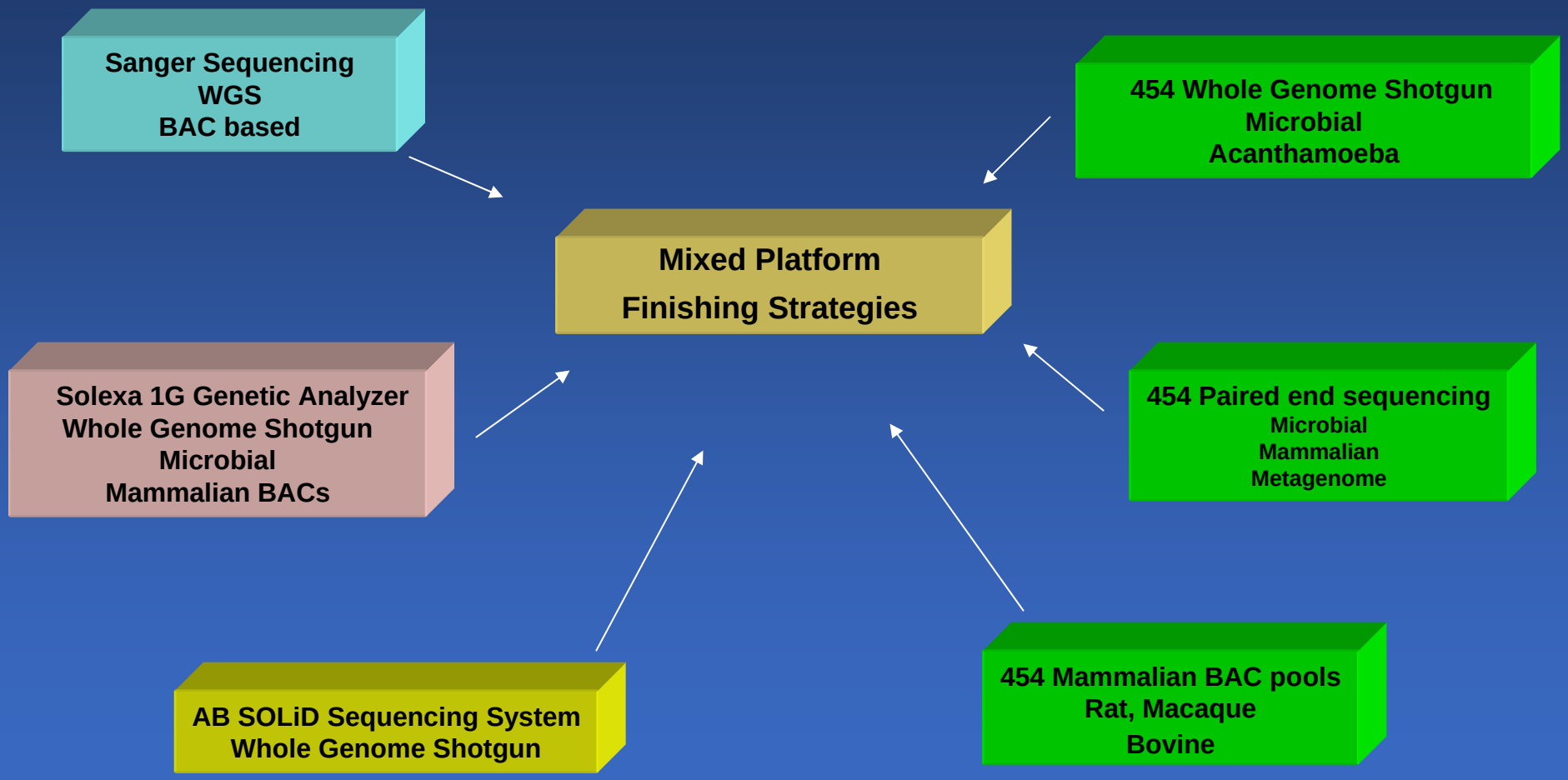
Transforming Genomes with New Sequencing Technology

Santa Fe, NM 06/19/2007

Donna M. Muzny

New Technologies and Finishing



Challenges to incorporate next Generation technologies into Finishing Strategies

- Shorter reads than Sanger
- Many platforms
- Translate error rates
- No cloned templates
- Standards for finishing

Develop Microbial and Mammalian Finishing Pipelines

Lots of work to do

BCM New Sequencing Technologies

454 Sequencing Pipeline

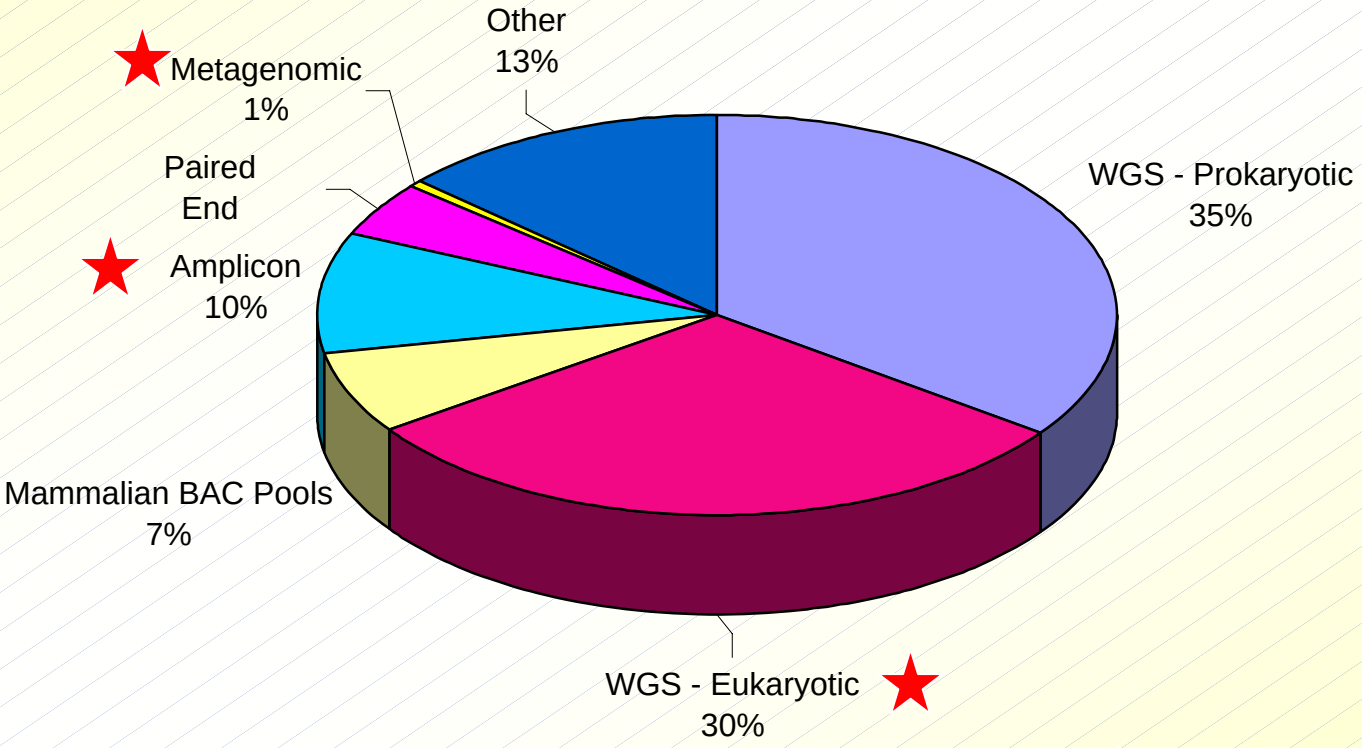
- Instrumentation
  - GS-20 - July 2005
  - 2 GS-FLX - Dec 2006
- Total of 139 - GS-20 runs for 4.6 GB of Sequence
- Total of 108 - GS-FLX runs for 9.1 GB of Sequence
- Pipeline and LIMS development using microbial projects

Solexa Sequencing

- 1G Genome Analyzer - Jan 2007
- Total of 13 runs - 12.5 GB of Sequence
- Generated sequence on 10 microbial Genomes from finishing and 454 pipeline.

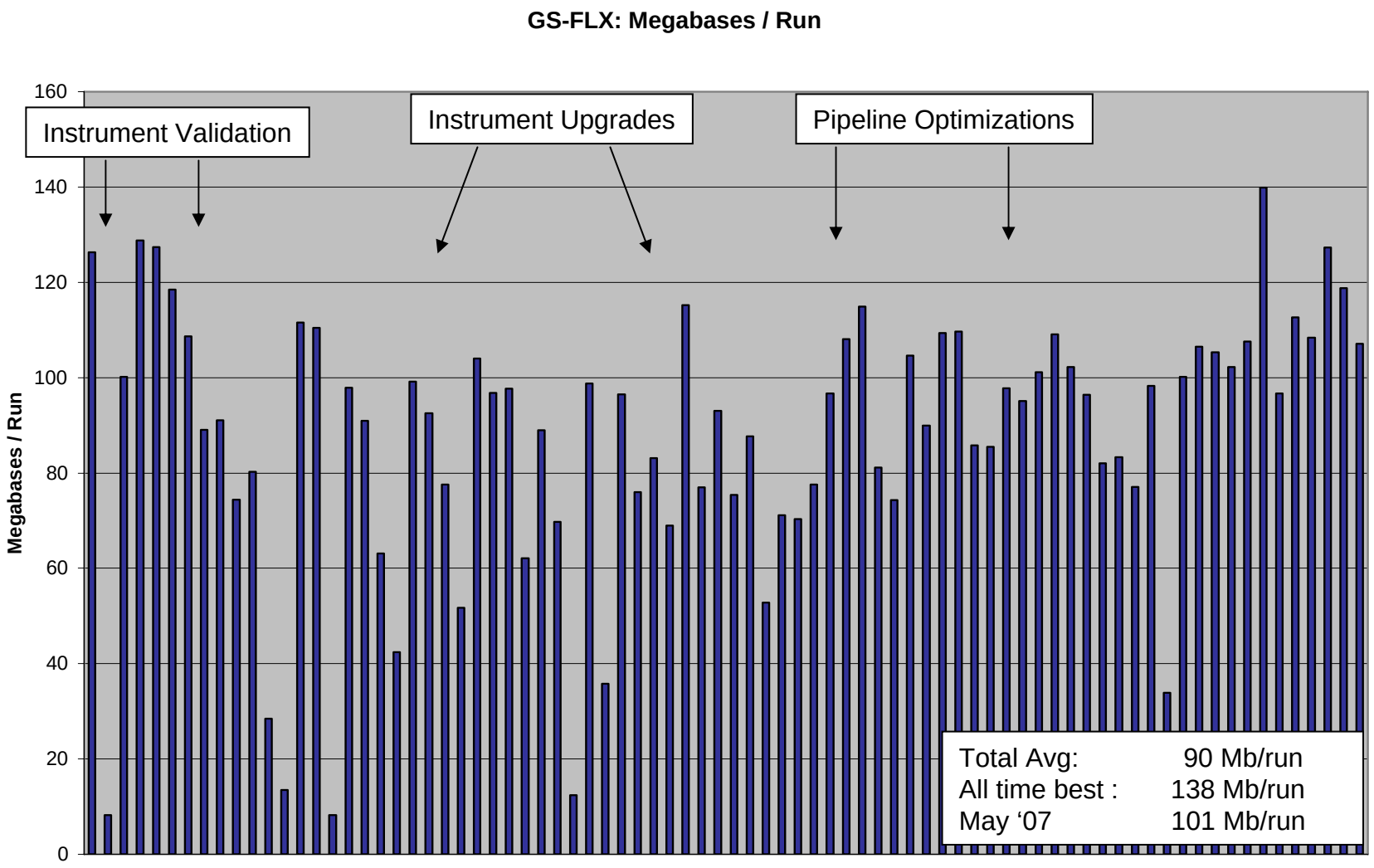
Distribution of 454 Production Bases

247 total runs
14 Gb of Sequence



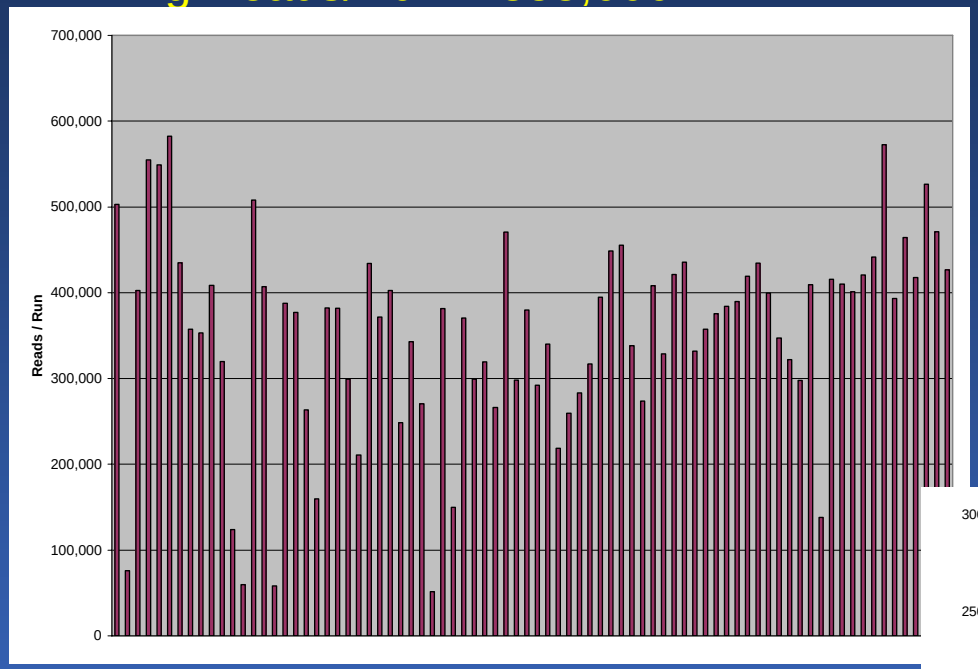
★ Areas to be expanded

GS-FLX Sequence Runs



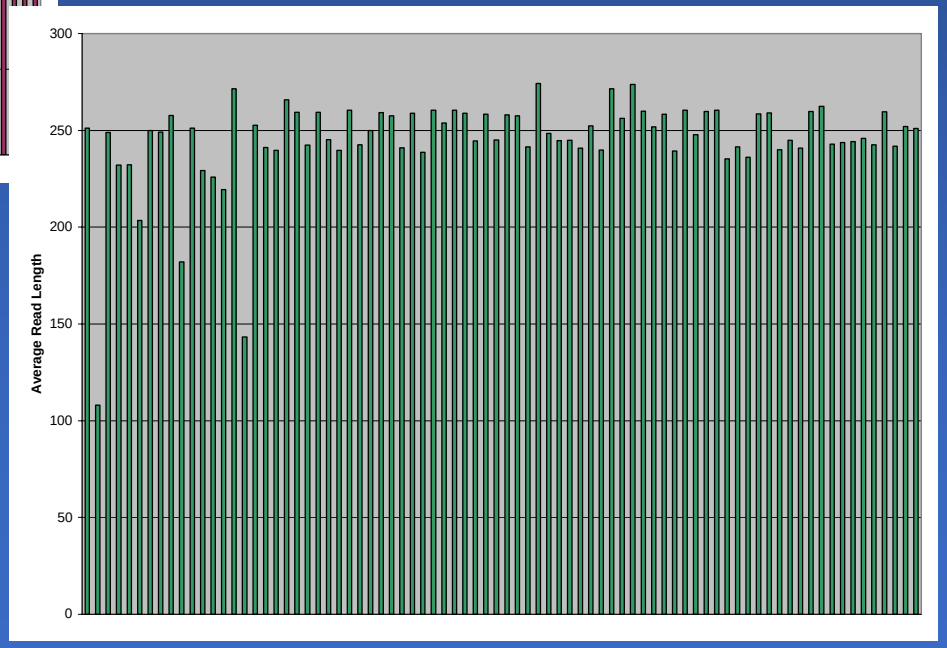
GS-FLX Run Statistics

Avg Reads/Run = 355,000



Run Metrics:
Avg Key Pass = 95%
Pass filter = 42%
TSQ% = 34%
Mixed%=19%

Average Read length/Run = 247 bp



GS-FLX Protocol Optimizations

- Instrument performance - Early Problematic Runs
  - Replaced priming pump
  - Filtered Sipper tubes
- Low bead recovery/Inconsistent recovery
  - Returned to 100ul emPCR protocols
  - Increased copies per bead to 3-4X
- High % Trimmed to Short Quality
  - Stringent software quality filter setting - “Valley Filter”
  - Loading 1.4-1.5 Million beads
  - Possible micro emPCR breakage

Microbial 454 Assemblies and Finishing

- 454 Microbial Assemblies (20-25X) behave similar to Sanger (8-10X)
- 454 Paired end and longer read lengths from the GS-FLX show significant increase in contig length and scaffolds
- Generally low error rates for draft assemblies but can vary by species
- Changes to finishing pipeline

454 Assembly Statistics for Bacterial Genomes

454 Paired-end Reads GS-20 VS FLX

454 Assembly	*F. tularensis*	*E. faecalis*		*S. aureus*	*S. aureus*
	KO97	OG1RF		USA-300	USA-300
Instrument	GS-20	GS-20		GS-20	GS-FLX
Genome Size	1.8 Mb	2.8 Mb		2.9 Mb	2.9 Mb
Total reads	443,124	560,817		431,656	353,192
Coverage	25x	21.5X		16X	32X
Average read length	105	106		105	257
Avg.454 Quality	26.4	26.5		26.6	26
Total trimmed contigs	126	128		266	94
Avg. contig length	14.0 Kb	21.1 Kb		10.6 Kb	30.5 Kb
N50 size	23.7 Kb	49.6 Kb		29.0 Kb	107.7 Kb
Longest contig	77,480	181,200		80,331	490,089
Paired-End Reads	212,571	225,208			
Total Scaffolds	20	8			
Average Scaffold Size	91 Kb	339 Kb			
Scaffold N50	116 Kb	597 Kb			

Microbial Finishing of 454 Assemblies

Microbial 454/Sanger Mixed Assembly Strategy

454 Reads Newbler Assembly → 454 Assembled Scaffolds/Contigs → Sanger Reads used to capture 454 Contigs by Blast → Scaffolds created: 454 paired-end Info Phred/Phrap/Consed

Finishing utilizing: Autofinish, Multiplex PCR
Direct genome sequencing, Long Range PCR

Statistics for Microbial Genome in BCM Finishing Pipeline							
	Sanger Coverage	454 Coverage	Total Reads (Sanger/454)	Total Size(Mb)	Initial Scaffolds	Current Scaffold(s)	Status
S. aureus	9.3 X	16.3 X	37925/431656	2.8	5	1	complete
E. faecalis OG1RF	4 X	20 X	19101/560817	3	26	1	complete
P. stewartii	19 X	9.6 X	144906/503412	5	99	28	in progress
B. pumilus	10 X	15.5 X	72300/572802	3.8	14	4	in progress
T. paraluiscuniculi	0	59.1 X	0 /643072	1.1	77	1	in progress
M. bovis	1.8 X	23.9 X	6491/560327	2.4	59	33	in progress
S. iniae	1.5X	28 X	4554/526264	2	16	16	in progress

Finishing Tools for 454/Sanger Microbial Assemblies

- Pipeline techniques
  - Utilization of 454 Paired end data
  - Autofinish - Adjusted primer walks and PCR design
  - Repeat databases for each Microbe
  - Tagging of repeats
  - Adapting human TPF tools for tracking microbial genome assemblies

Finishing Tools for Mixed Assemblies

- Lab techniques developed
  - Direct walks from GenomiPhi

  (60% -88% Pass, 600 Q20 read lengths, *P. stewartii*, *B. pumilus*, *S. aureus*)

  - Multiplex PCR + sequencing with individual primers

  (closed 20 gaps in *P. stewartii*)

  - Long Range PCR + Shatter libraries

  (rRNA regions)

Direct Microbial Genome Sequencing

Protocol for GenomiPhi Amplification and Direct Sequencing of Microbial Genomes

Bacterial Genomic DNA 10ng/ul (Genome size: 2-5Mb)

Sample buffer 9ul, 95°c X 3 minutes, then cool to 4°c

↓

Add mix reaction buffer 9ul, enzyme 1ul(all on ice)

Incubate at 30°c for 18 hours

↓

Inactivate enzyme at 65°c for 10 minutes,

cool to 4°c

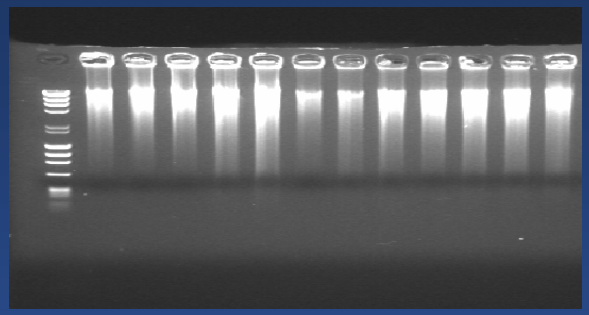
↓

EXO-SAP 2ul for 15 minutes

↓

Sequencing with Premix AB Big Dye 1:8,

DNA 1:1 dilution



See posters by
Yan Ding and Shannon Dugan

Direct Genome Walk

Mammalian Finishing
Targeted finishing in draft genomes

- A critical component for both new genome sequences and upgraded draft sequences
  - BAC-based targeting
- Resolve segmental duplications (esp. in WGS assemblies)
- Refine gene structures (target genic regions)
- Upgrade rapidly evolving regions
- Define breakpoints (comparative genomics)

Genome Refinement Resources

	BACs	Mapped BACs	Fosmids	Refseq genes	mRNAs	ESTs	Human orthologs	Mouse orthologs
RAT	20,697	16,126	None	9,580	23,627	838,663	22,035	19,066
MACAQUE	-	77,613	193,558	None	1,856	63,290	24,284	18,652
BOVINE	19,680	None	None	None	11,408	1,233,439	24,044	18,726

Genboree: Macaque Site

RefSeqs

Rhesus:ESTs

Rhesus: BES

Rhesus: FES

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

Mammalian 454/Sanger Assemblies

Mammalian BAC pools of 50-100 BACs

Reduced up front library effort

Cost effective per run

Challenges:

- What is the Optimum 454 read coverage needed for an effective top up?
- Repeats– increased complexity in mammalian genomes yields 454 contig N50 lengths of approximately ~4kb.
- Equal BAC representation in the pooling process.
- Screening for overlapping regions or segmental duplications in the pool.
- Mapping of 454 reads/contigs – can we accurately map the 454 contig back to the appropriate BAC from the pool?

Mammalian Bac Assembly Statistics: Effects of Paired-Ends and Longer Reads

Rat BAC Pool of 9 BACs	GS-20	GS-20 +	GS-FLX	GS-FLX+
454 Assembly		Paired ends		Paired ends
Total Size: 2.1Mb				
Total reads	481,739	5		
Coverage	21X			
Average read length	97			
Avg.454 Quality	24.5			
Total trimmed contigs	1193			
Avg. contig length	1.5Kb	1.		
N50 size	3.5Kb	3.5Kb	4.8Kb	3.9Kb
Longest contig	25,998	26,005	41,796	27,104
Paired-End Reads		125,081		125,081
Total Scaffolds		134		112
Average Scaffold Size		13.4Kb		16.5Kb
Scaffold N50		35Kb		54Kb

150 Bovine BAC Stats

Coverage: 17X

Read Length: 254 bp

N50 size: 4.9Kb

Longest Contig: 52,939

Mammalian BAC pools

Pooled BAC Experimental sets:

1. Macaque pool of 100 BACs
   - all finished BACs, genic regions

2. Rat pool of 100 BACs
   - finished/unfinished, genic regions

See poster by
Christian Buhay
Mike Holder

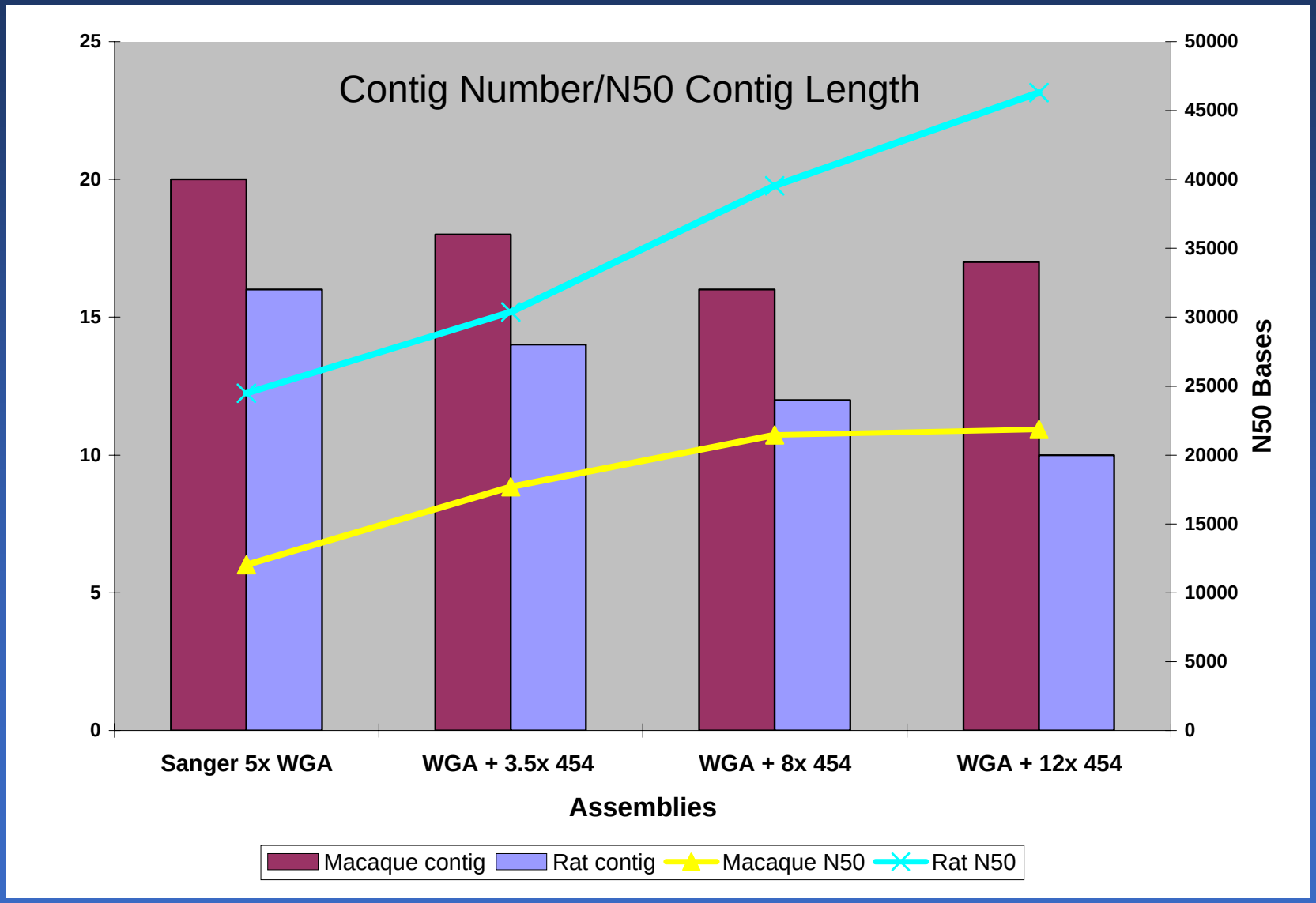
454 Reads → Newbler assembler → 454 contigs

4X Sanger BAC Assembly Reads → Blast → 454contigs + Sanger Reads per BAC → Phred/ Phrap → 454/Sanger Mixed Assm

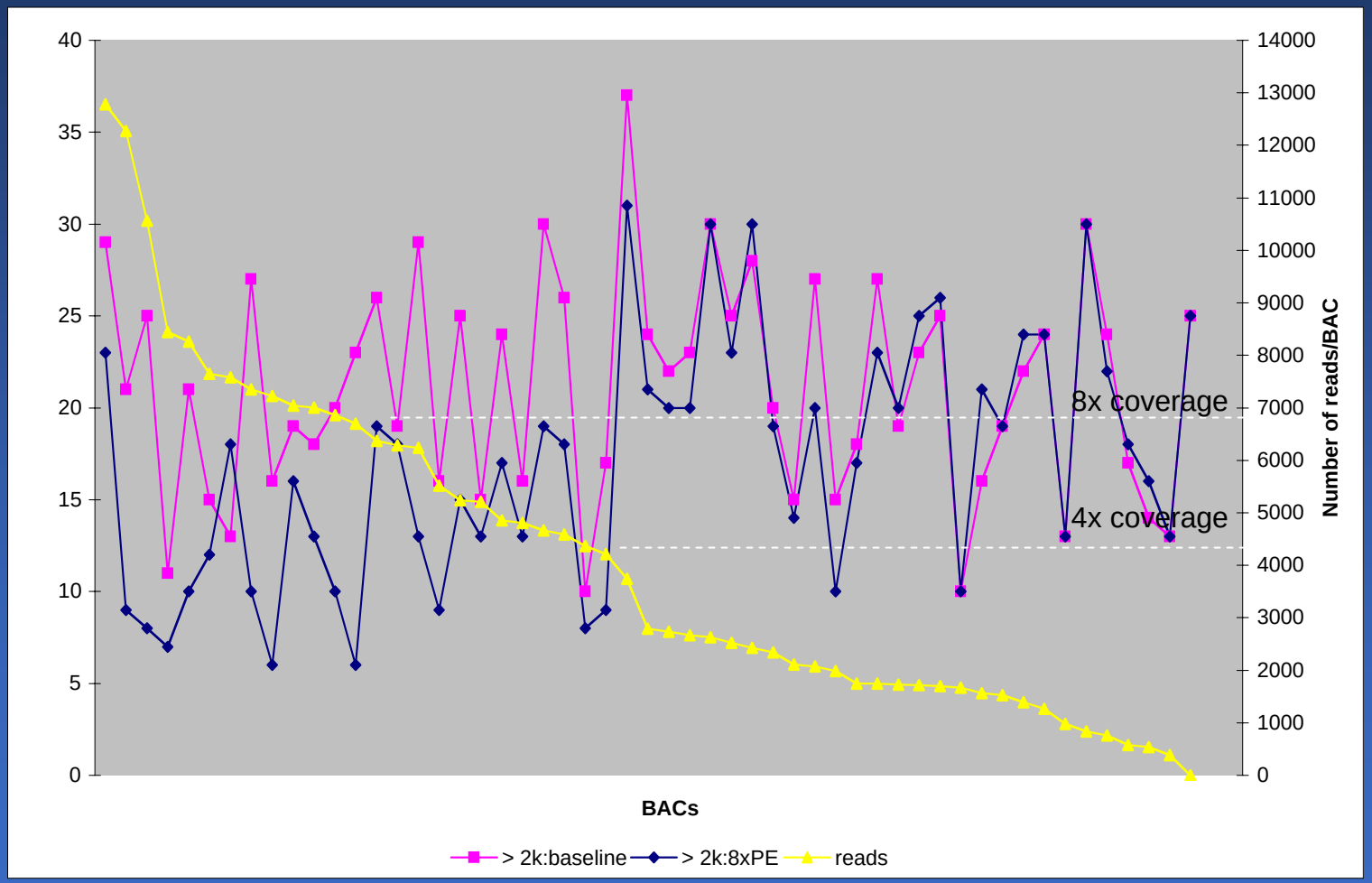
- Used 4-5X Existing Whole Genome Shotgun reads or BAC top up reads for Sanger Reads

- Added increasing coverage of 454 reads (4,8 and 12X) to determine optimal coverage

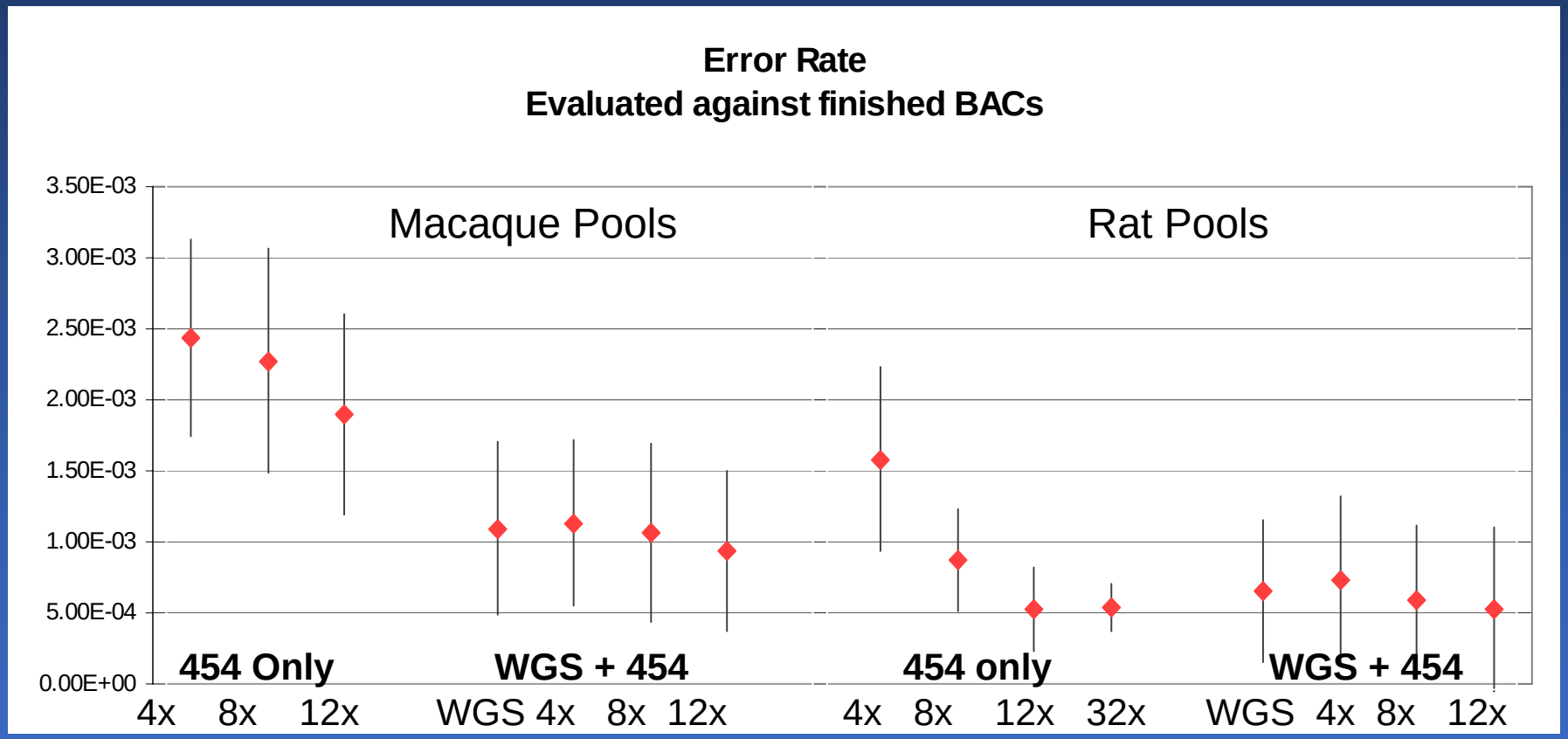
Mixed Assembly Results



100 Macaque BAC pool
4X Sanger Coverage + 8X 454 Pool Coverage
Distribution of individual BAC Coverage and Contigs



454 Mammalian Error Rate Estimates



Mammalian BAC Summary

- 5-6X 454 coverage per BAC is optimal for gap closure. Target 8X 454 pool coverage.
- Assembly performance issues between species
- Now have ~20 BACs in finishing from the Rat pool

- Balance pooling strategy with smaller pools
  - Pool in groups after library construction
  - Pool after emPCR
- Evaluate 454 mapper tool for sorting reads.

Acknowledgements

Richard Gibbs

George Weinstock

HGSC Finishing Group

Shannon Dugan-Rocha
Yan Ding
Christian Buhay
Mike Holder
Xiang Qin
Aniko Sabo
Alicia Hawes
Hua Shen
Guan Chen
Qiaoyan Wang

Microbial genomes

Joseph Petrosino
Sandor Karpathy
Sarah Highlander

HGSC 454/Solexa Group

Christie Kovar-Smith
Huyen Dinh
Sandra Lee
Pete Blyth
Bei Bei Song
Larhonda Jackson
Rica Pauzo
Anna Sneed
Viktorya Korchina

Lynne Nazareth
Yi Han
Jeffrey Reid

Assembly Analysis

Erica Sodergren
Kim Worley
Xiang Qin
Huaiyang Jiang
Yufeng Shen
Yue Liu