unclonable regions sequenced by 454

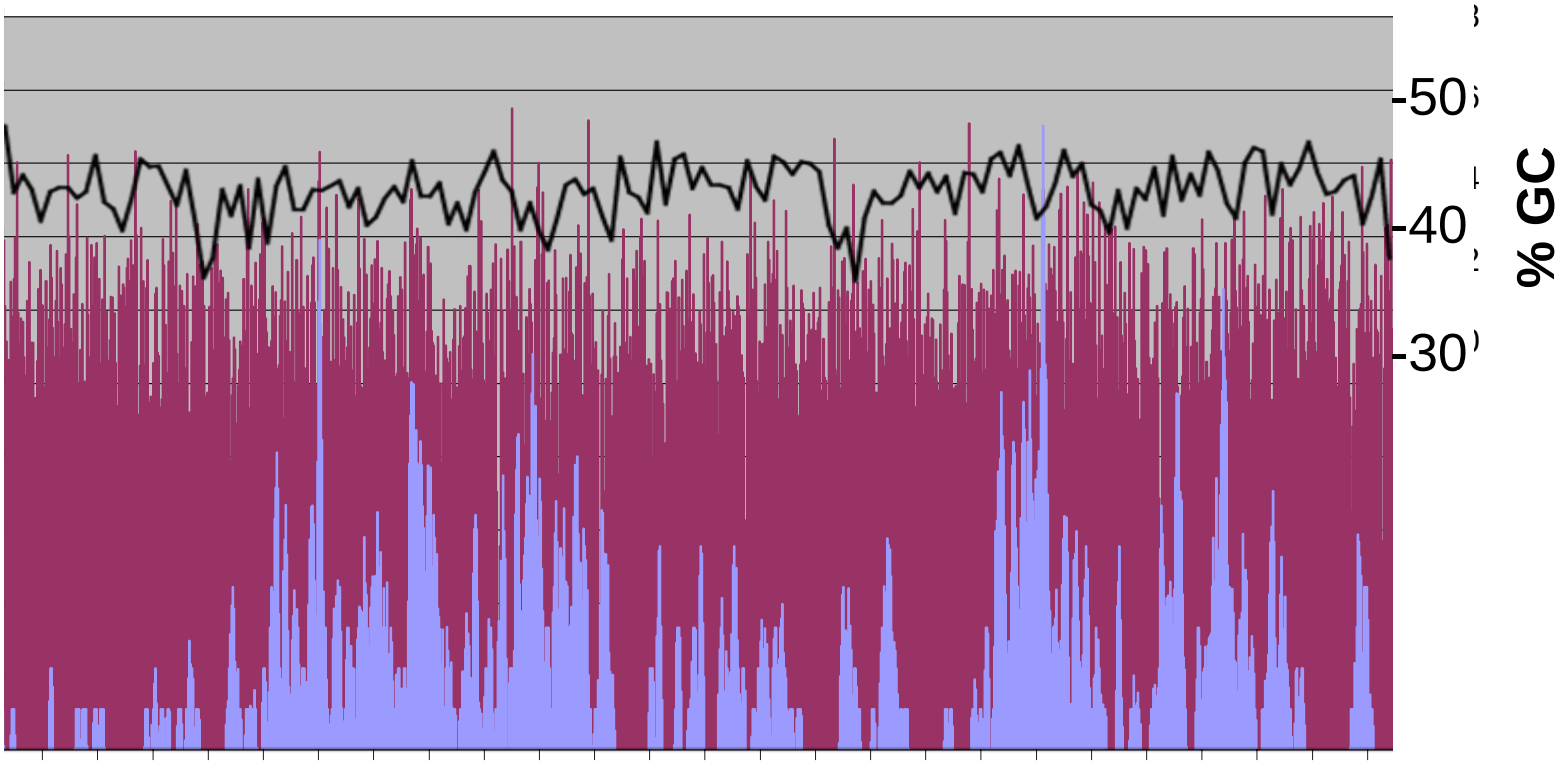
No

- not correlated to GC content

454 Captures Regions of Cloning Bias



Bias Does Not Correlate To GC



Coverage distribution (500 kb)

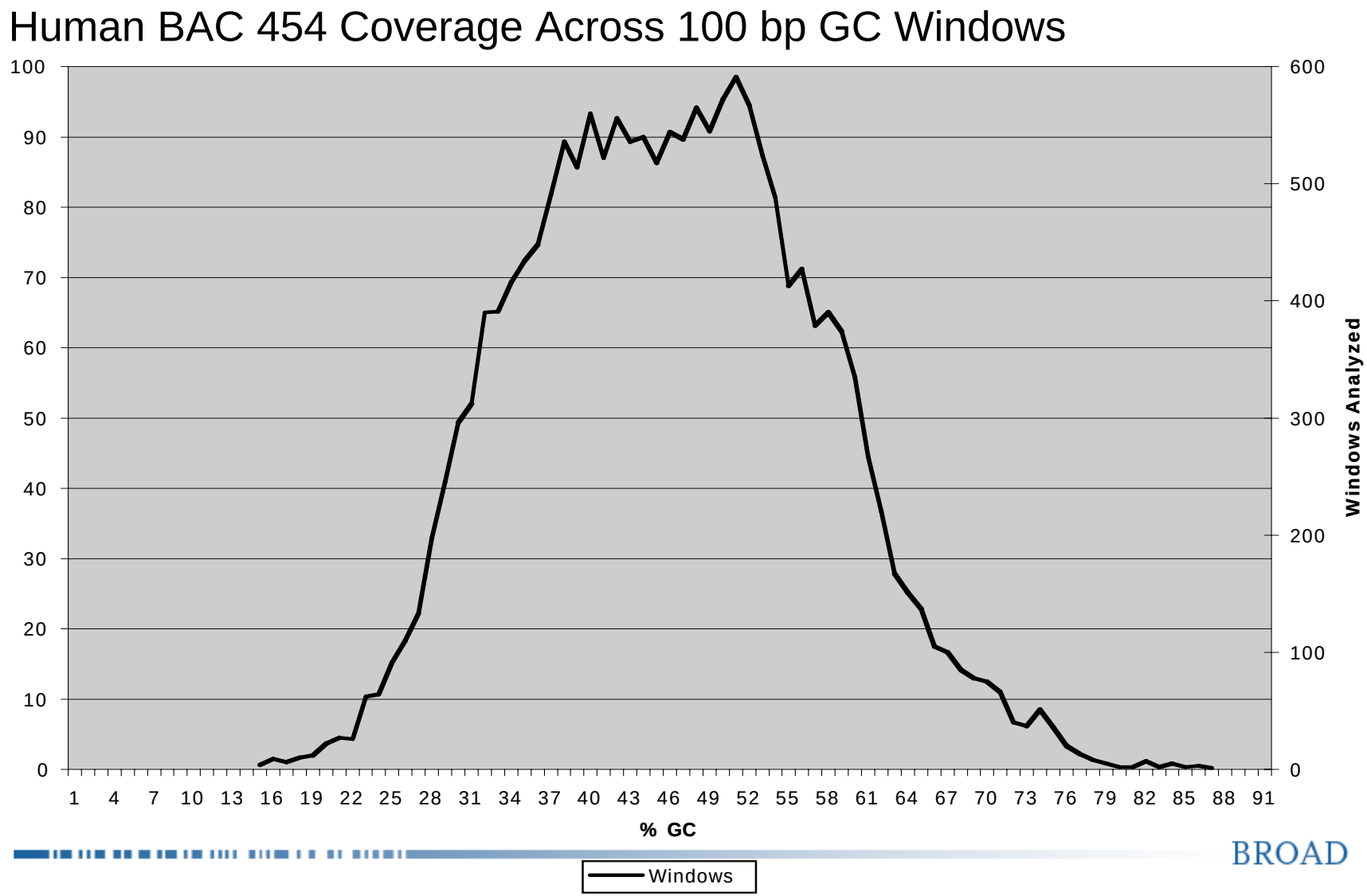
454 Works…End Of Story?

Not Really…

<u>Some Questions Remain:</u>

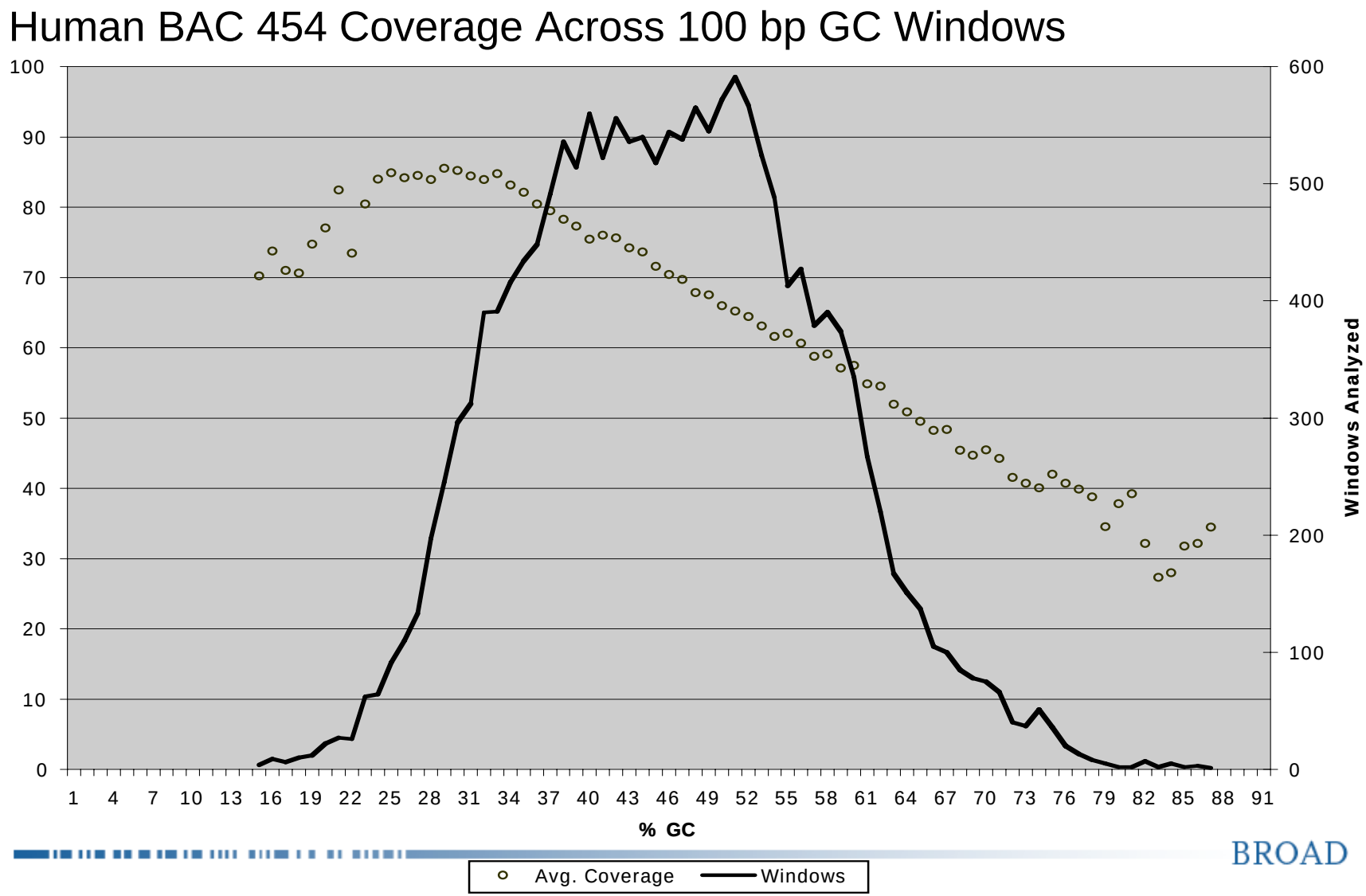
- Different Bias?
- Accuracy?

454 Coverage: A Closer Look



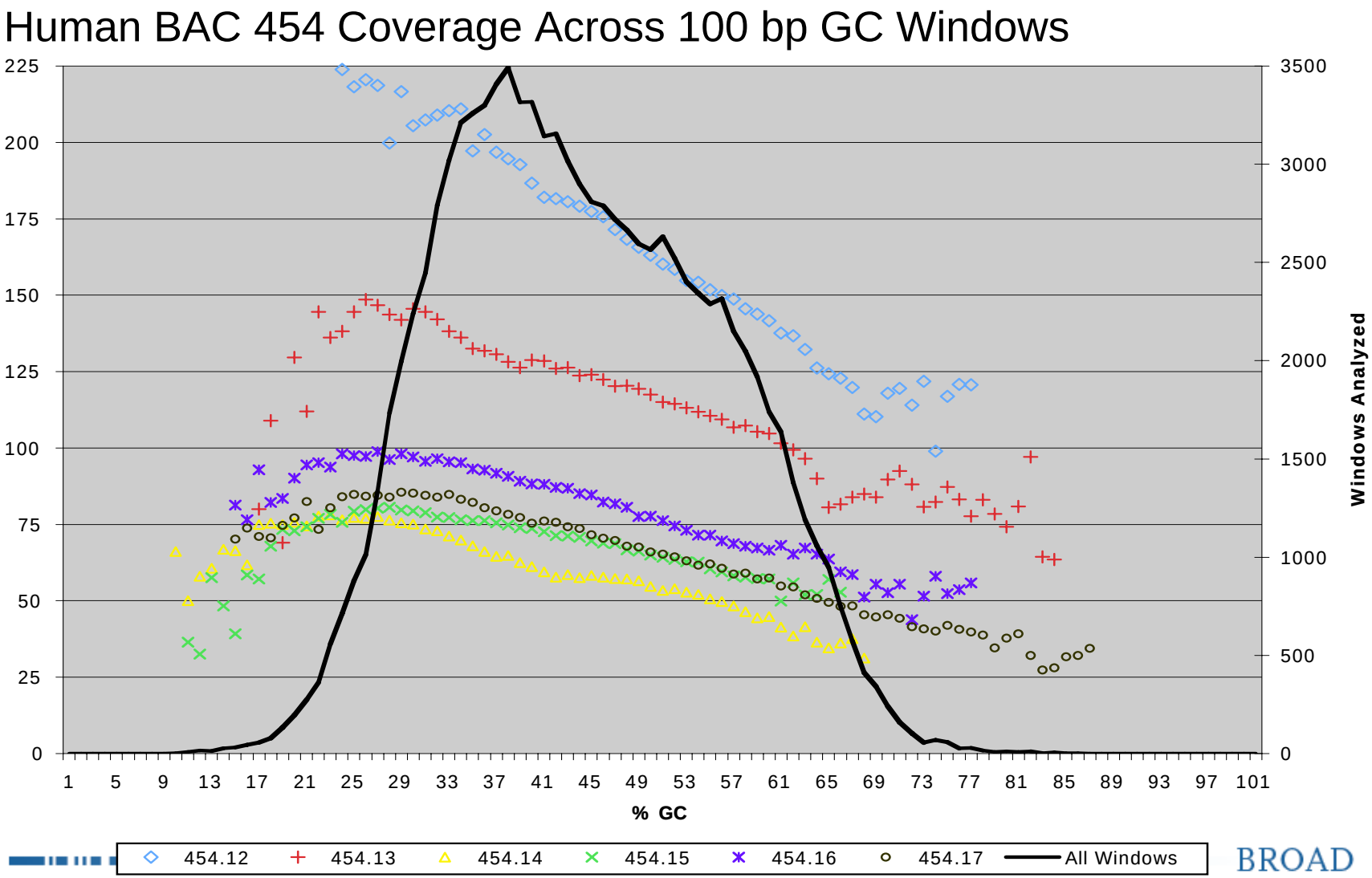
BROAD

454 Coverage: A Closer Look



BROAD

Is Coverage Drop An Isolated Case?



BROAD

Revisiting 454 Questions

- Different Bias?
  - Yes

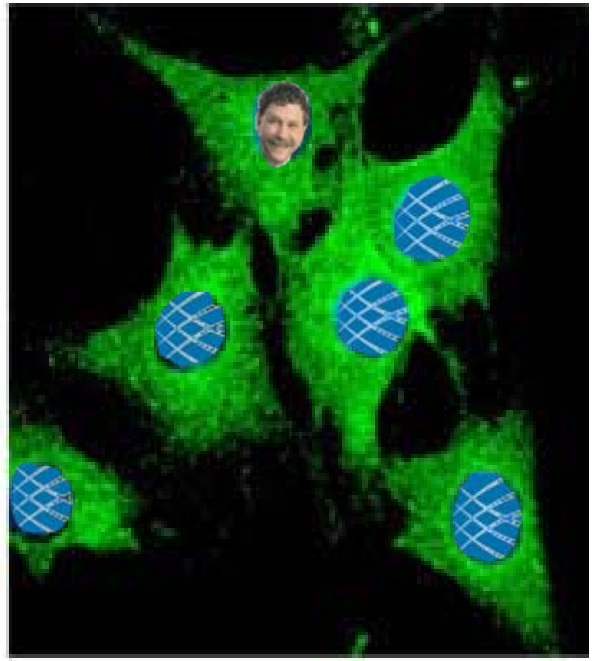
- Accuracy?

Assessing Sequence Accuracy of 454

Compare 454 assembly to finished sequence

- Base errors rare in consensus (~q40)
- >95% of errors are insertion/deletion
  - Majority in homopolymer runs
  - ~1% error rate in 5+ bp homopolymer runs

Hypothetical *Broaderia* Genome



- 3 Mb genome
- 3000 genes
- 1/10 kb Indel errors
- 300 errors/genome
  - 270 in coding region
    - frameshift in gene

9% gene models broken

Frameshifting Indels In *Listeria*

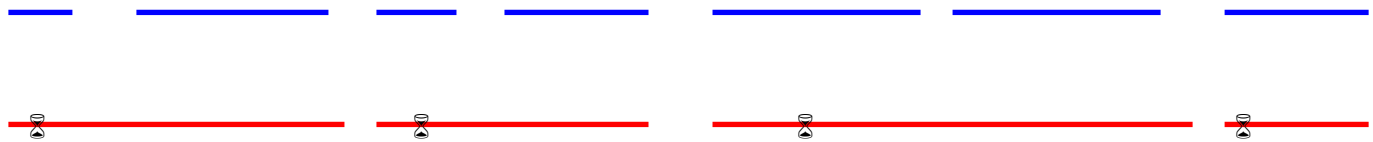
	Gene Models Broken	
	Total	Percent
10403s	47	1.57
J0161	130	4.33
J2818	238	7.93

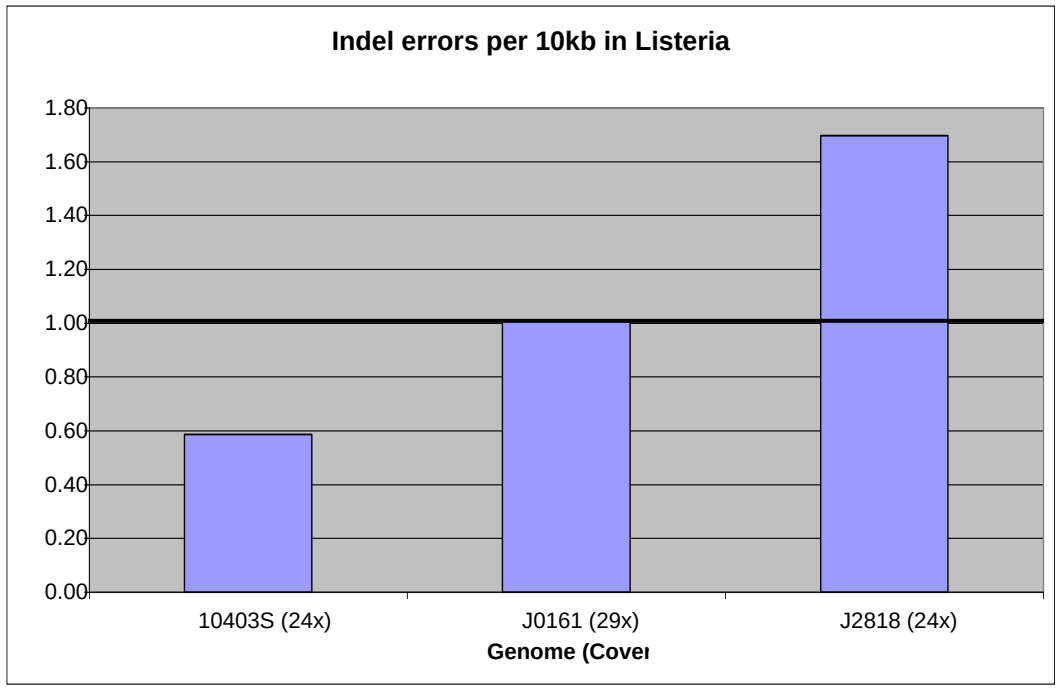
- Collaborators less than delighted
- Validation showed only 8/41 correct

Error Rates In *Listeria* 454 Assemblies

ABI q40

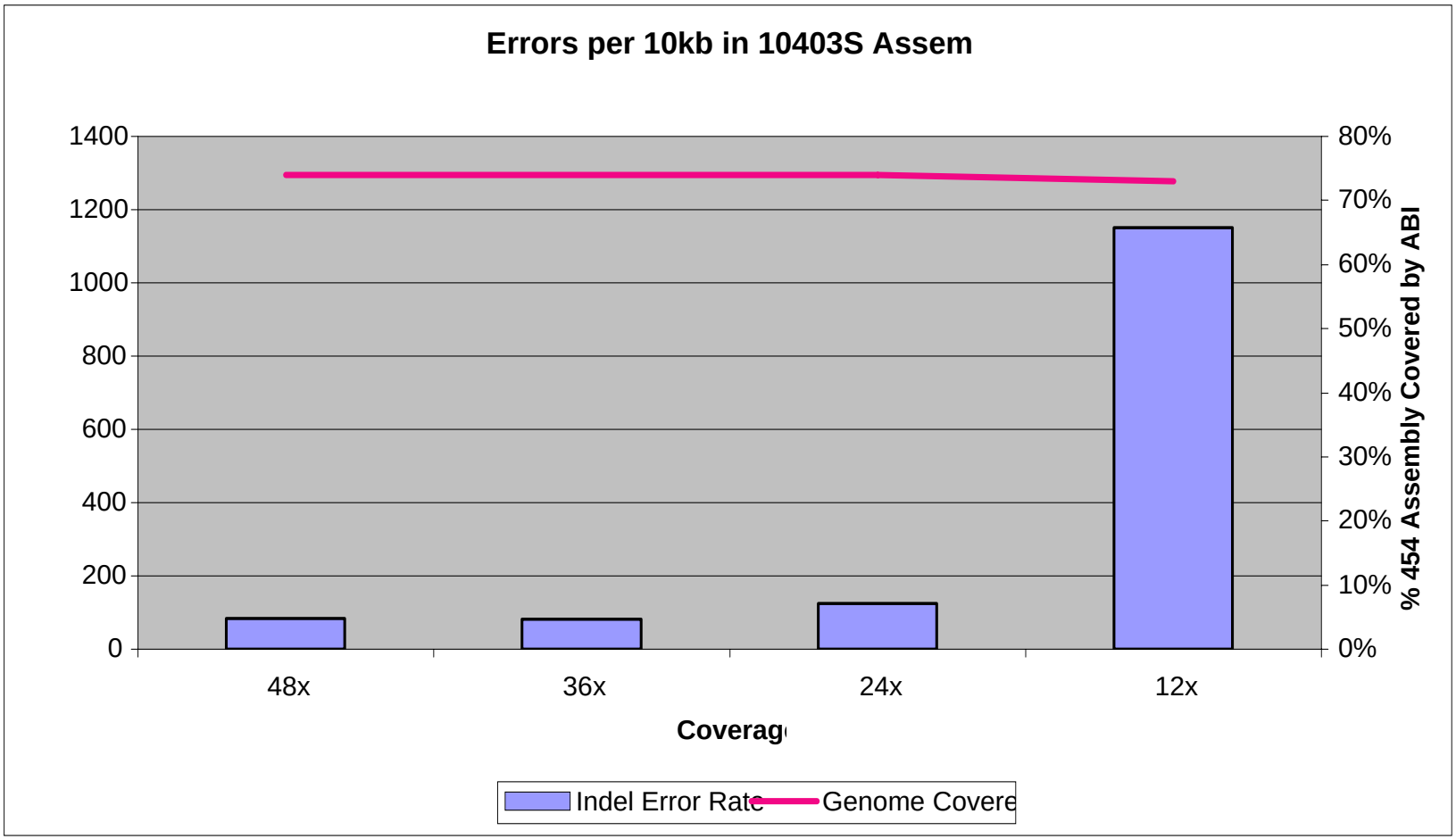
454 Assembly





9% Gene Models Broken

Increased 454 Coverage Increase Accuracy?

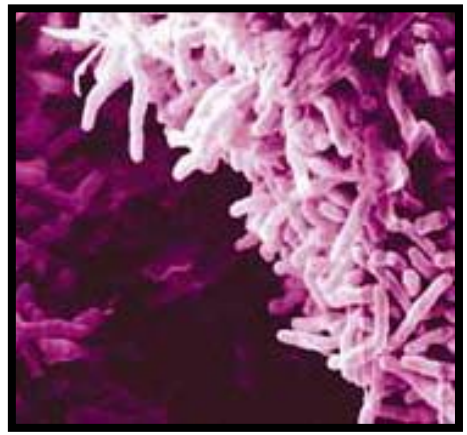


Revisiting 454 Questions

- Different Bias?
  - Yes
- Accuracy?
  - Indels break gene models

Solexa/Illumina Improve 454 Accuracy?

Solexa Accuracy



Sequence Tb by Solexa
Compare to finished sequence

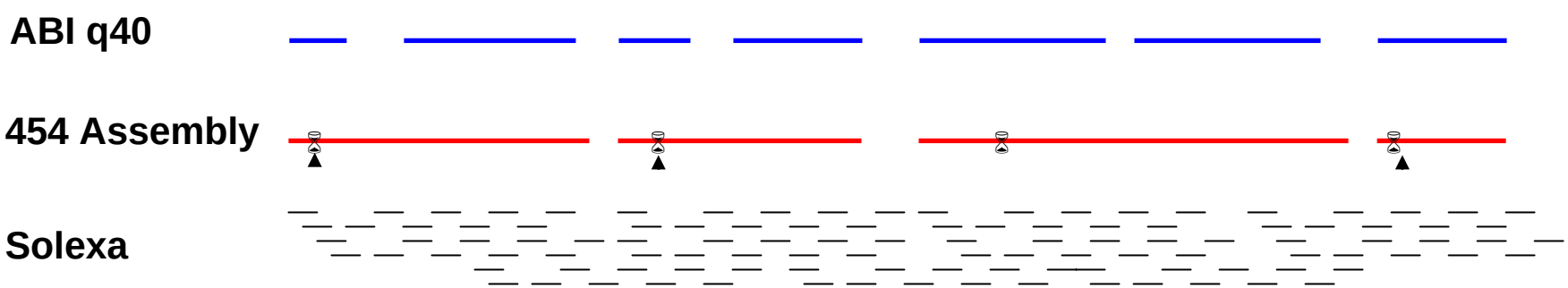
Sequenced 1 lane of F11 by Solexa
2.5M passing reads, 27b = ~14x coverage

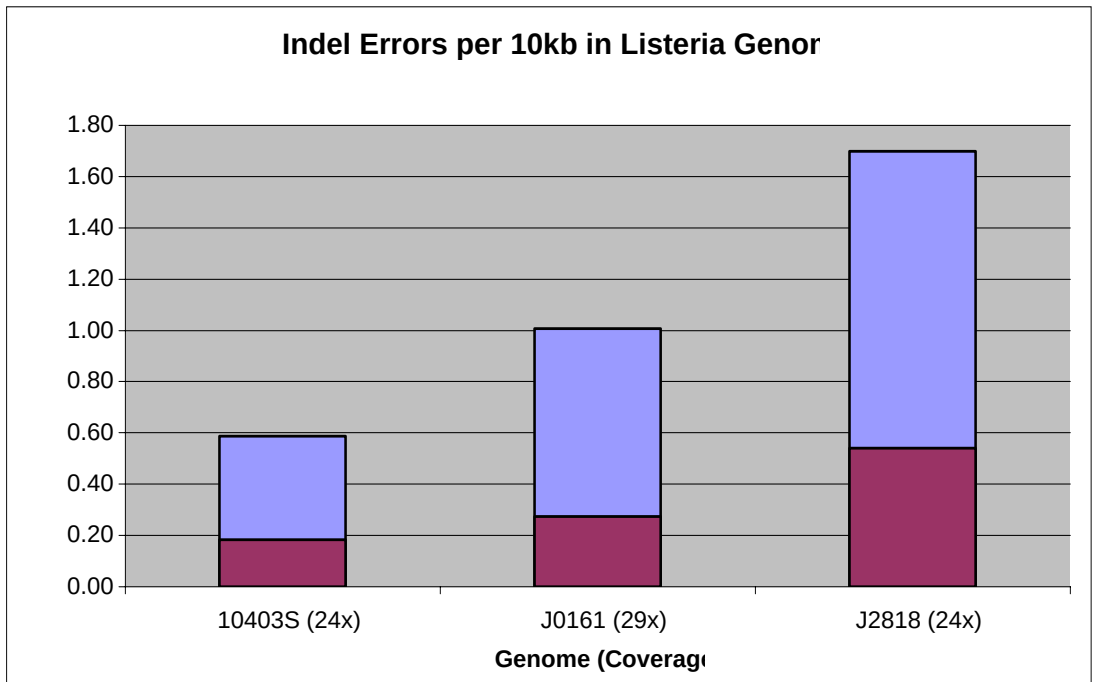
Aligned all reads to H37Rv genome

Identified 98% of 800 known SNPs
Identified 95% of indels
NO false positives
Missing SNPs in regions >80%GC

High sensitivity and specificity

Can Solexa Fix 454 Errors In *Listeria*?





Conclusions

- 454 captures regions of ABI cloning bias, but other biases still exist
- 454 indels impact gene annotation
- Solexa data correct indel errors in 454 data
- Significant improvement in consensus quality for relatively small cost

Acknowledgements

Chad Nusbaum
Bruce Birren

Assembly Analysis Team
- **Sarah Young**
- Aaron Berlin

WGA Team
- Will Brockman
- Pablo Alvarez
- **David Jaffe**

Genome Annotation
- **Chandri Yandava**
- Chinnapa Kodira

Technology Development
- **Georgia Giannoukos**
- **Will Lee**
- Sheila Fisher
- Jim Meldrim
- Nabil Hafez
- Diana Shih
- Tracey Honan
- Todd Sparrow
- Michael Weiand
- Zac Zwirko
- Nicole Stange-Thomann
- Alex Melnikov
- Andreas Gnirke
- Pat Cahill
- Robert Nicol

Computer Finishing
- **Mike FitzGerald**
- Lynne Aftuck

Scientific Project Mgt
- **Carsten Russ**