Prescreening Illumina Data Results in High-Quality Genome Polishing

Cliff Han

Joint Genome Institute -
Los Alamos National Laboratory

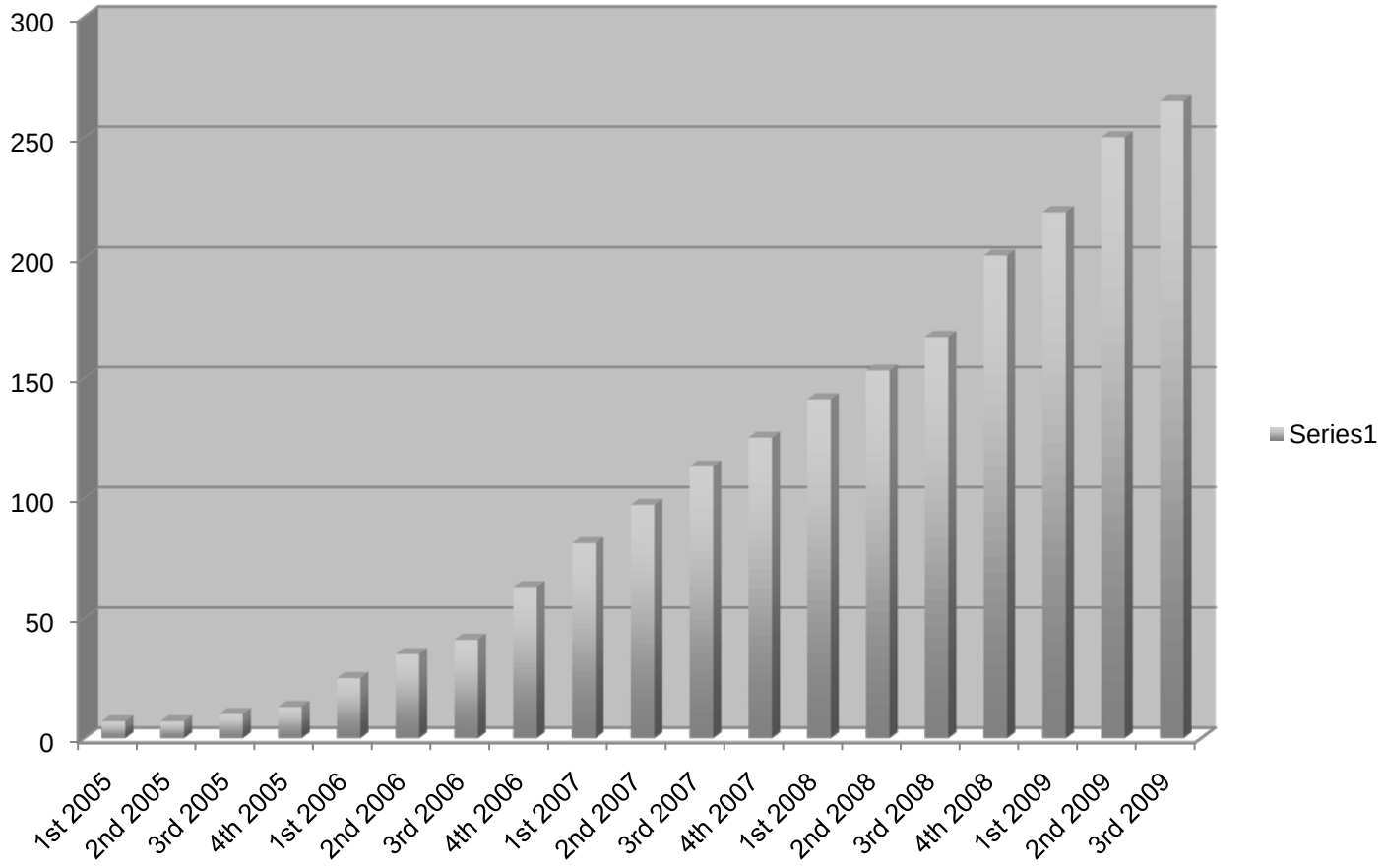
Overview

- **LANL finishing progress**
- **Prescreening of Illumina data**

Finishing microbes at LANL

- **Source:**
  - Community Sequencing Program (CSC)
  - Genomic Encyclopedia of Bacteria and Archaea (GEBA)
  - Work for others (WFO)
- **Genomes in finishing:**
  - 265 bacteria & archaea
  - 3 fungi
  - 18 phages
  - ~100 flu virus

Cumulative finished genomes at LANL



Projects types

Draft Product	Num Projects	Ave Genome Size (Mb)	Ave Contig count	Ave gap size (bases)
Sanger	285	4.1	81.2	1500
Sanger + 454 + Solexa	79	4.2	82	787
454 + Solexa	57	3.1	87.8	2338*

* From 454 assembly scaffold info, including repeats gap

Sanger and non Sanger Assemblies

Sanger	Non Sanger	Solution
20 contigs/Mb	25% more contigs per megabases	gapResolution to resolve repeats
Hundreds low quality regions /Mb	Less low quality bases	< 100 bubble PCR needed
5-10 scaffold	< 5, many 1 or 2	454 assembly structure is being kept in phrap assembly
Few high quality discrepancies	More -- likely from PCR errors	Ignore SNPs
8 – 10 x	Need higher paired end coverage to resolve repeats computationally	~10% New tech projects requested more paired end data.

Current finishing process

DRAFT MICROBIAL SEQUENCE
454(20-30 x including paired ends) & Illumina (2 lanes)

↓

gapResolution, polisher

↓

Convert Newbler assembly to Phrap assembly

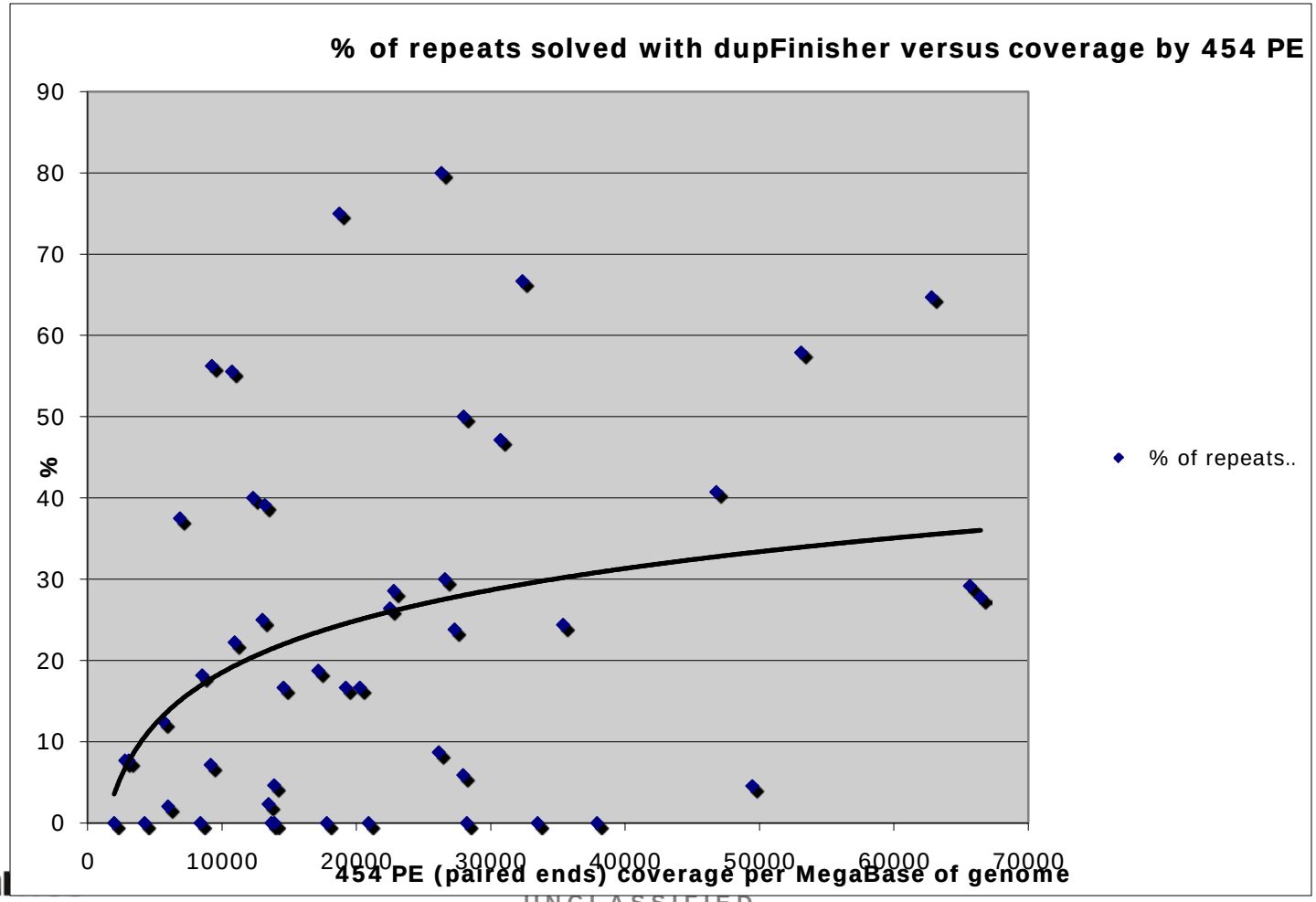
↓

dupFinisher, autofinish

Computational finishing

- Convert 454 assembly into phrap assembly
  - 454 fakes, 454 pairs, Illumina assembly,
- Using gapResolution to close duplication gaps (~30% without reactions)
- Using polisher to raise quality (<80% without reactions)
- Using consed autofinish to close unique gaps
- Automated project loading process is being developed

Repeats resolution and paired end data



Wet lab finishing

- Bubble PCR primer walk works well with non-high GC projects to close unique gaps and raise quality.
- Using PCR primer walk or subclone library to resolve repeats
- Bubble PCR need to be modified for high GC projects. PCR can be tried as well, but cost is likely high.

Finished new technology projects

Name	ProjectID	%GC	Size	# of finishing reads	# repeats solved
Aminobacterium colombiense DSM 12261	4085722	47	2	113	6
Methanocaldococcus fervens AG86	4085029	33	1.5	86	10
Thermosphaera aggregans DSM 11486	4086261	47	1.3	0	0
Hirschia baltica	4085752	41	3.5	5	2
Archaeoglobus profundus DSM 5631	4085236	41	1.6	12	4
Average for 250 finished projects	n/a	52	4.1	1374	55

Overview

- LANL finishing process
- **Prescreening of GAii data**

Polisher

- **Why?**
- **How?**
- **Performance**
  - A dozen corrections per Mb
  - Mostly indels
  - False correction < 5%, mostly substitutions

Polisher – early version

- **Performance**
  - 20 - 30 corrections per Mb
  - Mostly substitutions
  - False corrections > 50%, mostly substitutions

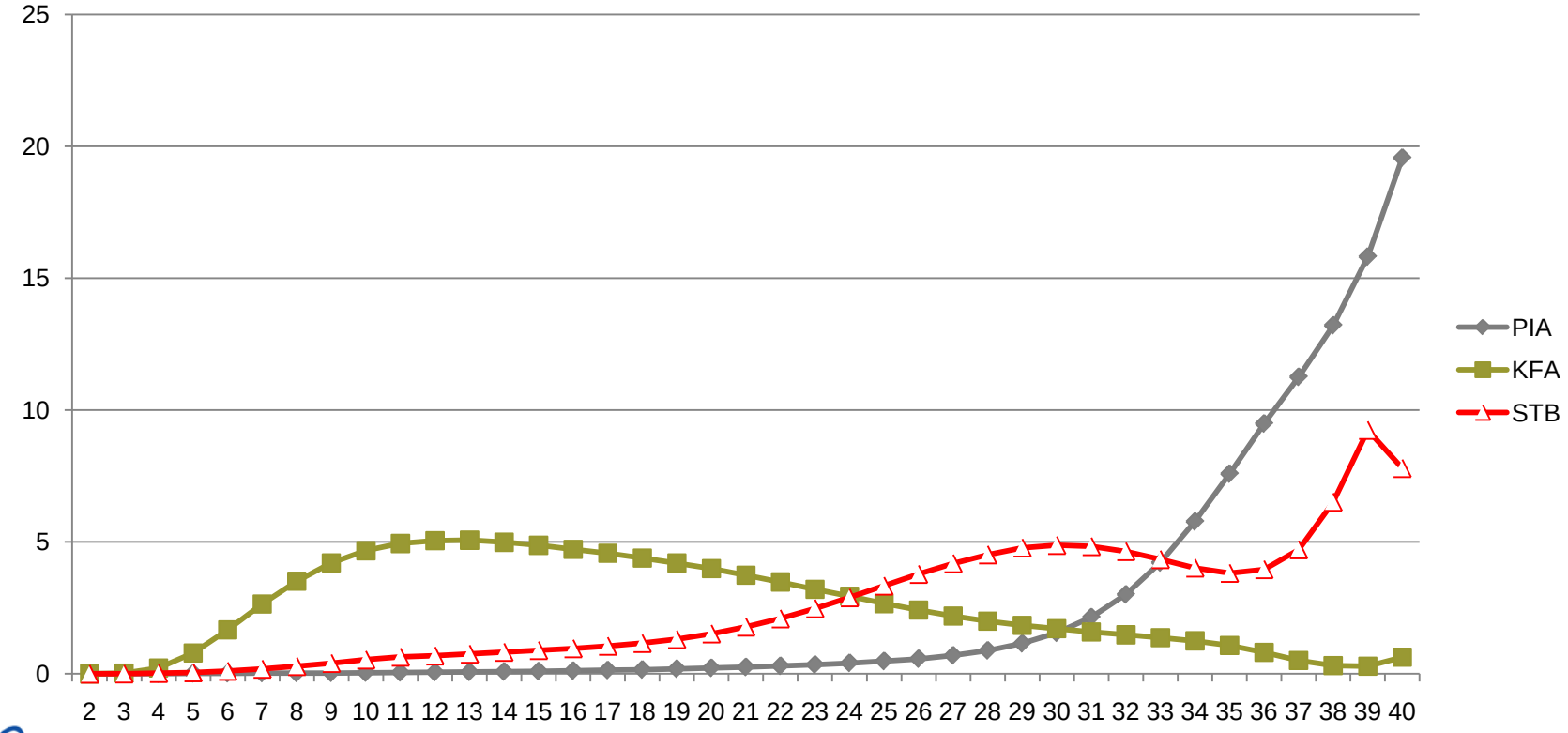
Possible reasons

- **Low quality reads**
- **Sequence artifacts**
  - Homopolymer reads
  - Stacking reads

The genomes

Genome	Size	GC%	Sanger	454	GAii (lane)
PIA	5.4	50	4x	20x	2
STB	4.5	34	4x	19x	2
KFA	7.6	70	4x	21x	3

The data from GAii



Polisher – parameters

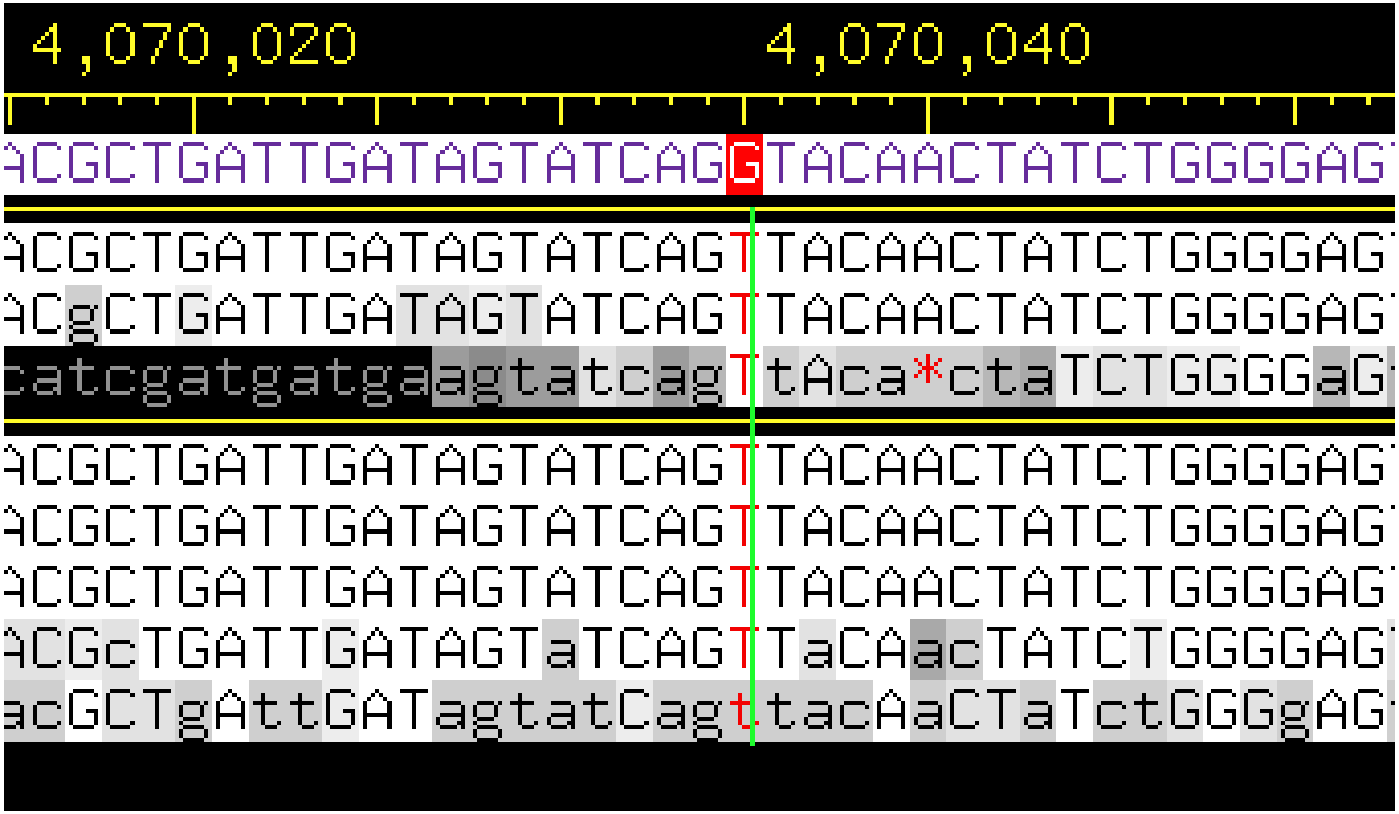
- errorDetector.coverageThreshold=5
- errorDetector.disagreeThreshold=0.5
- MAX_ERROR_PERCENT=8

Wrong corrections due to low quality reads

	0		20		30		35	
	Errors*	Coverage	Errors	coverage	Errors	coverage	Errors	coverage
PIA	3	31x	3	31x	2	29x	13	24x
STB	37	80x	42	72x	49	47x	221	8x
KFA	463	82x	426	28x	57	8x	1	2x

* Per 5 Mb genome

Polishing error confirmation



Wrong corrections due to low quality reads

	0		20		30		35	
	Errors*	Coverage	Errors	coverage	Errors	coverage	Errors	coverage
PIA	3	31x	3	31x	2	29x	13	24x
STB	37	80x	42	72x	49	47x	221	8x
KFA	463	82x	426	28x	57	8x	1	2x

* Per 5 Mb genome

Wrong corrections due to low quality reads

Average quality	Errors*
36	3
30	37
18	463

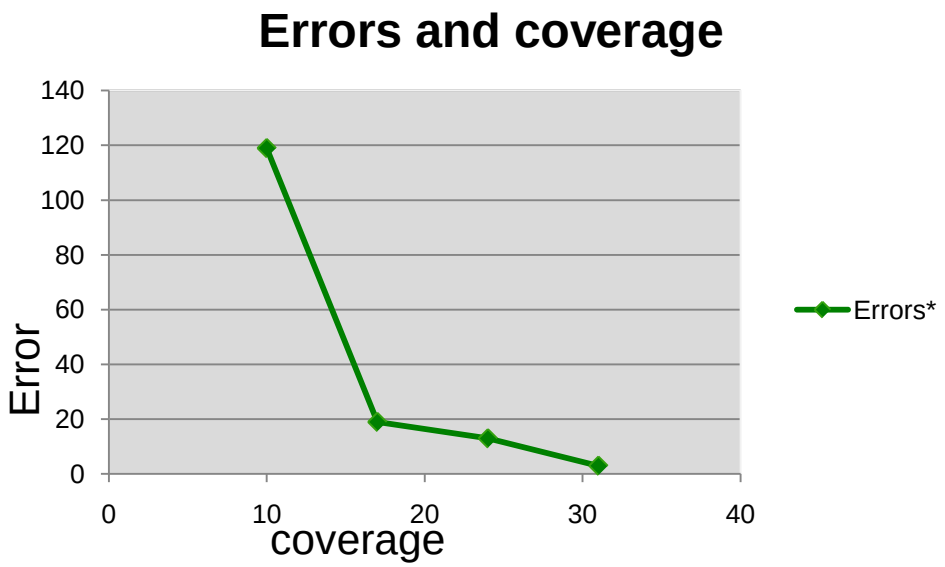
Wrong corrections due to low quality reads

	0		20		30		35	
	Errors*	Coverage	Errors	coverage	Errors	coverage	Errors	coverage
PIA	3	31x	3	31x	2	29x	13	24x
STB	37	80x	42	72x	49	47x	221	8x
KFA	463	82x	426	28x	57	8x	1	2x

* Per 5 Mb genome

Wrong corrections and coverage

Coverage	Errors*
10	119
17	19
24	13
31	3



* Per 5 Mb genome

Wrong corrections due to low quality reads

	0		20		30		35	
	Errors*	Coverage	Errors	coverage	Errors	coverage	Errors	coverage
PIA	3	31x	3	31x	2	29x	13	24x
STB	37	80x	42	72x	49	47x	**221**	**8x**
KFA	463	82x	426	28x	**57**	**8x**	1	2x

* Per 5 Mb genome

Reasons of false correction

- **Low quality reads**
- **Low coverage**
- **GC content?**
- **Sequence artifacts**
  - Homopolymer reads
  - Stacking reads

Current data screening

- Enough coverage
- Low quality screening
- Sequence artifact removal
- Stacking reads removal

Summary

- **LANL finishing process**
- **Low quality reads from GAii could cause finishing errors.**
- **Minimal 30 x of Gaii data needed for an error of < 1 per megabase**
- **Using non-amplified sequencing library will results low error rate in low coverage regions**

Acknowledgements

- **JGI – LANL Teams**
  - Tom Brettin
  - David Bruce
  - Lance Green
  - Cliff Han, Olga Chertkov, Karen Davenport, Hajni Kiss, Linda Meincke, Chris Munk, Liz Saunders
  - Chris Detter
- JGI-PGF
  - Alla Lapidus
  - Stephan Trong
  - Kurt Labutti
  - Brian Foster
  - Susan Lucas

Sponsors

- DOE-OBER
- DOD-DTRA