Assessing 454 Errors in Microbial Genomes

Stephan Trong

Bioinformatics, Microbial Genomics
DOE Joint Genome Institute

Sanger/454 Hybrid Assembly

Sanger/454 Hybrid Assembly at the JGI

- 9X Sanger (3 libraries), 20X 454/Roche GS 20 sequencing depth.
- Assemble 454 reads using Newbler assembler.
- Shred 454 Newbler contigs into 1000 bp fragments with 100 bp overlapping ends.
- Assemble 454 pseudo fragments with Sanger reads.

High Quality Finishing Standard

- All consensus bases >= Q30.
- Final error rate <= 0.2 per 10 Kb (1 error in 50,000 bp).
- No single clone coverage, i.e. minimum of 2X depth everywhere.
- Manually inspect and quantify single stranded regions.
- Check all high quality discrepancies.
- Final sequence should have a base at every position (no strings of xxxx anywhere).
- Verify all repeats (perform paired ends sequencing and PCR if necessary).
- Check correctness of final assembly. Confirm questionable areas with PCR.

454 Error Rate Analysis

GOAL:

- Assess error rate and type of errors produced by 454 sequence.

- Devise a strategy to target 454 errors in Sanger/454 hybrid assembly for quality improvement.

  - detect errors by motif?
  - correlate Sanger quality with Newbler quality?
  - adjust quality based on 454 read depth?

454 Error Rate Analysis

METHOD:

- Evaluated 9 finished microbial genomes sequenced with Sanger/454.
- Masked repeats in finished sequence.
- Aligned 454 Newbler contigs with finished sequence.
- Identify base mismatch between 454 and finished sequence.
- Categorize mismatches into homopolymers and non-homopolymers, where homopolymer is defined as a stretch of 2 or more identical bases.
- Categorize errors by read depth.

Results from Analysis

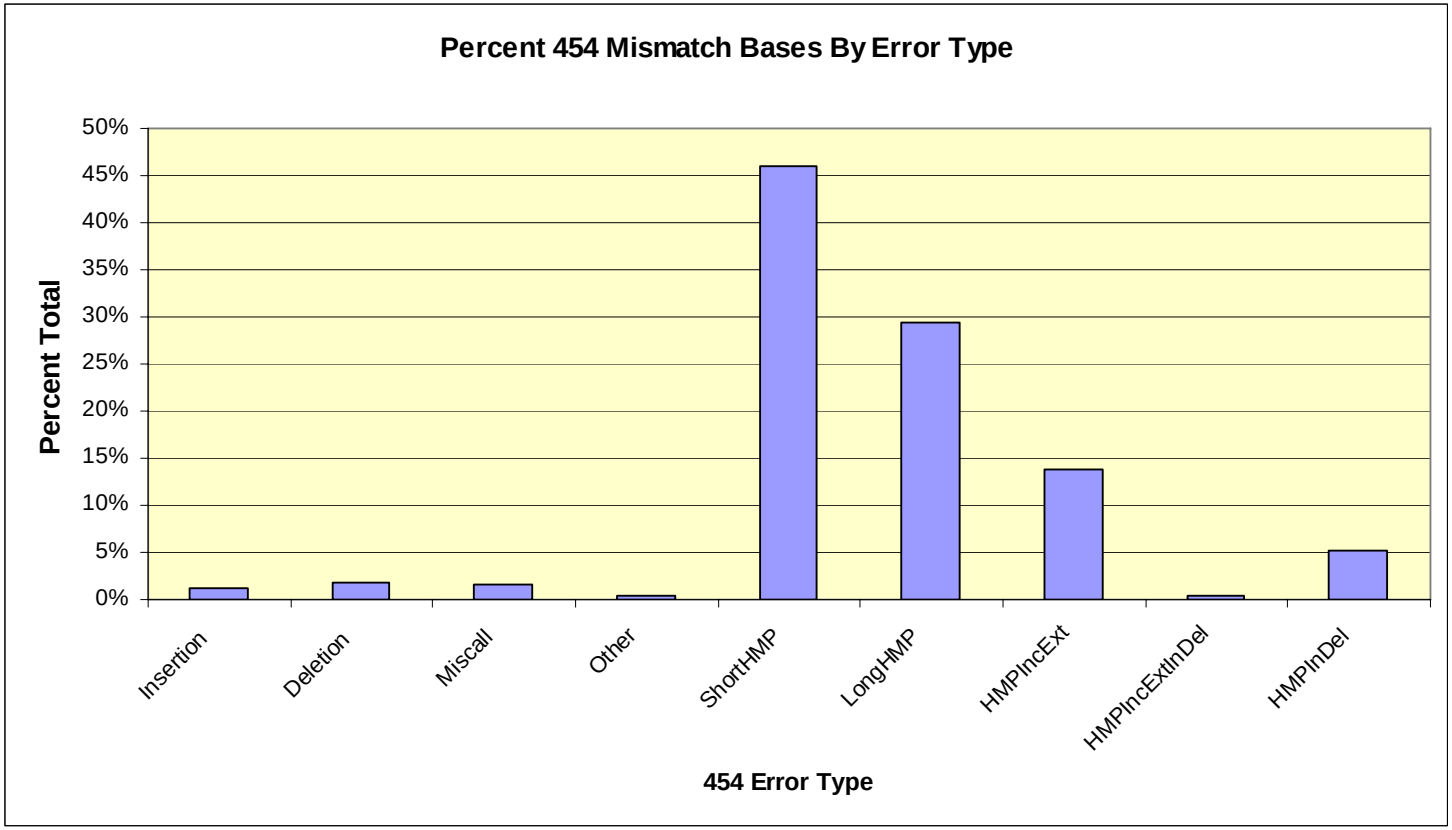
RESULTS:

Number of Finished Genomes:	9
Total 454 Bases:	23,940,257
Total Sanger Bases:	24,639,388
Total Bases Aligned:	23,650,082
Total Errors	5965
Total Error Rate	1 in 3,965
Homopolymer Errors:	5,666 (95%)

- Of all homopolymers detected in finished sequence, 1 in 833 (0.12%) contain errors

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

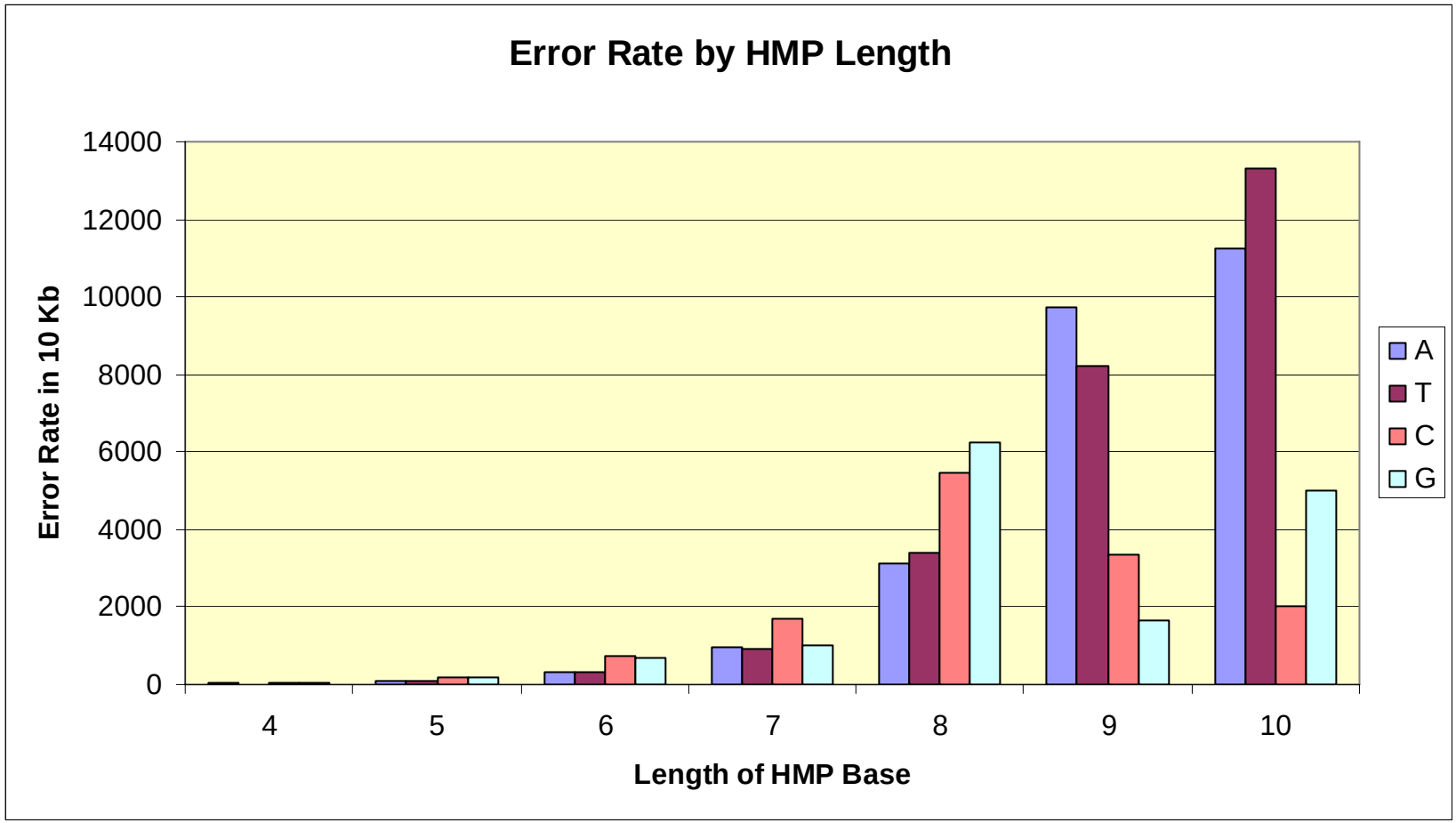
454 Mismatch Bases by Error Types



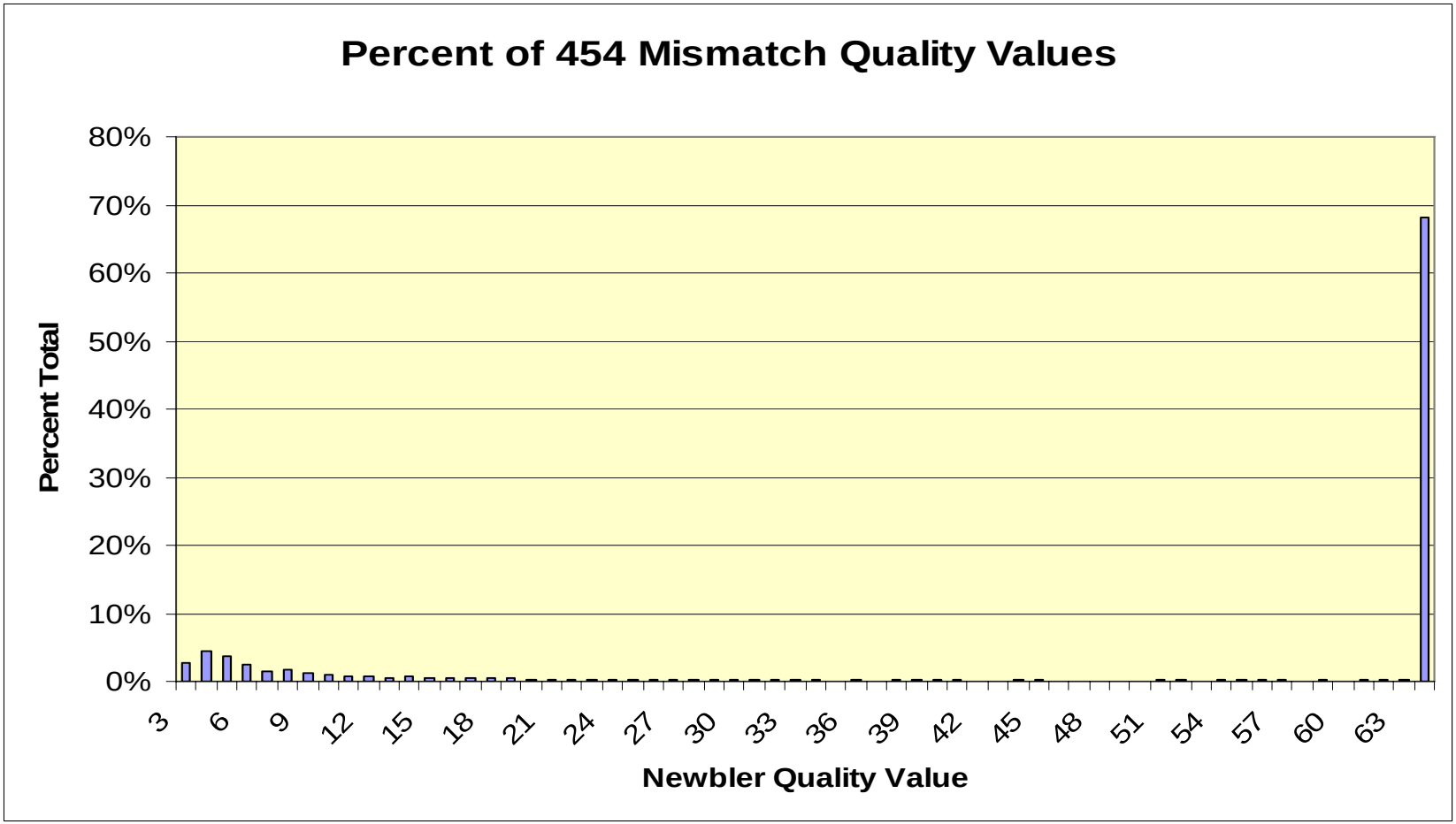
Non-homopolymer

Homopolymer

454 Error Rate by HMP Base and Length

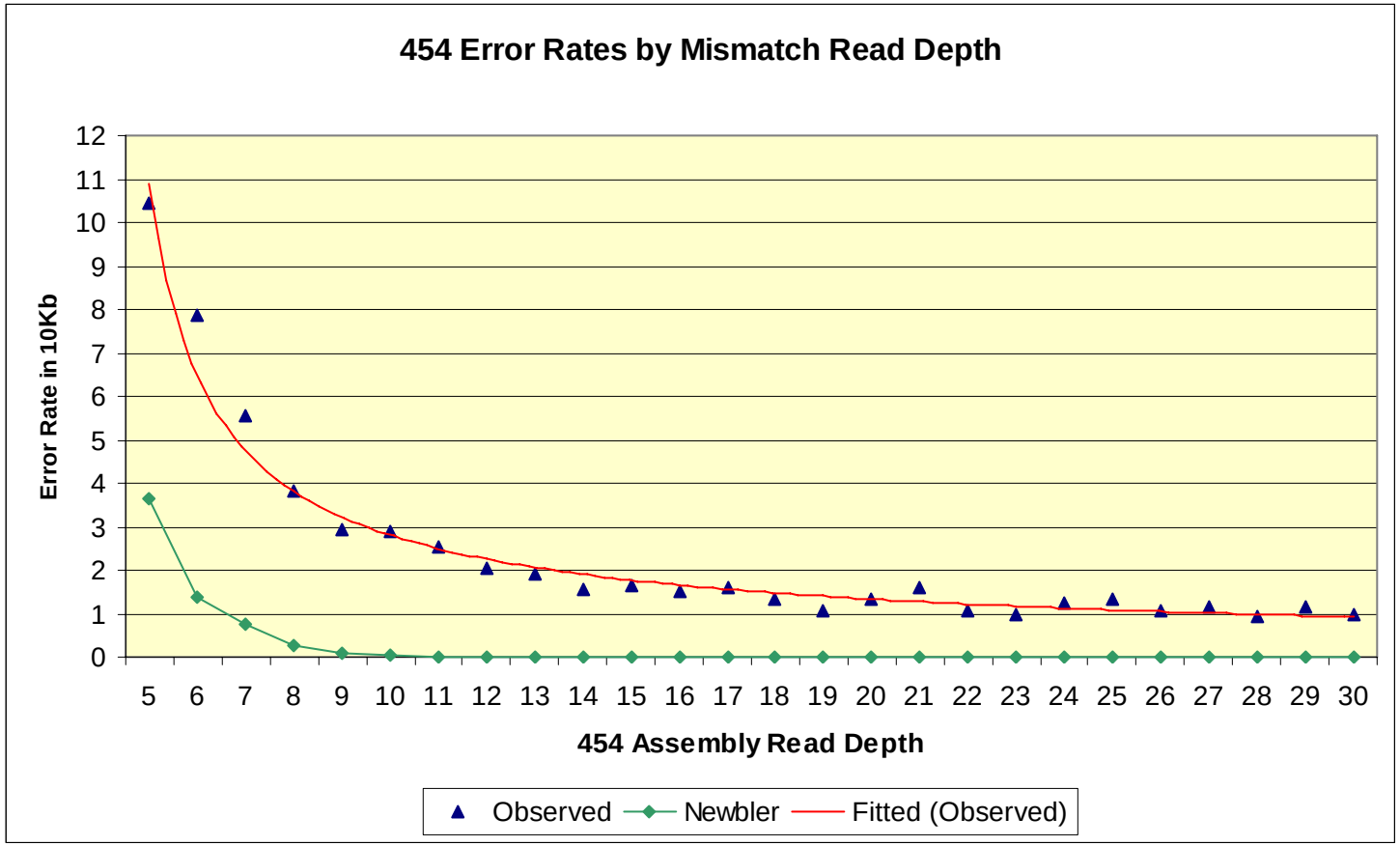


454 Mismatch Quality Values



Approximately 69% of 454 errors are assigned a Newbler consensus quality value of 64, the highest value assigned by Newbler.

Error Rate by 454 Mismatch Read Depth

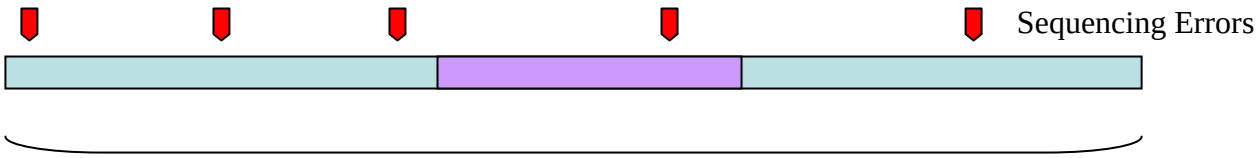


$$\text{Error Rate} = \frac{\text{Number of bases at read depth}}{\text{Number of mismatches at read depth}}$$

How Do We Use Read Depth to Target 454 Errors?

10% 454 Only Coverage

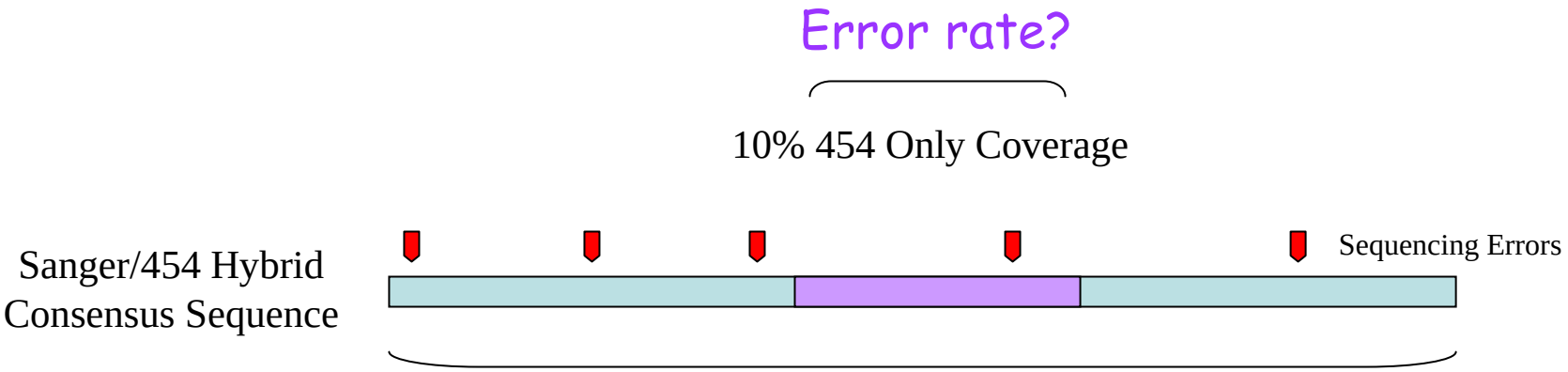
Sanger/454 Hybrid Consensus Sequence



Overall error rate <= 1 in 50 Kb

How Do We Use Read Depth to Target 454 Errors?

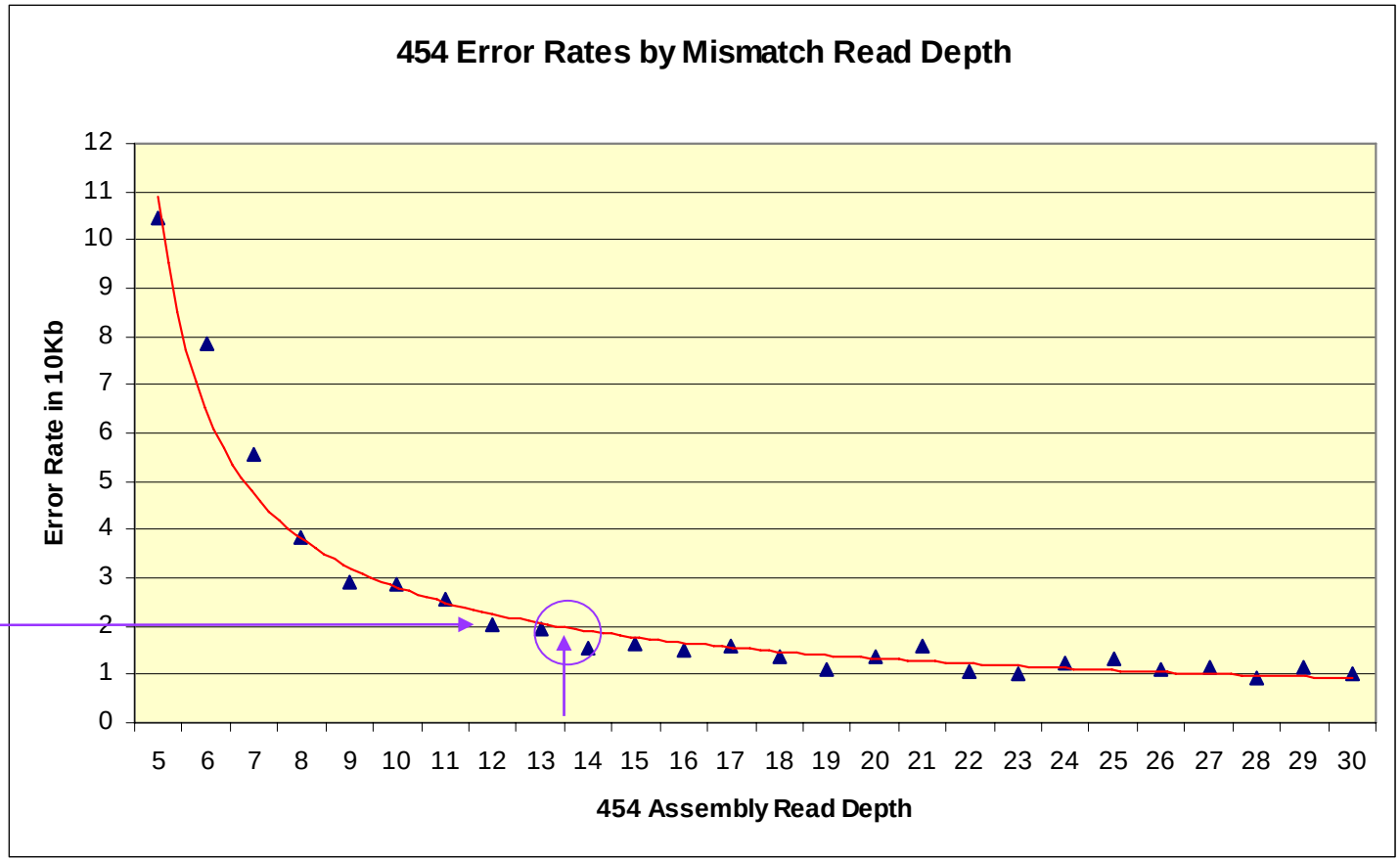
What is the error rate of the 454 only coverage such that the overall error rate is 1 in 50 Kb (assuming no errors outside 454 only coverage)?



Overall error rate <= 1 in 50 Kb

Error rate = 0.10 x 50 Kb = 5 Kb <u>OR</u> 1 in 5 Kb

Determining Minimum 454 Read Depth to Achieve Finishing Error Rate

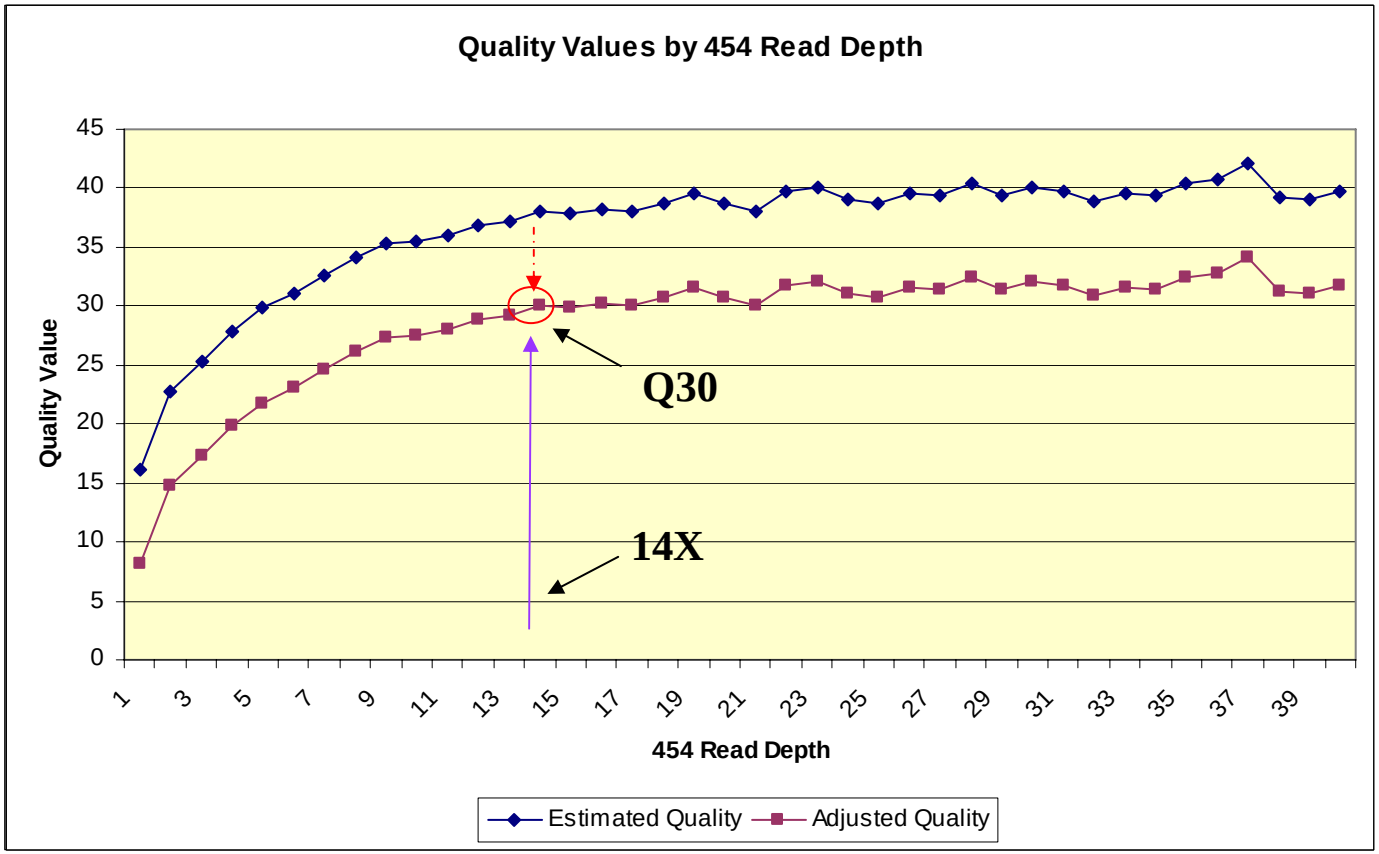


Error Rate in 5 Kb

Given Sanger/454 hybrid assembly containing 10% 454 only coverage, the minimum 454 read depth to achieve an overall error rate of 1 in 50 Kb is 14X.

Adjusted Quality Values Based on 454 Read Depth

Given Sanger/454 assembly containing 10% of areas covered only by 454 sequence



Alternative Strategies for Quality Improvement

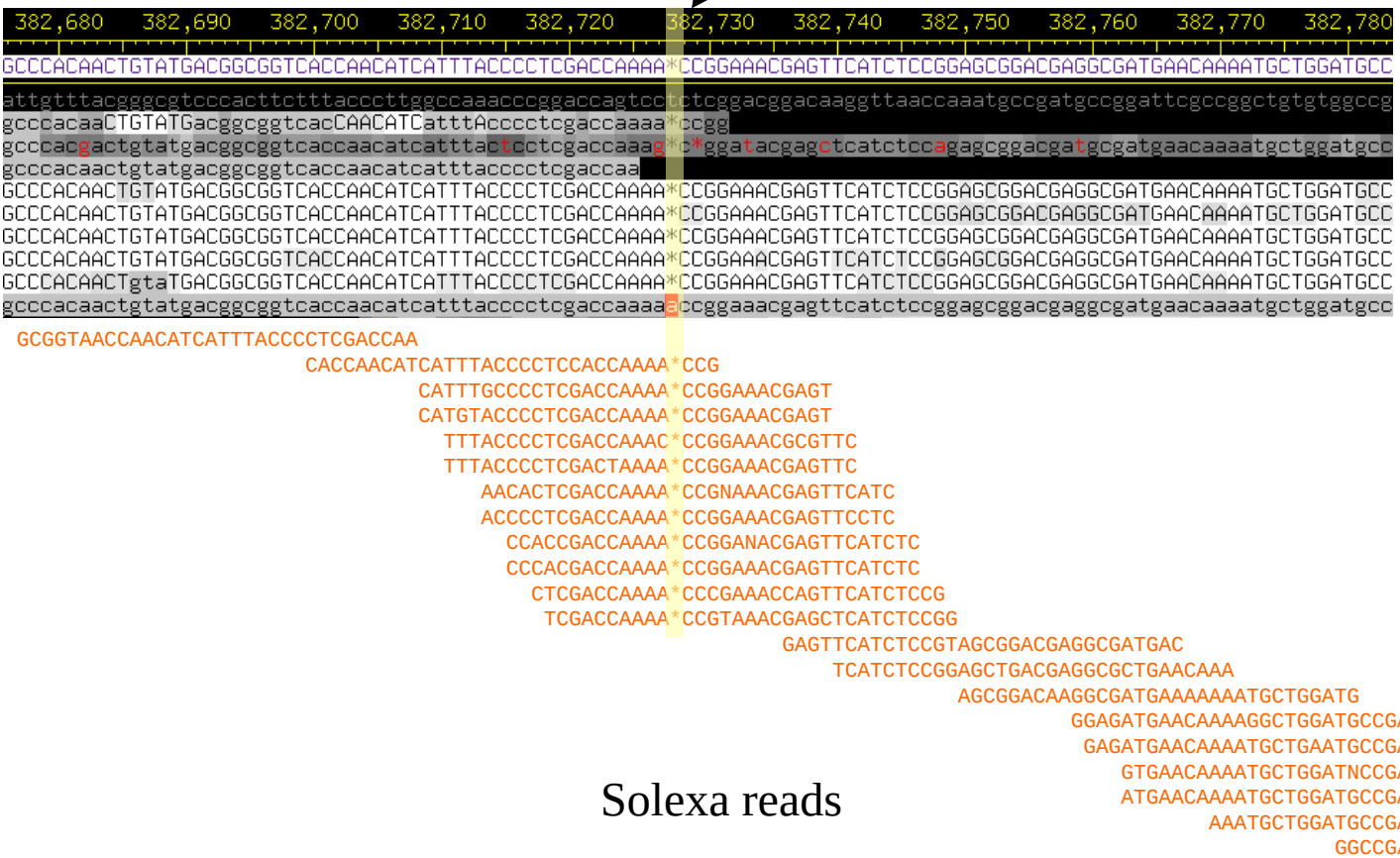
Solexa Sequencing:

- Identify and correct mismatches between Solexa and 454 only coverage

Frameshift Detection:

- HMM based frameshift detection tool (in collaboration with GeneMark developers at Georgia Institute of Technology)

Example: Solexa corrects frame shift (FS) in 454



Frame shift detected

```
 382,680   382,690   382,700   382,710   382,720    382,730   382,740   382,750   382,760   382,770   382,780
GCCCACAACTGTATGACGGCGGTCACCAACATCATTTACCCCTCGACCAAAA*CCGGAAACGAGTTCATCTCCGGAGCGGACGAGGCGATGAACAAAATGCTGGATGCC
attgtttacgggcgtcccacttctttacccttggccaaacccggaccagtcctctcggacggacaaggttaaccaaatgccgatgccggattcgccggctgtgtggccg
gcccacaaCTGTATGacggcggtcacCAACATCatttAcccctcgaccaaaa*ccgg
gcccacgactgtatgacggcggtcaccaacatcatttactcctcgaccaaag*c*ggatacgagctcatctccagagcggacgatgcgatgaacaaaatgctggatgcc
gcccacaactgtatgacggcggtcaccaacatcatttacccctcgaccaa
GCCCACAACTGTATGACGGCGGTCACCAACATCATTTACCCCTCGACCAAAA*CCGGAAACGAGTTCATCTCCGGAGCGGACGAGGCGATGAACAAAATGCTGGATGCC
GCCCACAACTGTATGACGGCGGTCACCAACATCATTTACCCCTCGACCAAAA*CCGGAAACGAGTTCATCTCCGGAGCGGACGAGGCGATGAACAAAATGCTGGATGCC
GCCCACAACTGTATGACGGCGGTCACCAACATCATTTACCCCTCGACCAAAA*CCGGAAACGAGTTCATCTCCGGAGCGGACGAGGCGATGAACAAAATGCTGGATGCC
GCCCACAACTGTATGACGGCGGTCACCAACATCATTTACCCCTCGACCAAAA*CCGGAAACGAGTTCATCTCCGGAGCGGACGAGGCGATGAACAAAATGCTGGATGCC
GCCCACAACTgtaTGACGGCGGTCACCAACATCATTTACCCCTCGACCAAAA*CCGGAAACGAGTTCATCTCCGGAGCGGACGAGGCGATGAACAAAATGCTGGATGCC
gcccacaactgtatgacggcggtcaccaacatcatttacccctcgaccaaaaaccggaaacgagttcatctccggagcggacgaggcgatgaacaaaatgctggatgcc
 GCGGTAACCAACATCATTTACCCCTCGACCAA
                       CACCAACATCATTTACCCCTCCACCAAAA*CCG
                                CATTTGCCCCTCGACCAAAA*CCGGAAACGAGT
                                CATGTACCCCTCGACCAAAA*CCGGAAACGAGT
                                  TTTACCCCTCGACCAAAC*CCGGAAACGCGTTC
                                  TTTACCCCTCGACTAAAA*CCGGAAACGAGTTC
                                     AACACTCGACCAAAA*CCGNAAACGAGTTCATC
                                     ACCCCTCGACCAAAA*CCGGAAACGAGTTCCTC
                                       CCACCGACCAAAA*CCGGANACGAGTTCATCTC
                                       CCCACGACCAAAA*CCGGAAACGAGTTCATCTC
                                         CTCGACCAAAA*CCCGAAACCAGTTCATCTCCG
                                          TCGACCAAAA*CCGTAAACGAGCTCATCTCCGG
                                                             GAGTTCATCTCCGTAGCGGACGAGGCGATGAC
                                                                 TCATCTCCGGAGCTGACGAGGCGCTGAACAAA
                                                                           AGCGGACAAGGCGATGAAAAAAATGCTGGATG
                                                                                    GGAGATGAACAAAAGGCTGGATGCCGAGATGC
                                                                                     GAGATGAACAAAATGCTGAATGCCGAGATGCT
                                                                                        GTGAACAAAATGCTGGATNCCGAGATGCTGCC
                                                                                        ATGAACAAAATGCTGGATGCCGAGCTGCCGCC
                                                                                               AAATGCTGGATGCCGAGATGCTGCCGTACTGT
                                                                                                         GGCCGAGATGCTGCCGTTGTGTGTTGGAGCGG
```

454 contig

Solexa reads

- -Frame shift detected in 454 contig.
- -Sanger and Solexa support the finished reference.
- -Development of both GeneMark based FS detector (in collaboration with GeneMark developers) and corrective tool are in progress.

100% of positive frame shifts detected in 454 contigs were confirmed by Solexa reads

454 Error Rate Analysis

SUMMARY:

Sanger/454 sequencing:

- detect 454 errors by read depth and percent 454 only coverage.

Sanger/454/Solexa sequencing:

- detect mismatches between Solexa and 454 to correct 454 errors.
- use frame shift detection to identify discrepancies between Sanger and Solexa sequence.

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Acknowledgements

PGF Microbial Genomics:

Alla Lapidus (Group Lead)
Eugene Goltsman
Brian Foster
Kurt M. LaButti

LLNL Microbial Finishing:

Patrick Chain (Group Lead)

Technology group:

Ed Kirton
Sirisha Sunkara
Feng Chen

Technology Dpt. Head:

Paul Richardson

Collaborators:

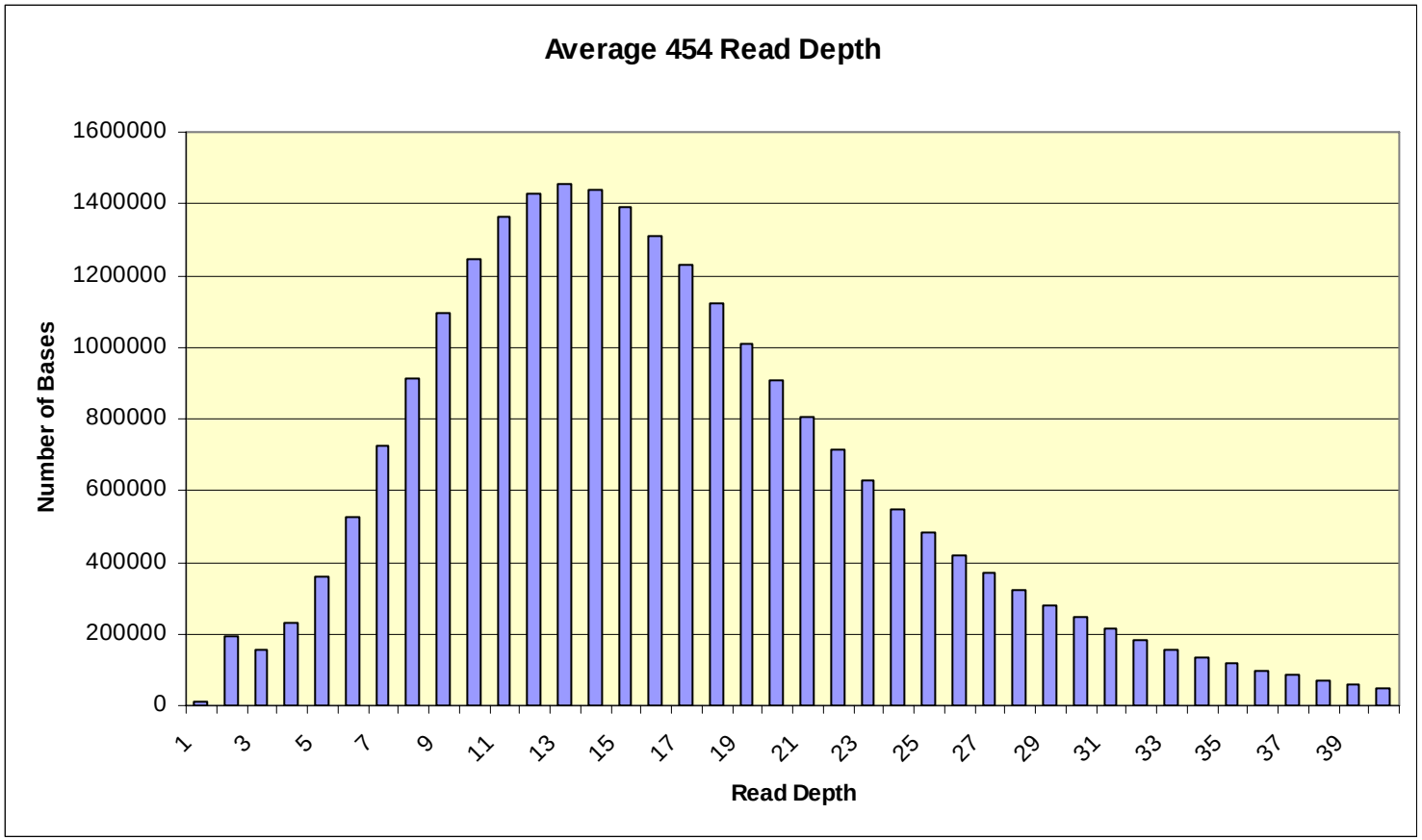
Prof. Mark Borodovsky
(Georgia Institute of Technology)

Informatics:

Pat Kale (Group Lead)

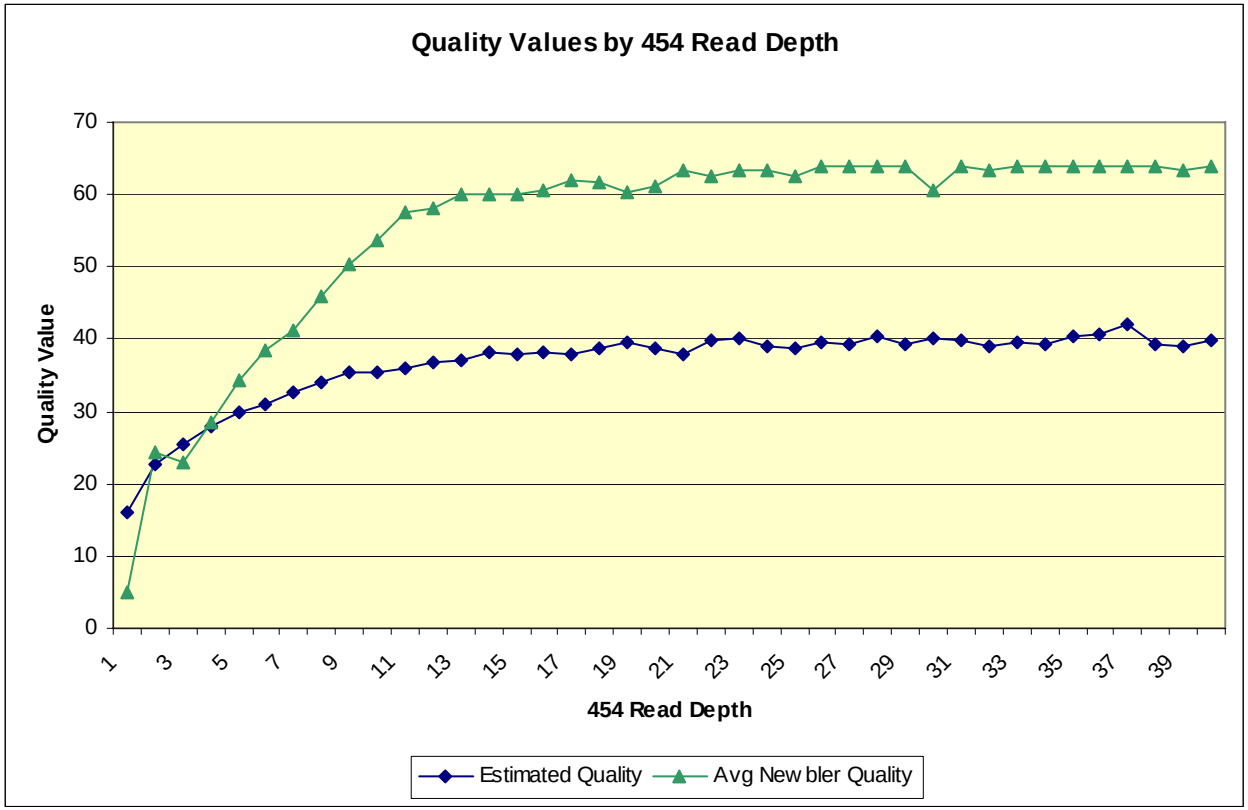
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Average 454 Read Depth



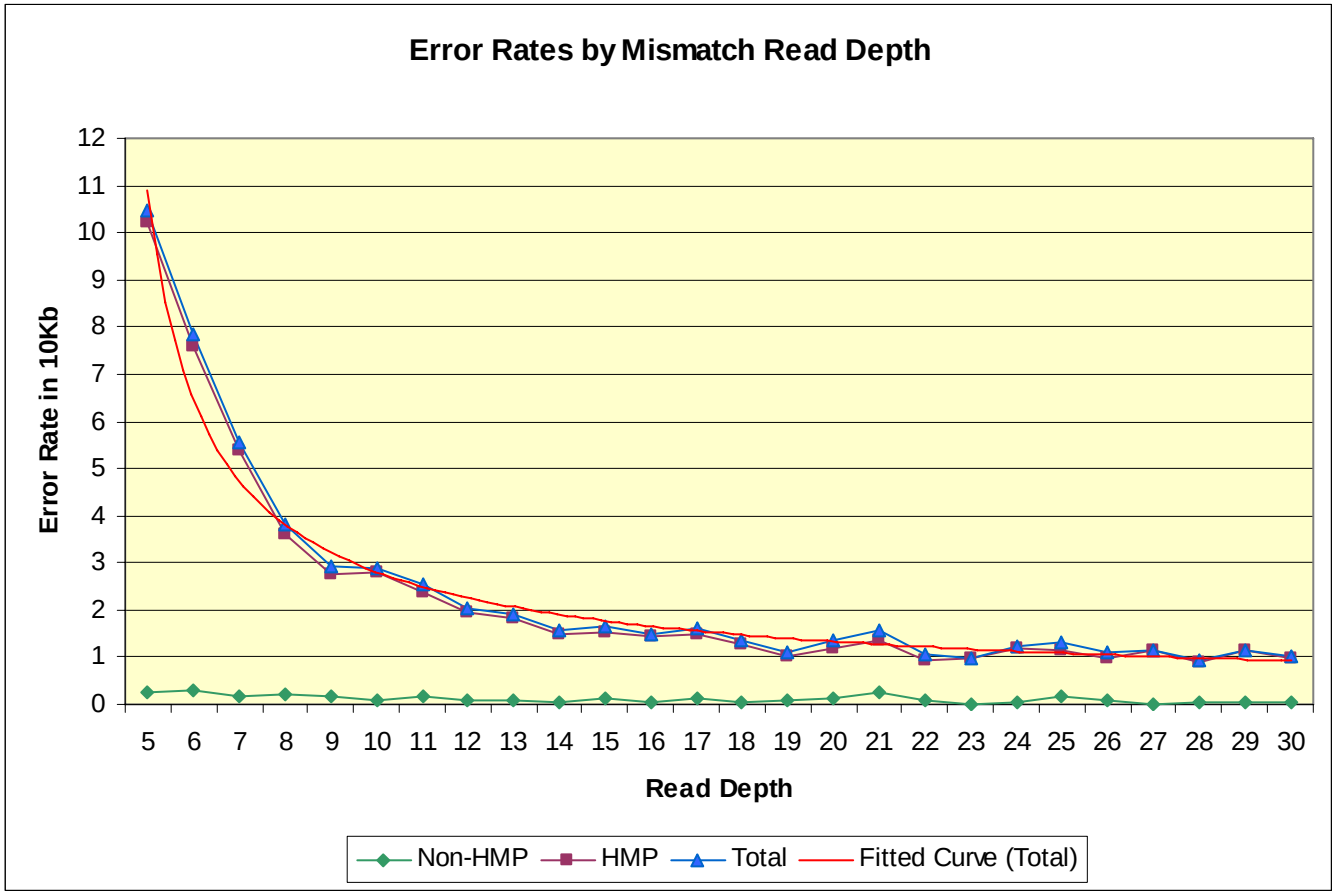
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Quality Values Based on 454 Read Depth



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Error Rates by Mismatch Read Depth



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Distribution of 454 Only Coverage