The instruments, the runs, the QC metrics, and the output

Peter Schweitzer, Director, DNA Sequencing and Genotyping Lab

- Overview
- Roche/454 GS-FLX
  - 454 (GSRunbrowser information)
  - Evaluating run results
  - Errors produced
  - Data files that are produced and distributed
- Illumina (Solexa)
  - Types of runs available
  - Evaluating run results
    - Summary stats page, error graphs, images
  - Errors produced
  - Data files produced and distributed
- Both platforms:
  - Multiplexing options
  - Sequence capture technologies
    - Nimblegen and other microarray-based
    - Agilent bead-based
    - Raindance Technologies PCR-based

Resources:

- Cornell DNA Sequencing listserver
  - To subscribe, send an email to: dna-sequencing-l-request@cornell.edu with the word “join” in message
- seqanswers.com

- Illumina website (www.illumina.com)
  - Publications (www.illumina.com/publications)
- 454 Website (www.454.com)
  - Publications (www.454.com/publications-and-resources)

Similarities:

Library preparations very similar (A and B adaptors on dsDNA fragments)
Clonal amplification of single molecules
Sequencing by synthesis
Both produce fasta files with quality scores

Differences:

454 sequencer

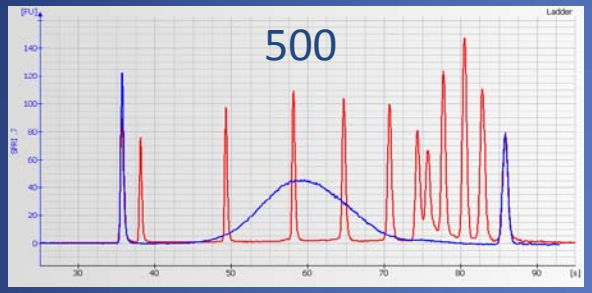
Read Lengths = 400 nt (average)
~1 million reads per full run
~ 400 Mb per full run
Errors are insertions/deletions
sff files also produced

Illumina sequencer

Read Lengths = 43, 86, 129 nt
~100-150 million reads per full run
Errors are substitutions

Sample Preparation (generation of small dsDNA fragments)

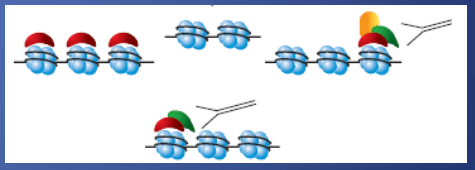
Random fragmentation (whole genome sequencing)
(5 ug = ~1.5 million copies of human genome)



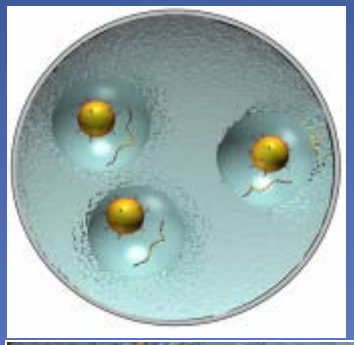
Micro RNAs (23-nucleotide, regulatory RNAs)

RNA (transcriptome)

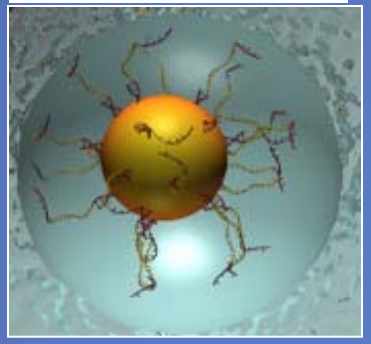
Chromatin immunoprecipitation (ChIP)



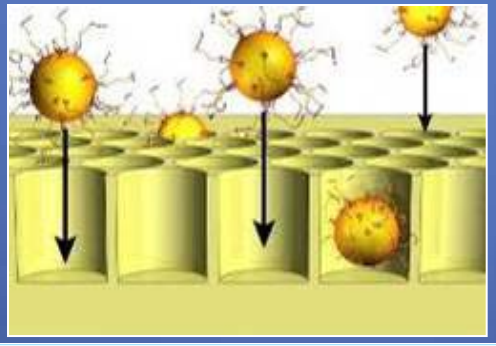
Roche/454 Genome Sequencer FLX

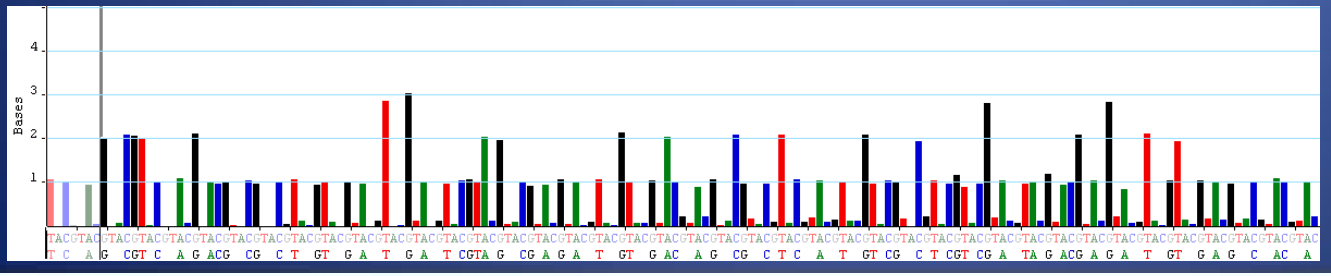


~1 copy per bead



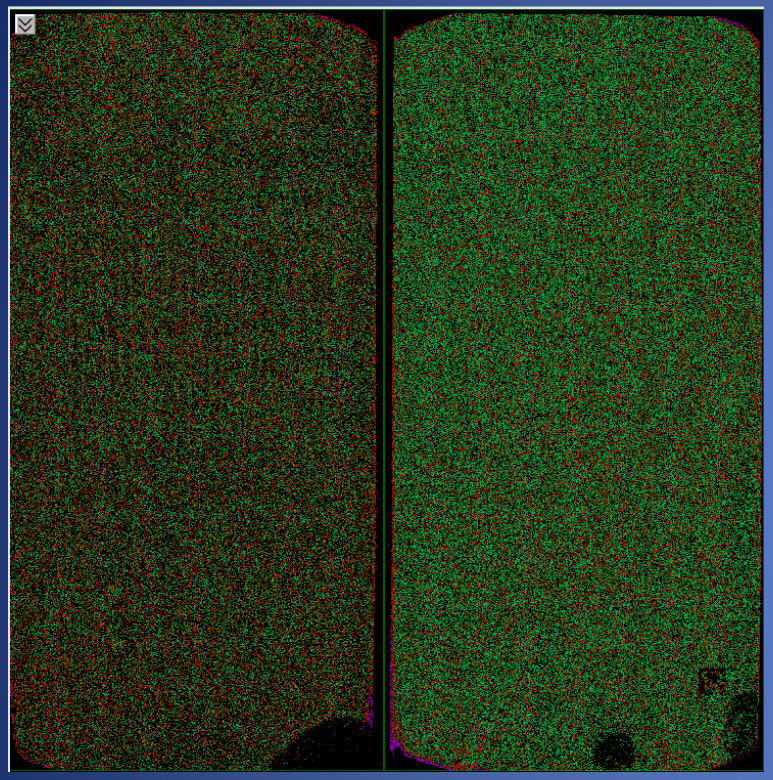
Clonal amplification





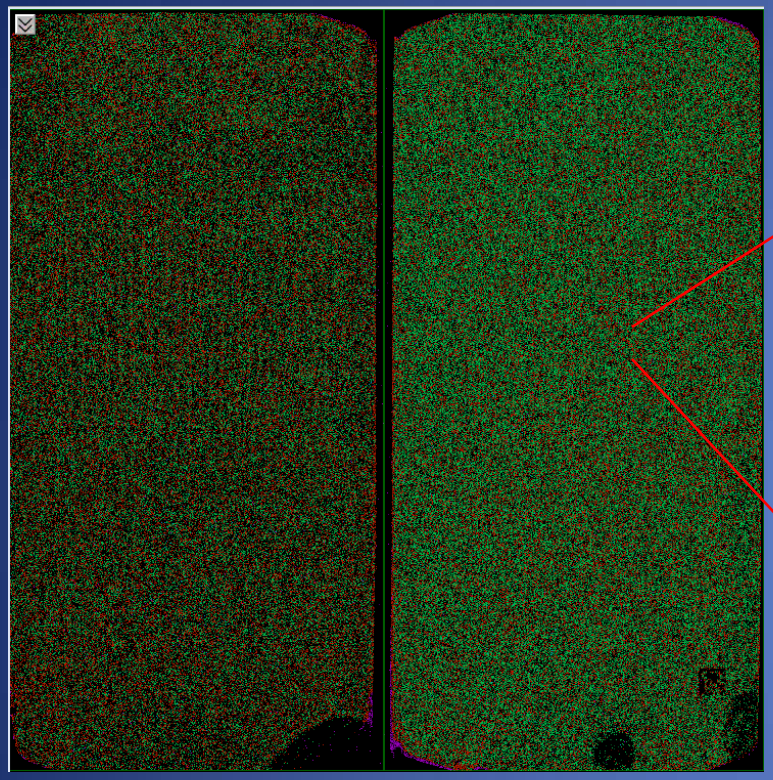
Parallel, clonal sequencing (pyrosequencing)

454: GS Runbrowser

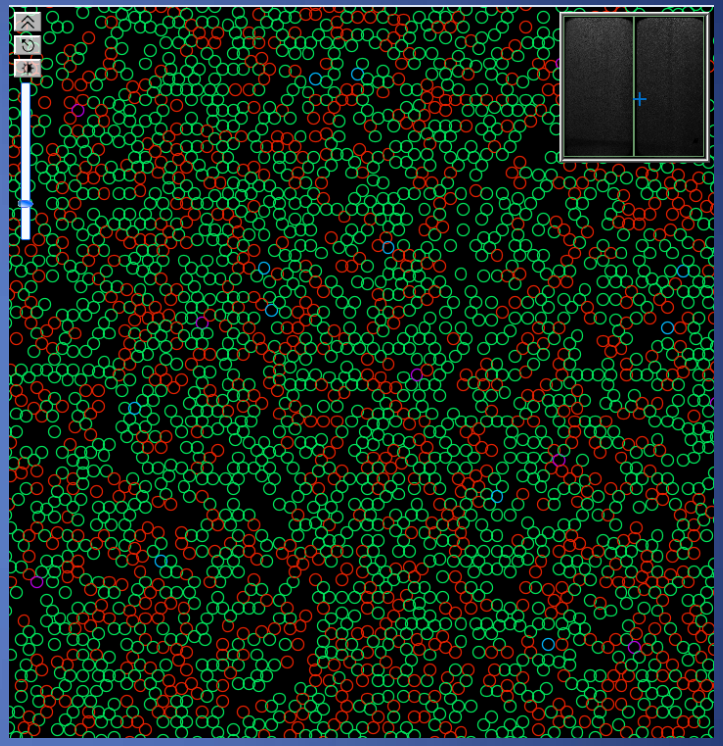


454 PTP Image Display

454: GS Runbrowser

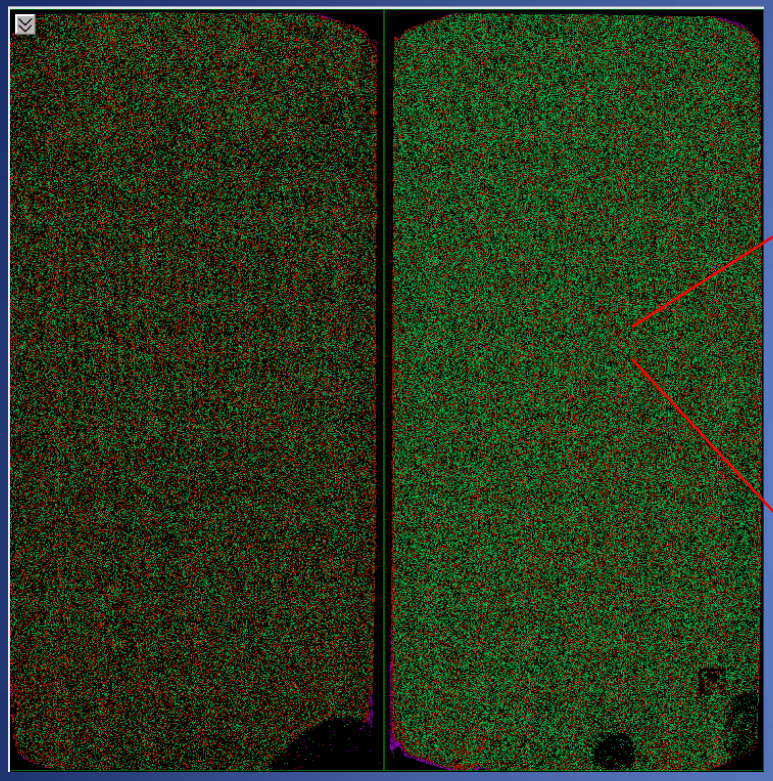


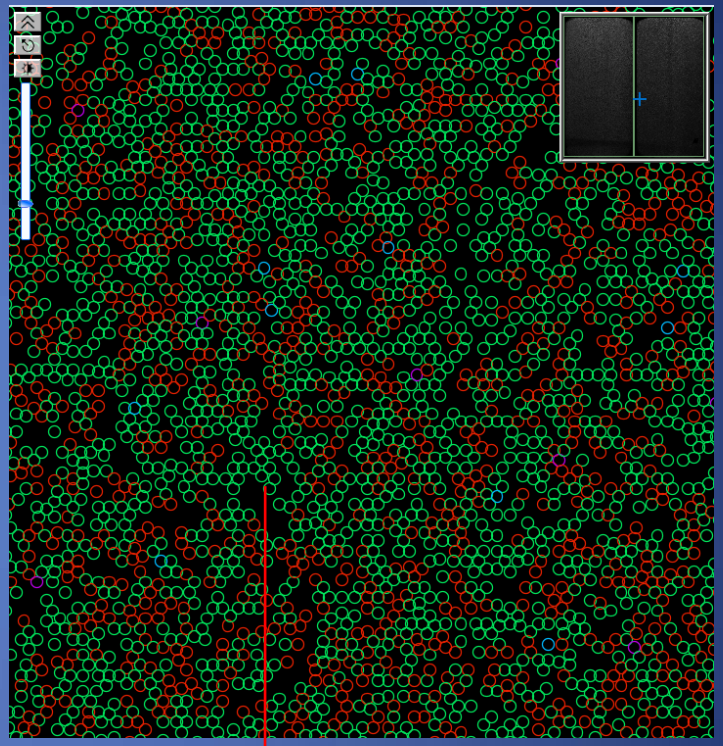
454 PTP Image Display

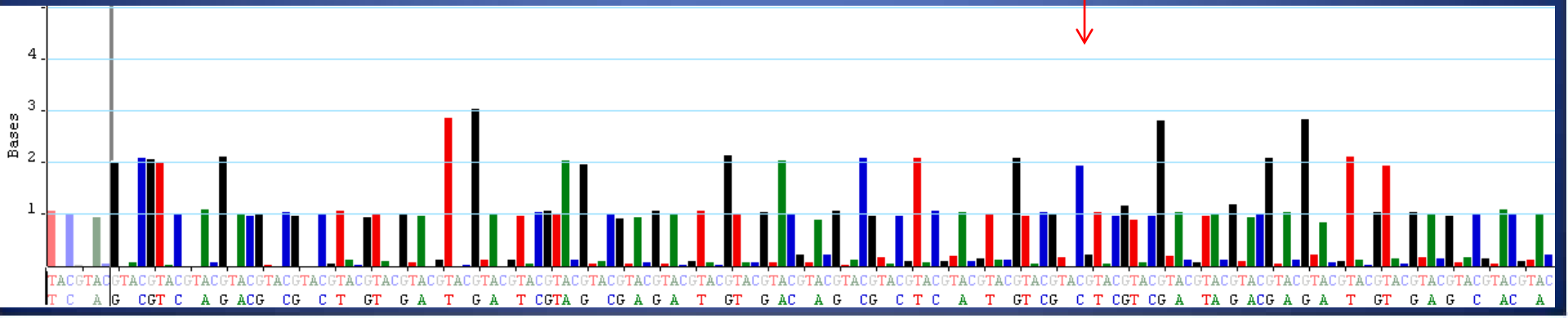


454: GS Runbrowser

454 PTP Image Display



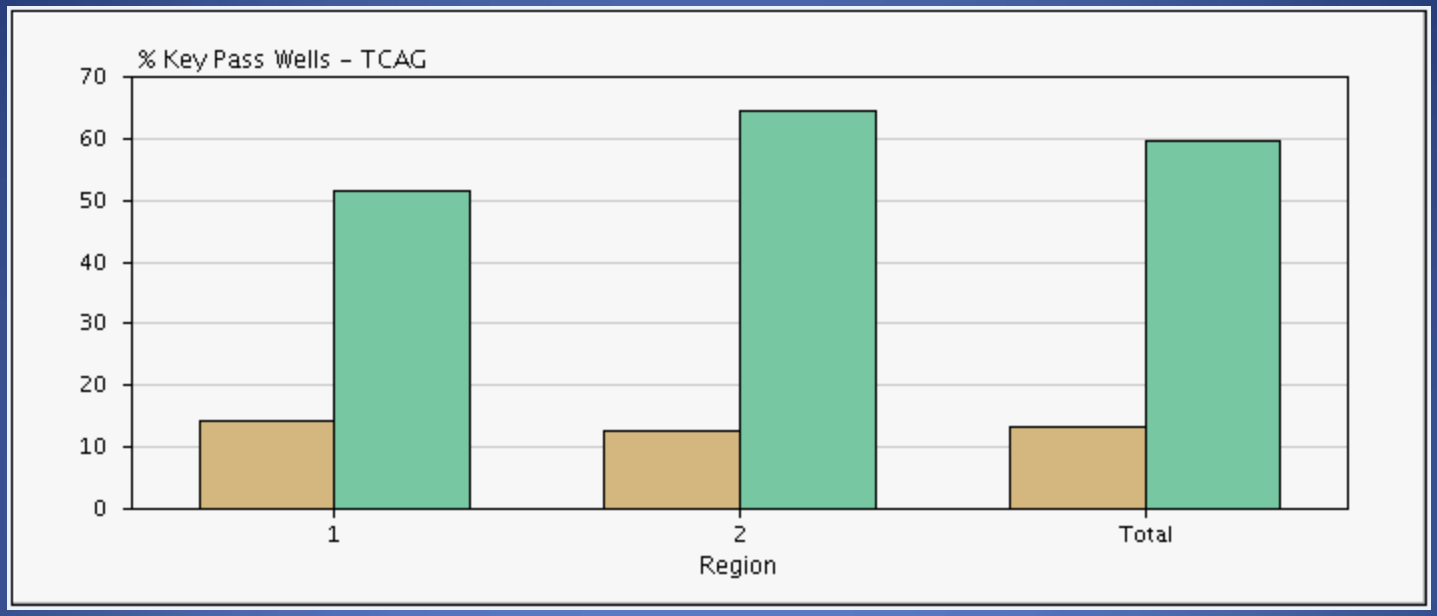




454 GS RunProcessor
Standard Filters/Trimming

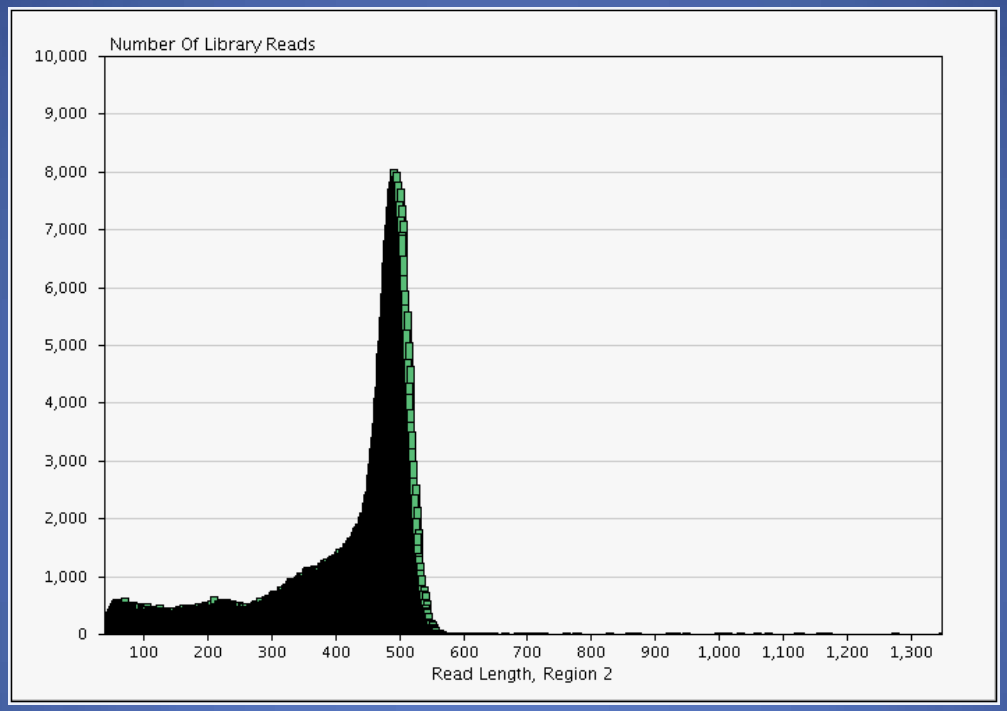
- Filtering
  - Bead index
    - library beads = TCAG
    - control beads = CATG
  - Dots filter (too many cycles with no base call)
  - Mixed filter (too many incorporations/half incorporations)
- Trimming
  - Quality filter (trims 3’ end for low quality)
  - Primer filter (trims any 454 adaptor sequences)
  - Valley filter (trims 3’ end or rejects for too many half-signals)

454 Read Filtering Stats



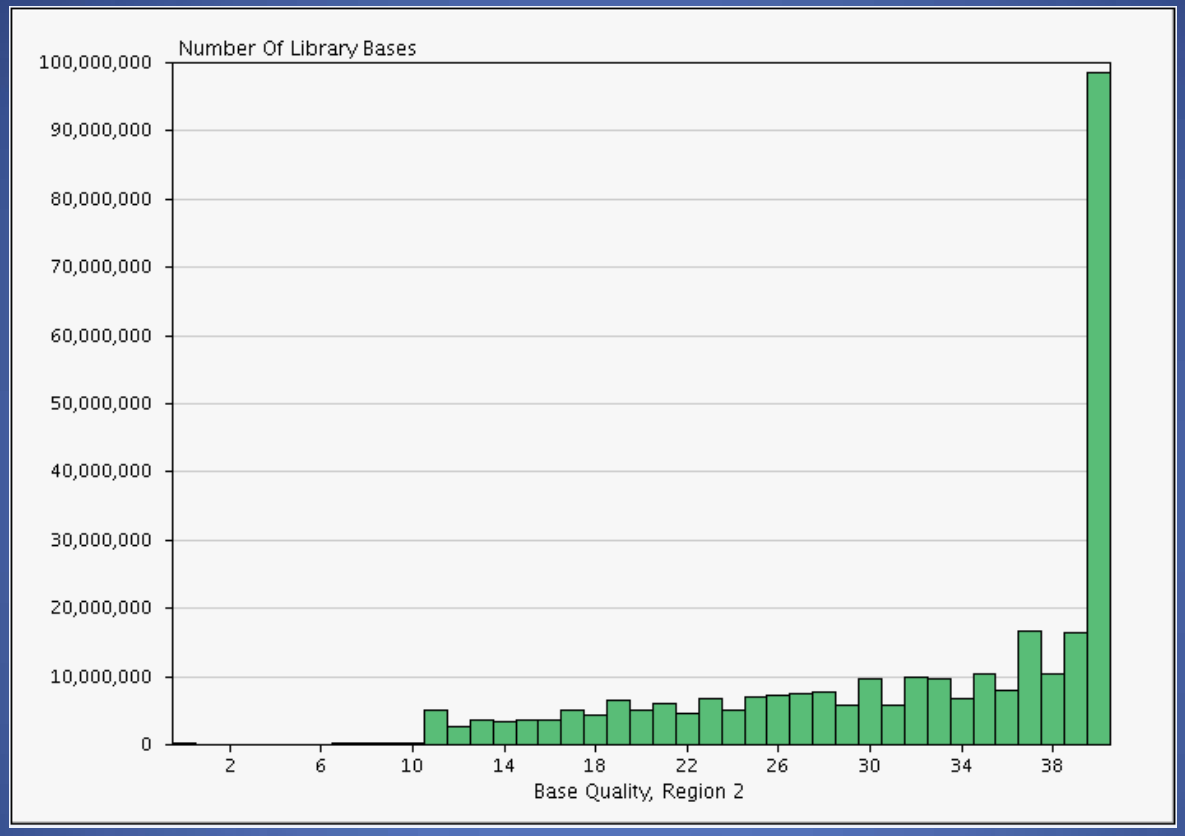
TCAG (Library)		Region 1	2	Total
	Raw Wells	737,992	1,161,887	1,899,879
	Key Pass Wells	722,069	1,148,532	1,870,601
Failed	Dot	69,970	57,459	127,429
	Mixed	32,198	87,658	119,856
	Short Quality	247,885	260,926	508,811
	Short Primer	97	1,043	1,140
	Passed Filter Wells	371,919	741,446	1,113,365
	% Dot + Mixed	14.15	12.63	13.22
	% Short	34.34	22.81	27.26
	% Passed Filter	51.51	64.56	59.52

454 Read Length Distribution



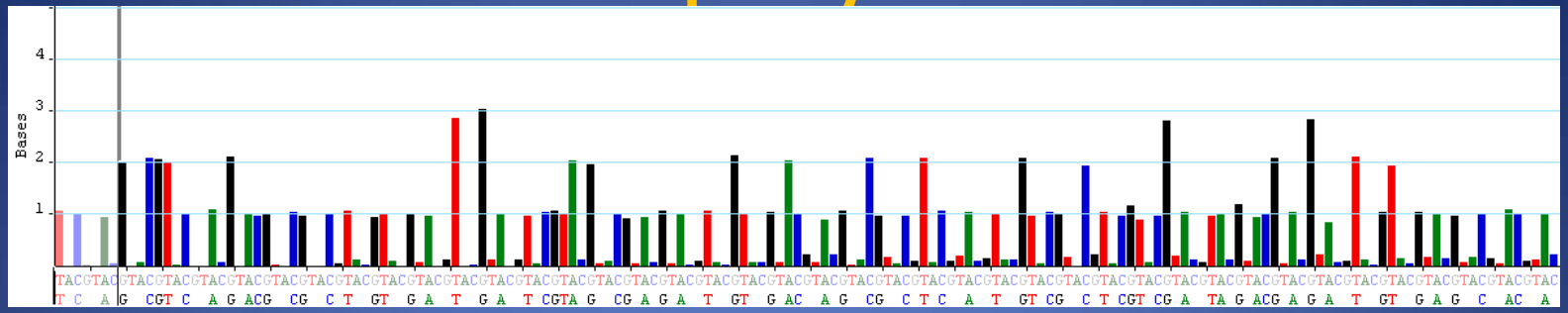
TCAG (Library)	Region 1	Region 2	Total
Raw Wells	737,992	1,161,887	1,899,879
Key Pass Wells	722,069	1,148,532	1,870,601
Passed Filter Wells	371,919	741,446	1,113,365
Total Bases	120,383,340	303,041,628	423,424,968
Length Average	323.90	408.79	380.31
Length Std Deviation	150.29	123.95	
Longest Reads Length	1,364	1,347	1,364
Shortest Reads Length	40	40	40
Median Reads Length	356.0	466.0	438.0

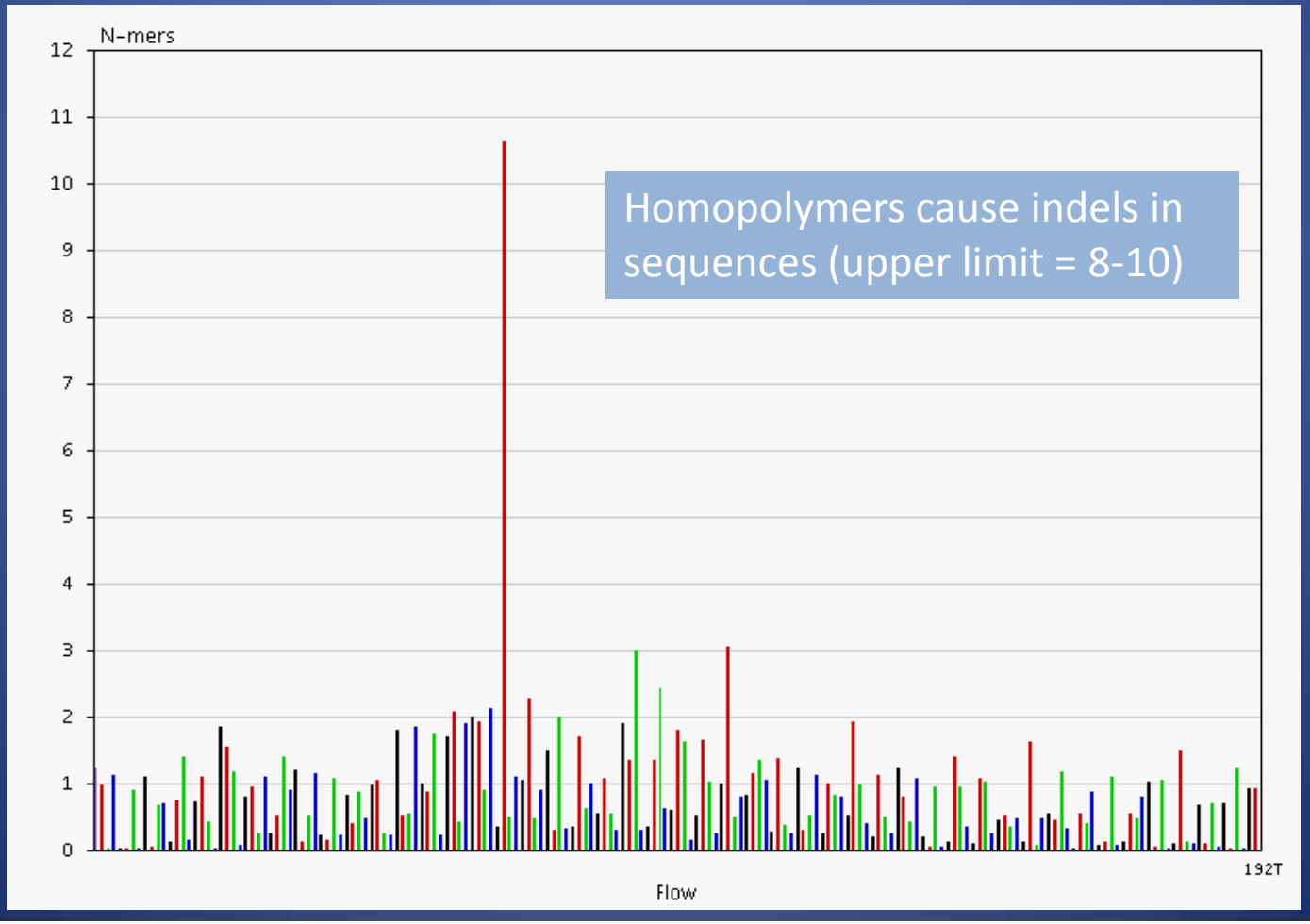
454 Base Quality Stats

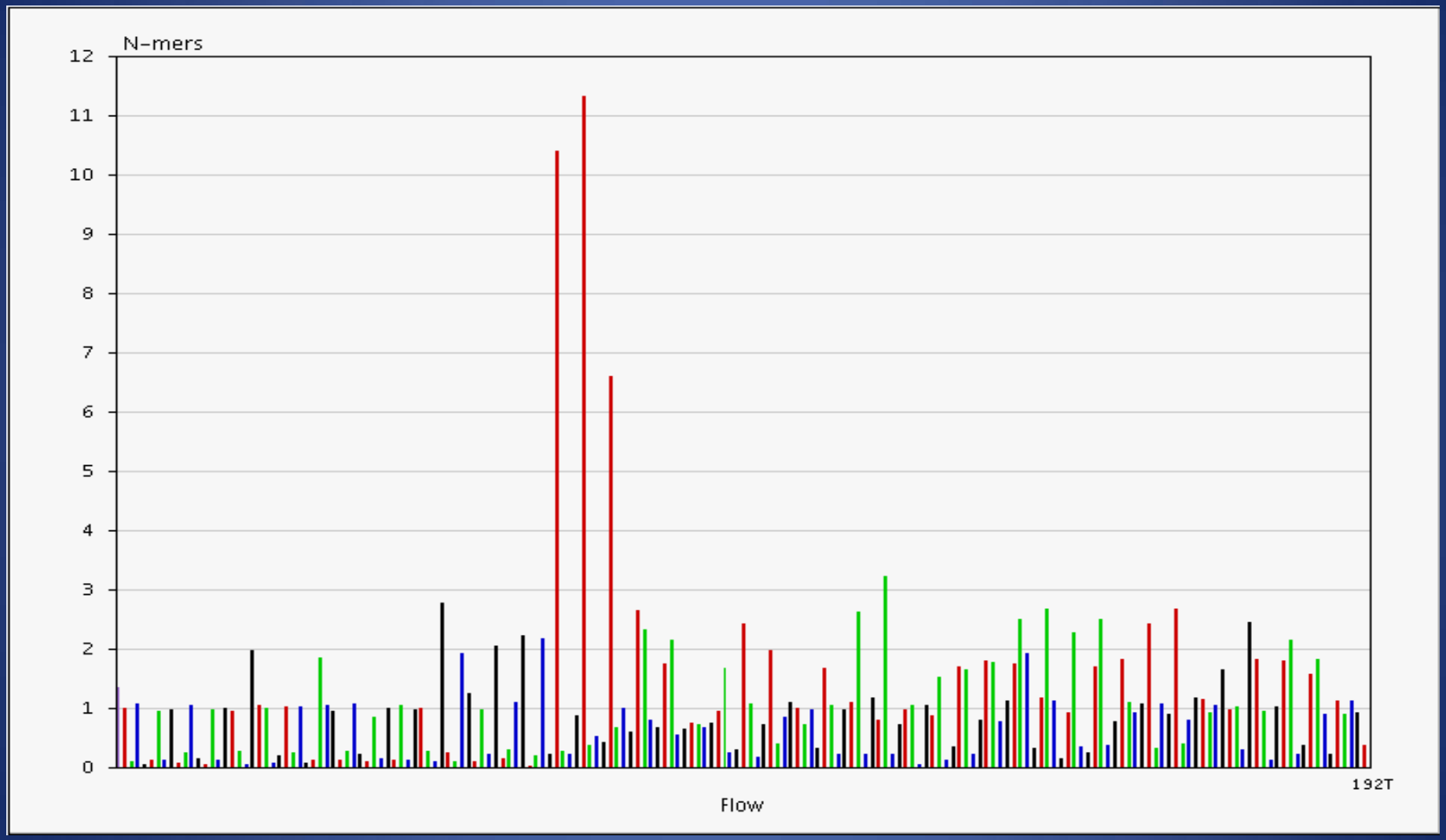


TCAG (Library)	Region		
	1	2	Total
Raw Wells	737,992	1,161,887	1,899,879
Key Pass Wells	722,069	1,148,532	1,870,601
Passed Filter Wells	371,919	741,446	1,113,365
Total Bases	120,383,340	303,041,628	423,424,968
Quality Average	31.78	32.09	32.00
Quality Std Deviation	9.14	8.67	

454: Homopolymer Errors






N-mers
12
11
10
9
8
7
6
5
4
3
2
1
0
192T
Flow

454: Data Files Produced

1.CAT.454Reads.fna	20-Jul-2009 19:18	3.3M
1.CAT.454Reads.qual	20-Jul-2009 19:18	9.0M
1.TCA.454Reads.fna	20-Jul-2009 19:20	137M
1.TCA.454Reads.qual	20-Jul-2009 19:20	365M
2.CAT.454Reads.fna	20-Jul-2009 19:23	2.8M
2.CAT.454Reads.qual	20-Jul-2009 19:23	7.6M
2.TCA.454Reads.fna	20-Jul-2009 19:29	335M
2.TCA.454Reads.qual	20-Jul-2009 19:29	908M
454BaseCallerMetrics.csv	20-Jul-2009 19:29	18K
454BaseCallerMetrics.txt	20-Jul-2009 19:29	57K
454DataProcessingDir.xml	20-Jul-2009 18:02	351
454QualityFilterMetrics.csv	20-Jul-2009 19:29	937
454QualityFilterMetrics.txt	20-Jul-2009 19:29	2.0K
454RuntimeMetricsAll.csv	20-Jul-2009 19:29	16K
454RuntimeMetricsAll.txt	20-Jul-2009 19:29	31K
gsRunProcessor.log	20-Jul-2009 19:29	9.4K
gsRunProcessor_err.log	20-Jul-2009 19:29	252
regions/	20-Jul-2009 19:17	-
sff/	20-Jul-2009 19:20	-

Region 1 fasta and qual files, control beads

Region 1 fasta and qual files, sample reads

Region 2 fasta and qual files, control beads

Region 2 fasta and qual files, sample reads

Basecaller metrics

Filter metrics

sff files

.fna file

```
>FZAOHM402I4584 rank=0083510 x=3632.0 y=1366.0 length=528
TCGAAAGGATGAAAACAGAAGATCATCCTTTAACGGCACTTTGAAGCGCGAGAAGGCGTA
GGCCGCCGCGGTCCCAAGCACAATACACAGGACCGTCGATCCAAAACCGATGATGACGGA
ATTGGTAAAGCGCTCCGCAAAGCGCGACGGGCCAGCGATGACAAAGCCATCCTTGCGAAC
AATGCGATCATACCACGTCGTCGGCTCGCCCAGCGTTTGCAGCTCCTCTGCGGACGCCCG
CGTGCGGCTGGTAAAAAGGTTGACATATCCTTCGAGCGTTGGCGAAAACAACGTCTTCGG
CGGATAGGCAATCGCATCCGACGGCGACTTGAAGCCCGTTCCGATAATCCACAGCAGTGG
CAGGAGCGTTATGATCGCATAGACGACCACAAGGCTGCCTGCGAACCATTTTTTGACGGT
TCGAGGGTTCTGTAACCGAATAGCTCATCTTTGTTTGACCTTGTTCAGGGCTTTTGACGT
AGATCGACGCGAGGCCGAACACGGTGACGAAGAGAATGACCGCGTAGG
>FZAOHM402IN29O rank=0083511 x=3437.0 y=3066.0 length=523
ATCGTCGAACAGCGAGATCTTGGCGTAGGACAGGAGTACGCCGATTTCCGGTCGCGTCAG
GGATTGACCGCCTGCATAGCGCTCGGCCAGCGCCGCATCATCGGGTAGGGTTTCCACCTT
GCGGTTCAGTCGGCCAGAAGCCTCCAAGACGGTCATCAGCCGGGACAGTTCGTCACGATT
TGCGCTGCCCTGACGCGCCACCAGCGAAATCGCCAGGGATTGGAGATAGTTGTTGCGCAA
TACCAGTTCGCCGACTTCCTCGGTCATCGAGGCCAGCAGTTTGTTGCGGGCCGGGAGATC
AAGGCGTCCGGTGGATACCGCCGTTGACAGGGCAATCTTGATATTGACCTCGACGTCGGA
GGAATTGACGCCCGCCGAATTGTCGATGGCGTCGGAATTGCAGCGTCCGCCCTTCAGCGA
AAAGGCGATGCGGCCCTTCTGGGGTGAGGCCAAGATTGGCGCCTTCGCCGATCACCTTGG
CGCCGACCTCGTCCGCCGTAATGCGGATCGGGTCGTTCGCACG
```

.qual file

```
>FZAOHM402I4584 rank=0083510 x=3632.0 y=1366.0 length=528
40 40 39 21 21 21 40 40 40 40 40 21 21 21 21 39 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 38 38 38 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 26 26 26 34 34 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 21 21 21 21 39 39 40 37 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 35 21 21 21 26 39 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 39 39 39 40
40 40 40 40 39 39 39 40 40 40 40 40 23 23 25 22 22 24 22 24 24 36 36 36 36 36 29 26 27 27 28 28 36 38 36 33 22 22 22 22 26 28 36 30 24 24 24 24 32 25 35 40 40 35 35 34 40 40 40 40
40 40 40 40 40 40 40 40 39 39 39 39 40 40 40 40 40 40 40 40 26 19 20 20 28 33 40 36 36 36 36 22 24 24 34 33 29 24 24 22 20 20 34 32 24 24 24 28 29 33 34 36 36 40 40 36 36 39 40 40
40 40 40 39 39 35 35 35 40 40 40 40 40 40 40 40 40 35 35 35 35 35 40 40 40 40 39 39 39 39 39 39 39 36 40 35 35 35 28 30 29 22 18 13 13 12 12 16 11 11 11 11 11 11 18 26 31 28 28 23
21 21 16 21 15 15 15 16 16 27 23 31 34 31 31 32 32 31 32 32 36 33 36 36 36 32 34 34 34 32 32 32 36 28 28 28 36 36 36 36 36 34 31 27 22 25 25 14 14 14 20 14 14 14 14 25 30 30 26 27
25 25 29 36 36 36 36 36 36 31 31 31 31 32 36 36 33 36 36 33 19 22 22 22 22 19 31 34 31 29 29 28 28 28 31 31 32 27 27 21 21 19 28 31 30 21 19 19
>FZAOHM402IN29O rank=0083511 x=3437.0 y=3066.0 length=523
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 39
39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 37 36 36 36 30 19 19 19 20 20 26 36 36 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39
39 40 40 40 40 40 40 38 38 38 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 30 21 21 20 30 34 34 34 34 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 39 39
39 39 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 39 40 40 40 40 40 39 39 40 30 21 21 36 38 33 33 33 22 22 14 14 16 24 25 36 36 36
36 36 36 36 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 34 34 35 40 40 40 40 39 39 24 24 24 35 39 40 40 40 40 40 40 40 40 40 39 39 39 40 39 39 39 39 36 40 40 40 40 40 40 40
22 22 24 22 31 31 35 36 32 26 17 17 17 22 20 20 26 31 31 31 31 33 33 34 39 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 39 39 40 40 40 39 30 30 30 34 34 34 40 40 32 40 40 40 40 27
27 25 25 26 26 27 30 31 31 31 27 23 23 20 20 20 23 23 12 16 11 11 11 11 12 10 24 19 19 16 16 18 18 21 28 32 34 34 34 34 31 36 36 36 36 35 36 36 36 31 31 31 36 36 34 34 34 34 34 31
22 19 18 18 18 27 31 31 36 33 36 36 36 36 36 40 40 36 36 36 34 32 34 36 36 36 36 36 27 24 22 22 26 26 24 34 34 34 29 28 27 22 22
```

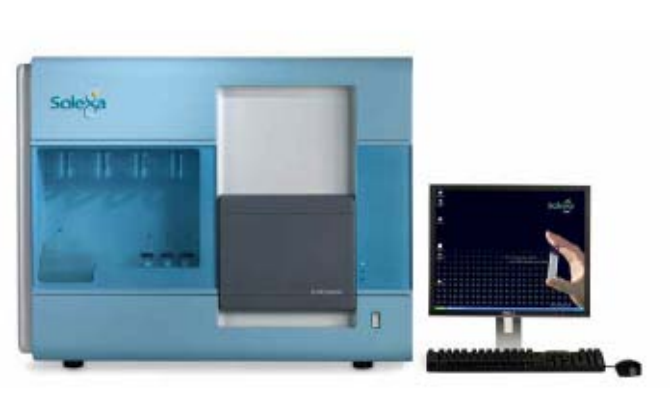
Distribution of Results

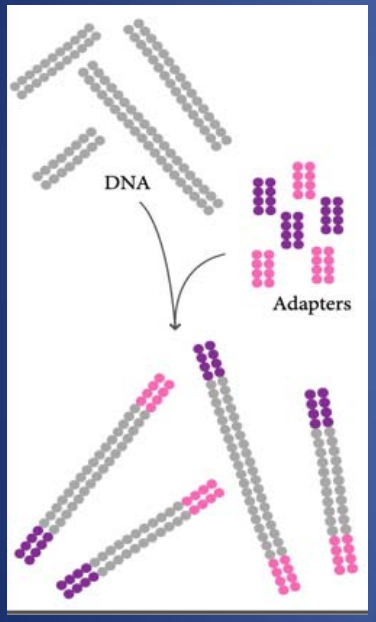
- Images captured of metrics using RunBrowser
  - Read length histogram, read length table, base quality histogram, quality table
- Data folder zipped, including:
  - Folder of sff files
  - fna and qual files
  - All metrics files and images
- Distributed to User using Cornell Dropbox

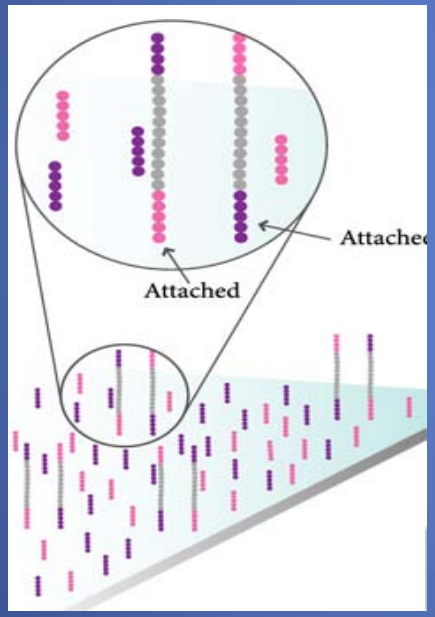
Major Applications for Roche/454 GS-FLX

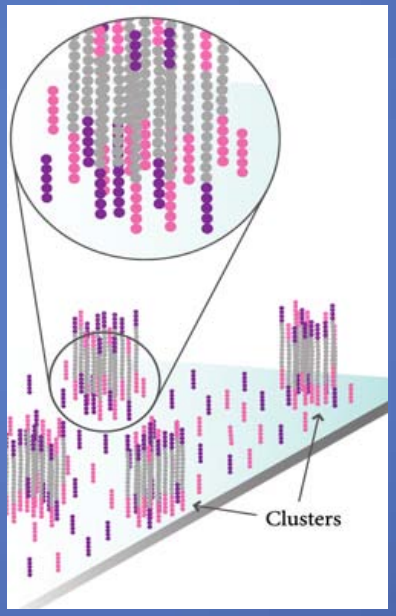
- Genomic *de novo* sequencing
- Metagenomics
- Transcriptomes

Illumina Genome Analyzer

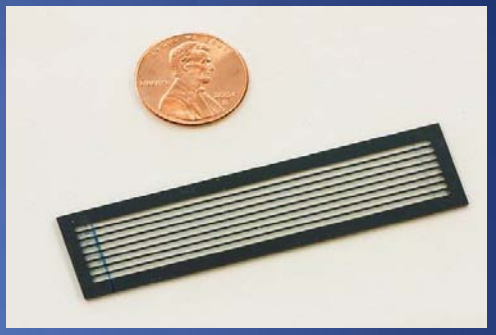




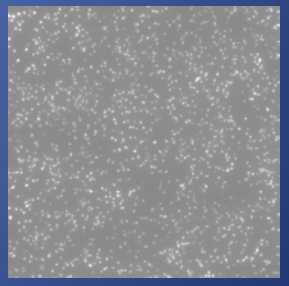


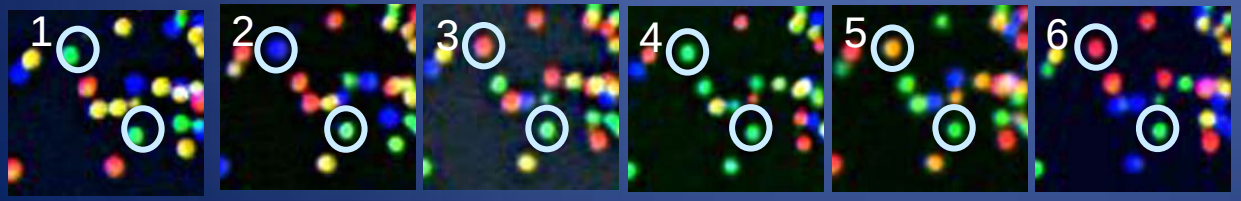


Clonal amplification



Parallel, clonal sequencing (reversible, fluorescent dye-terminator)



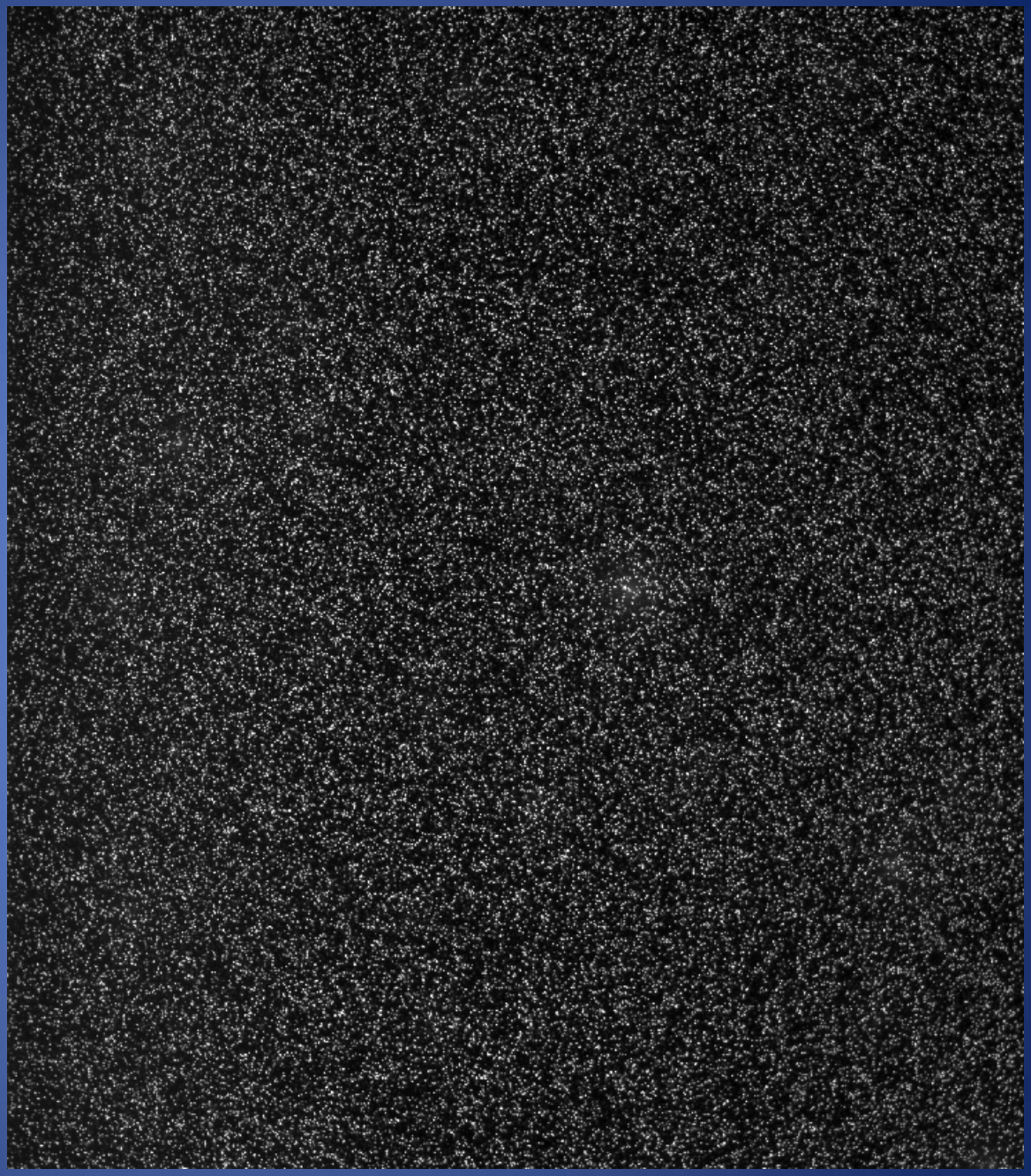


T G C T A C …

Illumina: Runs Available

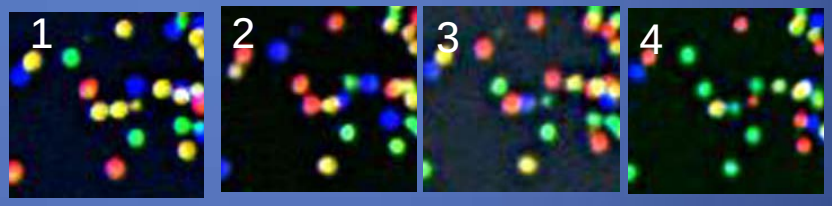
- Single-end 43 or 86 nt (129 nt possible)
- Paired-end reads from short fragments (200-250 bp fragments)
  - 2 x 43 nt
  - 2 x 86 nt
  - 2 x 129 nt
- Mate-pair large insert libraries
  - 2 kb to 5 kb “insert” size

- Each image (tile) contains 150K to 200K clusters
- Each lane has 120 tiles for GAIIx
- Current yield 10-20 million reads per lane
- ~330,000 images saved per 86 nt run = ~2 Tb



Illumina Filtering

- Identifies and discards clusters with mixed bases
  - Currently, first 4 cycles analyzed



- Reads with low signal intensity, or signal to noise ratio, removed (generally over the first 25 bases)

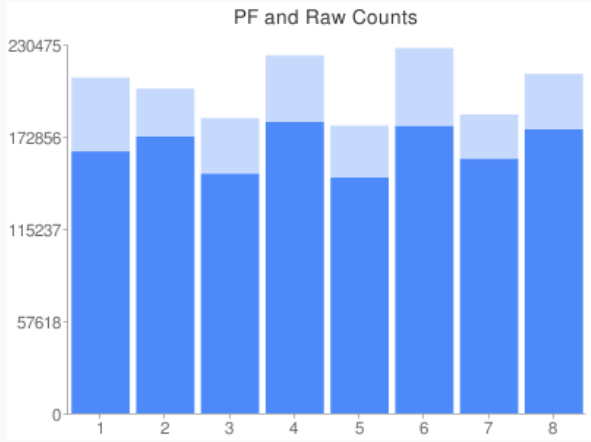
Illumina: Evaluating Run Results (phiX control alignment)

phiX control lane

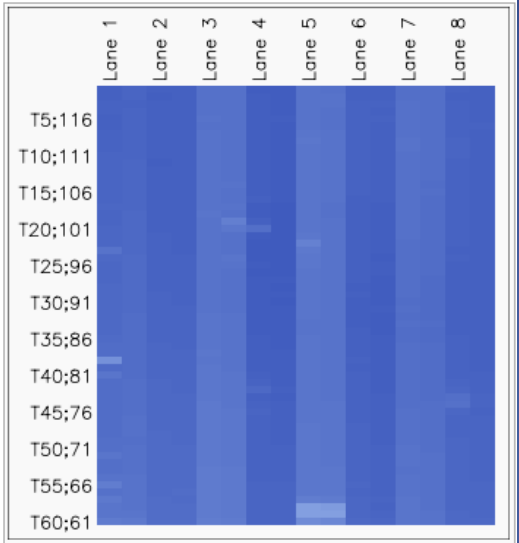
100301 HWI-EAS339 0004 61GGLAAXX

Date	2010-03-03 22:24:22
Machine	HWI-EAS339
RunFolder	**100301_HWI-EAS339_0004_61GGLAAXX**
Yield	6922448000
Raw	197257646
PF	160987282
#Reads	1

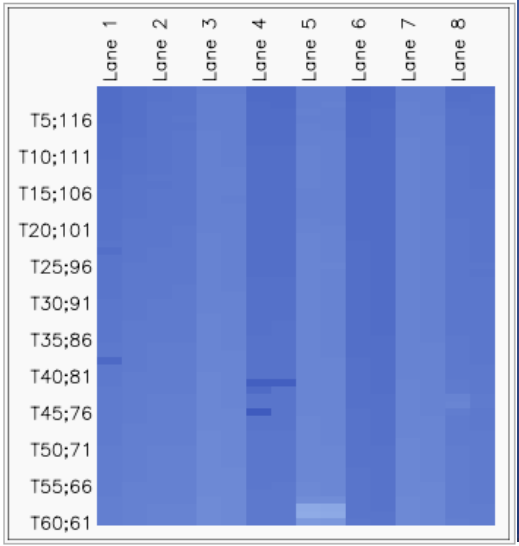
Lane	Aligned	Passed	Score	Errors
lane 1	%	77.87%		
lane 2	%	85.11%		
lane 3	%	81.40%		
lane 4	%	81.27%		
lane 5	%	81.74%		
lane 6	%	78.21%		
lane 7	%	85.10%		
lane 8	98.22%	83.52%	210.17	0.19



clusterCountPF



clusterCountRaw

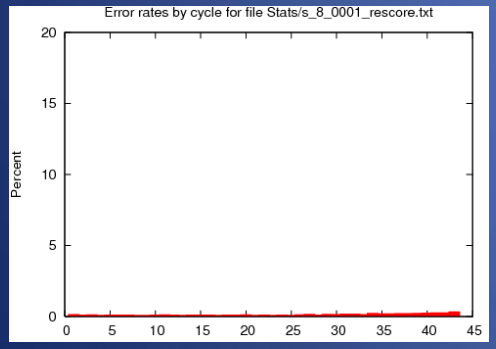


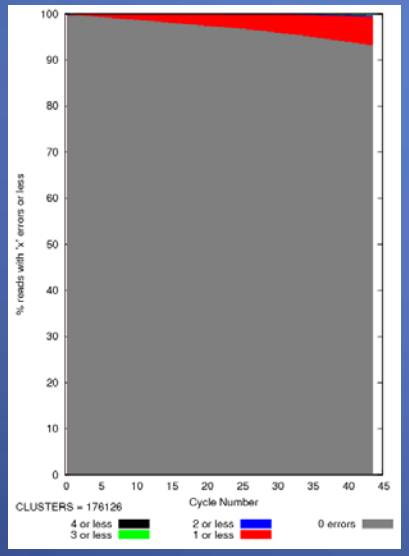
Illumina: Evaluating Run Results

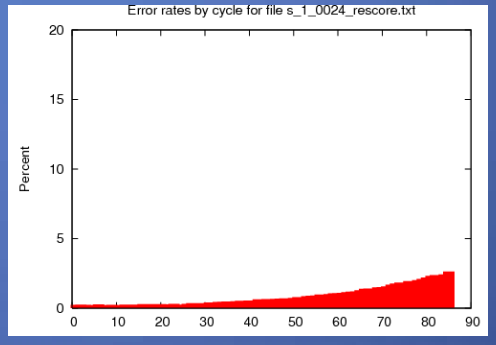
phiX control lane

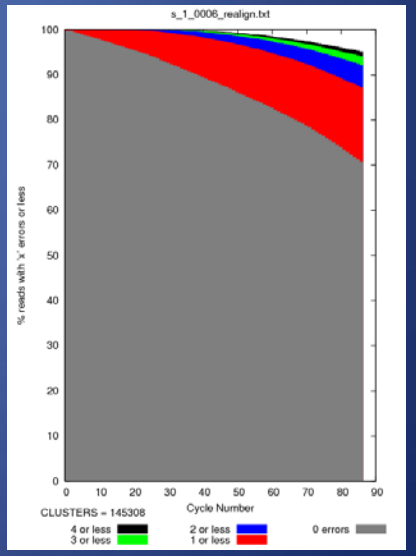
Lane Results Summary

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	853213	212552 +/- 12183	165352 +/- 10186	244 +/- 24	77.95 +/- 6.78	77.87 +/- 3.60	0	0	0
2	901469	205365 +/- 9580	174703 +/- 6509	264 +/- 11	82.92 +/- 0.95	85.11 +/- 0.99	0	0	0
3	781986	186267 +/- 7726	151548 +/- 5042	249 +/- 15	79.85 +/- 3.61	81.40 +/- 1.58	0	0	0
4	946567	226073 +/- 10819	183443 +/- 5169	245 +/- 17	81.85 +/- 2.96	81.27 +/- 3.30	0	0	0
5	769799	182729 +/- 16211	149186 +/- 12172	218 +/- 20	81.68 +/- 1.00	81.74 +/- 1.95	0	0	0
6	929700	230475 +/- 6455	180175 +/- 3539	227 +/- 19	80.66 +/- 0.99	78.21 +/- 1.67	0	0	0
7	823872	187669 +/- 4859	159665 +/- 3417	224 +/- 19	82.85 +/- 1.22	85.10 +/- 1.59	0	0	0
8	915842	212684 +/- 8466	177489 +/- 4287	209 +/- 19	83.09 +/- 2.70	83.52 +/- 1.99	98.22 +/- 0.03	210.17 +/- 1.31	0.19 +/- 0.13
Tile mean across chip									
Average		205477	167695	235	81.36	81.78	98.22	210.17	0.19





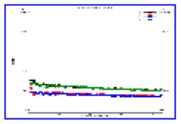
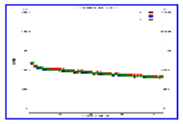
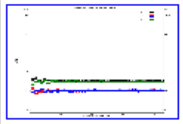
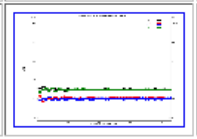
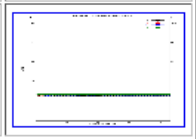
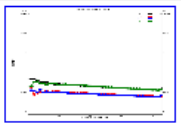
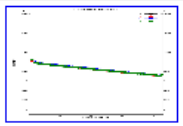
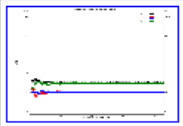
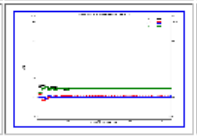
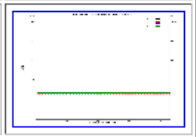
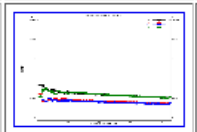
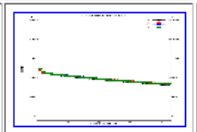
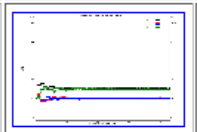
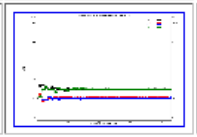
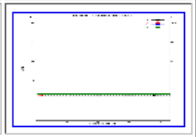
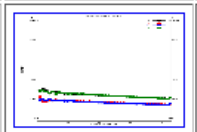
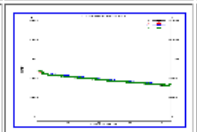
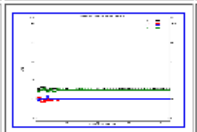
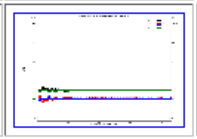
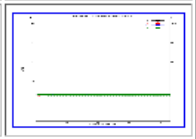
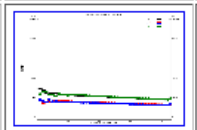
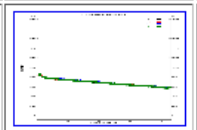
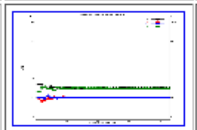
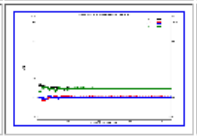
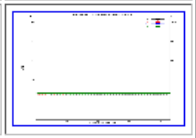
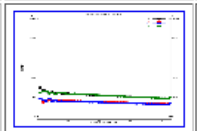
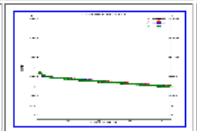
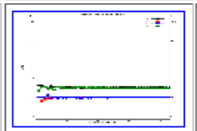
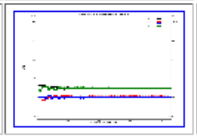
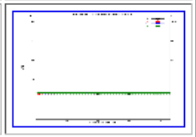
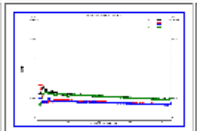
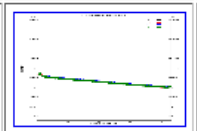
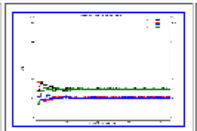
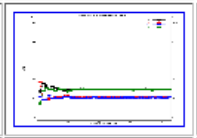
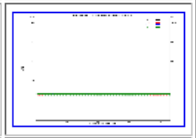
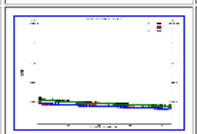
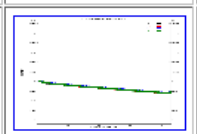
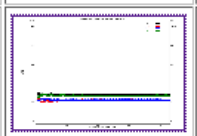
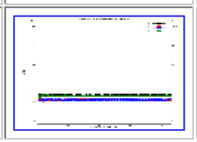
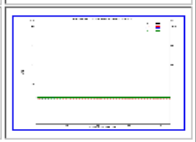




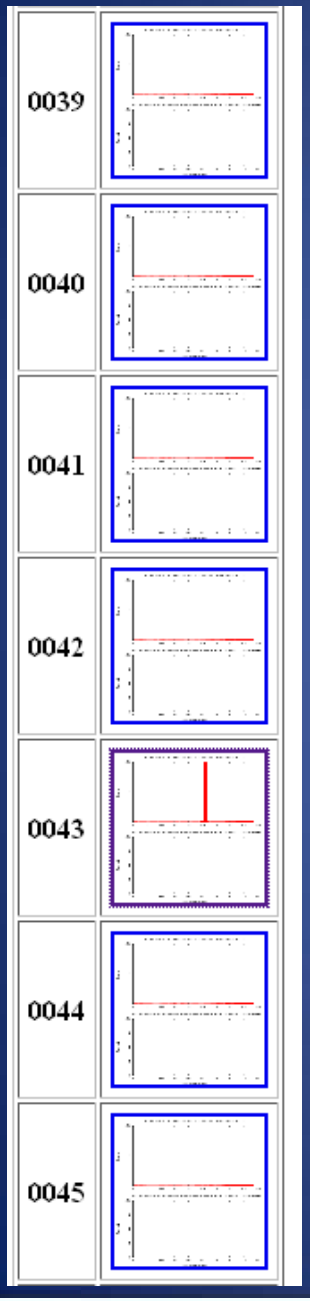
43 nt run

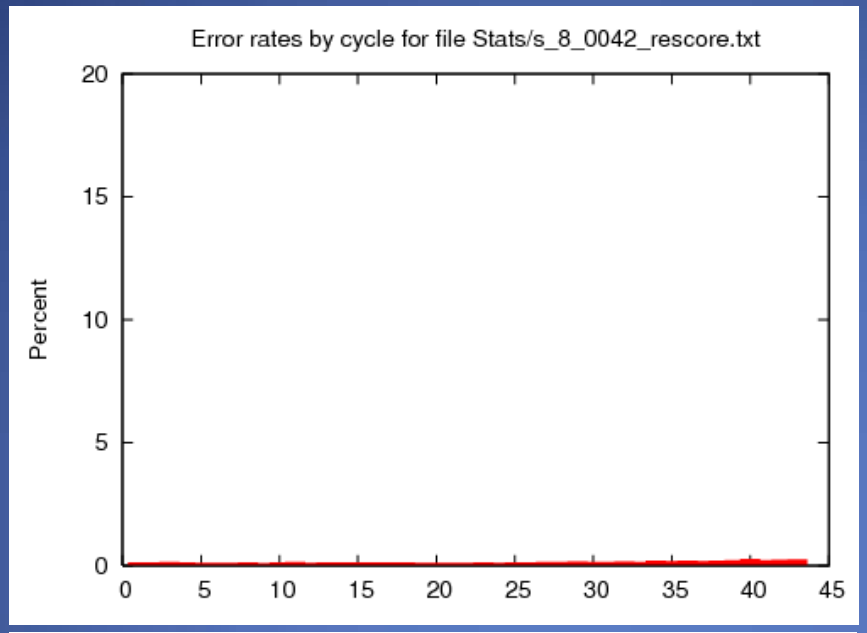
86 nt run

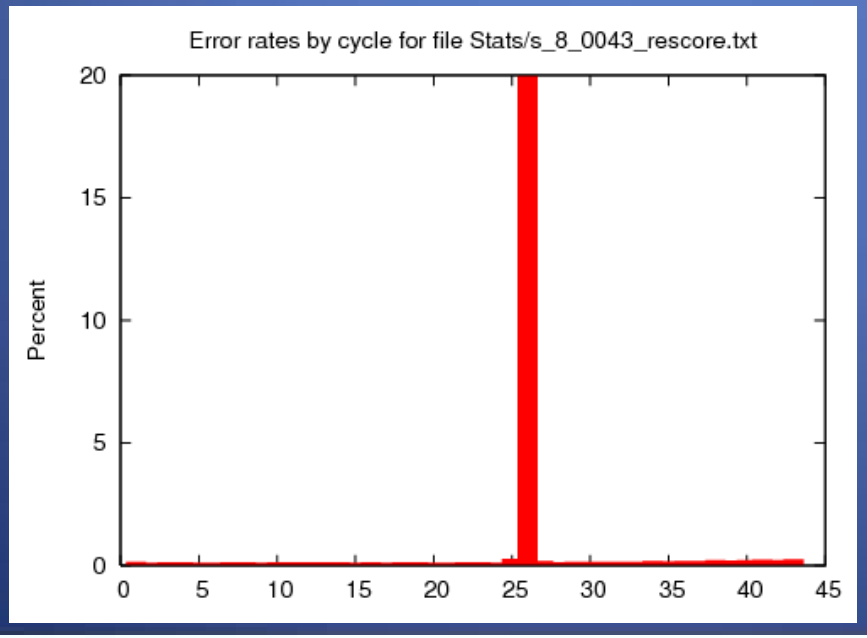
IVC plots

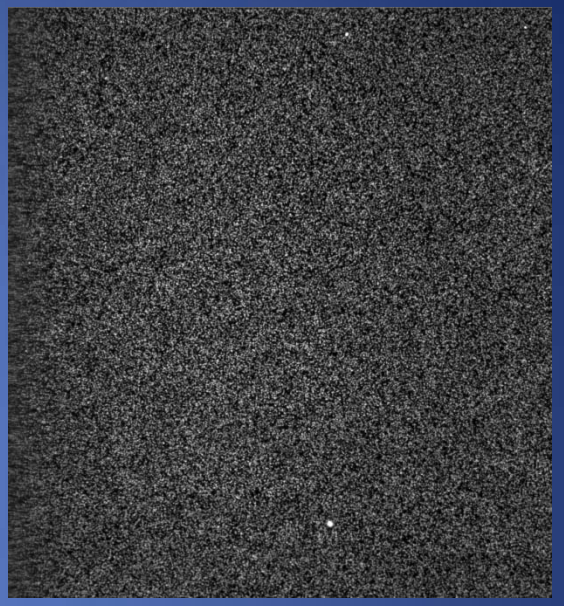
Lane	All	Called	% Base Calls	% All	% Called
1					
2					
3					
4					
5					
6					
7					
8					

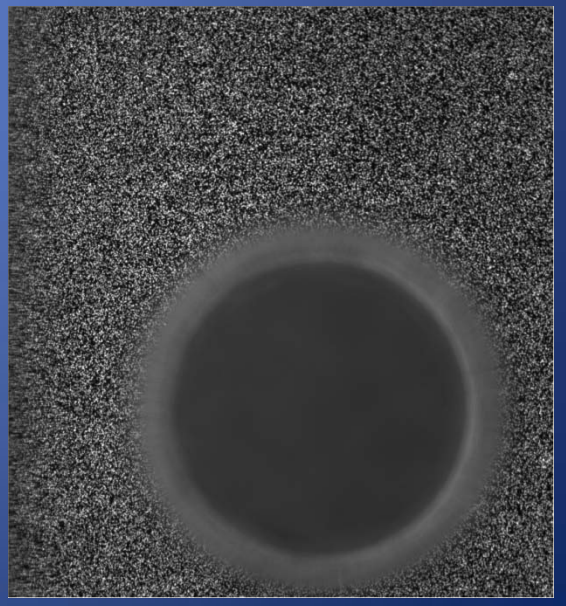
Illumina: Substitution Errors











Illumina: Data File

- Text file format (20M sequences, 43 bp = 3 Gb):

```
@HWI-EAS339_0001:3:1:1056:12207#0/1
CATTCTTCTCATGCATGTGGGCATCAGAGTCGTATGCCGTCTT
+HWI-EAS339_0001:3:1:1056:12207#0/1
bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbabbab\ba_bX^^b
@HWI-EAS339_0001:3:1:1058:3222#0/1
TGAATCGGAAGCGGGGCACTGATCTCGTATGCCGTCTTCTGCT
+HWI-EAS339_0001:3:1:1058:3222#0/1
bbbbbcbbabbcb``b\b^bbabZbabb^bb\Y`^^_bX\^R_
```

Sequence name
Sequence
Sequence name (again)
Quality scores (ASCII + 64)

Illumina distribution: bioHPC

Sequencing results for sample "RDTtest384"

Parameter	Value
Run Name:	100223_HWI-EAS83_0005_6148NAAXX
Lane#:	5
Analysis Software:	RTA 1.6.32.0
Sample Name:	RDTtest384
Lane Annotations:	Parameter / This Lane / Ctrl Lane Length / unknown / unknown Clusters_raw / 16.2M / 26.0M Clusters_PF / 13.5M / 10.5M
Order#:	10213178
Expiration Date:	4/1/2010

Files will be available for download until 4/1/2010 (26 days left)

File (click to download)	Size [bytes]	MD5 sum
10213178_6148NAAXX_s_5_sequence.txt.gz	516,435,051	c0d894bdb2caa61ec0b450c64fe5af1e

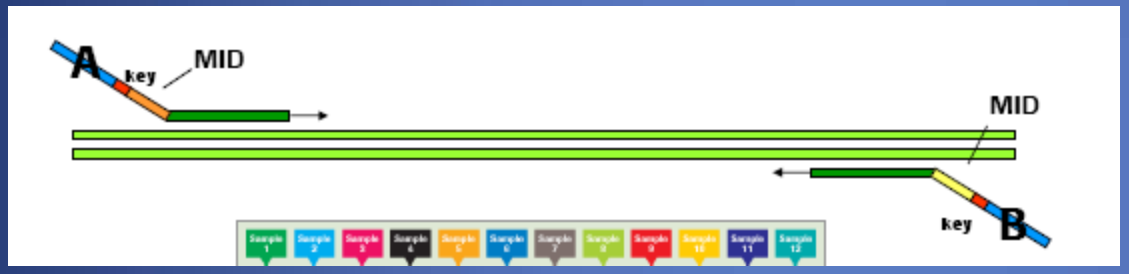
Prefer to download multiple files in batch mode?
Use **Lane Browser** at BioHPC Next Generation Data Anlysis site to generate a download script.

Major Applications for Illumina Genome Analyzer Sequencing

- Primarily, sequencing where reference genome is available
  - Genomic re-sequencing
  - Transcriptome analysis
  - SNP discovery
  - SNP genotyping
  - ChIP
  - miRNA
- Some *de novo* applications

Multiplexing (barcoding) Options

- Both platforms can utilize barcodes/indexes/MIDs to pool samples in a single lane or region.



Name	Sequence
RL1 *	ACACGACGACT
RL2 *	ACACGTAGTAT
RL3 *	ACACTACTCGT
RL4 *	ACGACACGTAT
RL5 *	ACGAGTAGACT
RL6 *	ACGCGTCTAGT
RL7 *	ACGTACACACT
RL8 *	ACGTACTGTGT
RL9 *	ACGTAGATCGT
RL10 *	ACTACGTCTCT
RL11 *	ACTATACGAGT
RL12 *	ACTCGCGTCGT

Incorporating barcodes into adaptors (Illumina shown):

ACACTCTTTCCCTACACGACGCTCTTCCGATCT P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

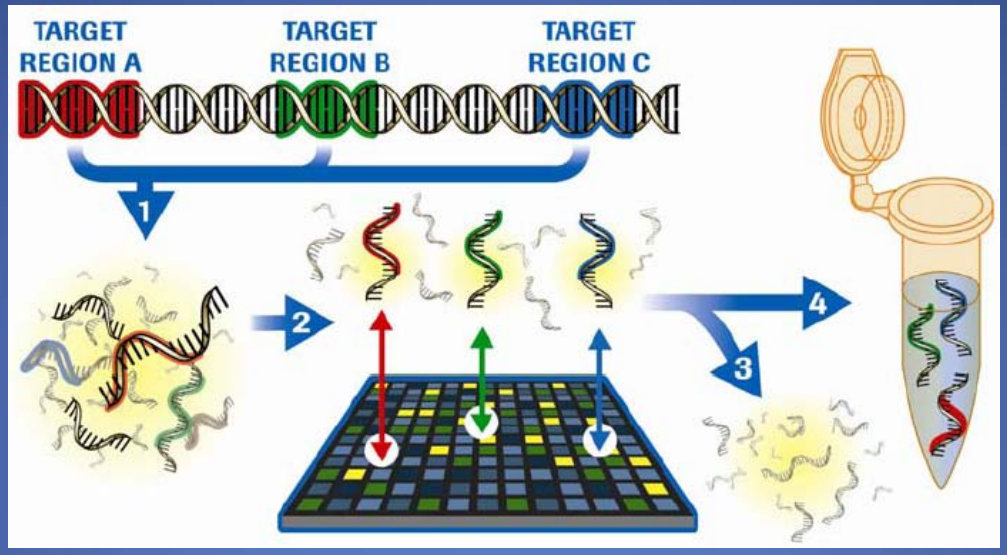
ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGT P-CGATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

ACACTCTTTCCCTACACGACGCTCTTCCGATCT (sequencing primer)

Targeted Resequencing
(for sequencing exons only or chromosomal region)

- Micorarray capture (Nimblegen, Agilent, Febit)
- Bead-based capture (Agilent)
- PCR-based (Raindance Technologies)

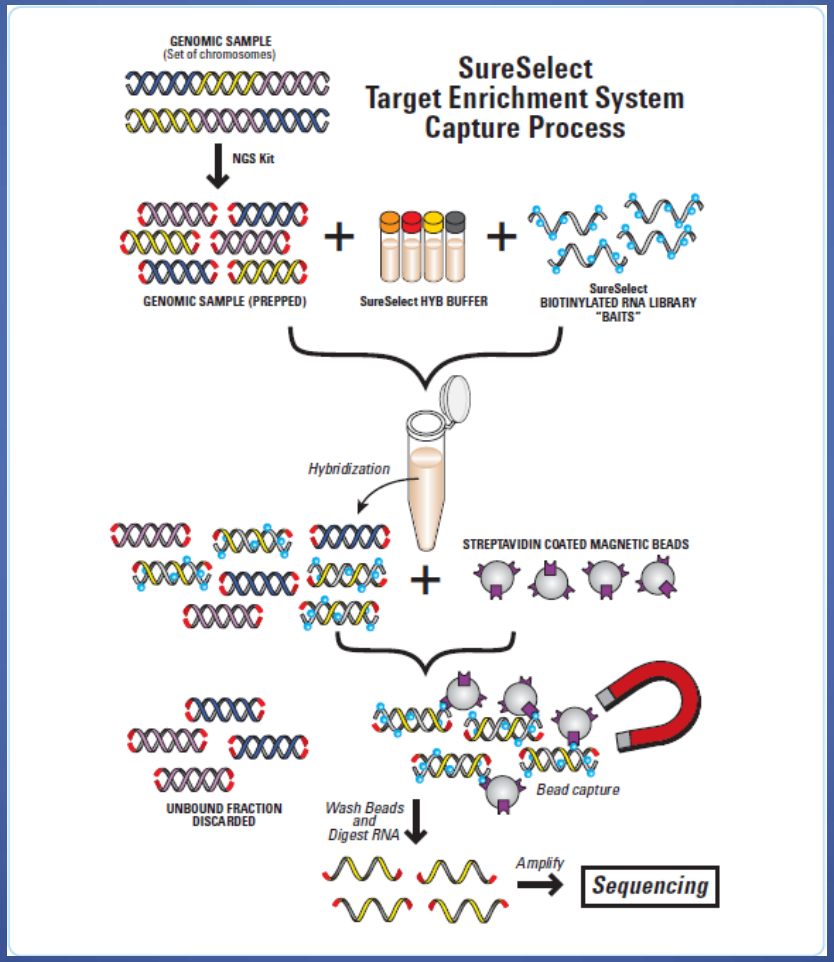
Microarray (e.g. Nimblegen)



- Customized genomic enrichment:
  - 385K custom array, which captures up to 5MB of sequence.
  - 2.1M custom array, which captures up to 30MB of sequence

- 2.1M Human Exome Array:
  - captures the entire human exome (180,000 exons) on a single array

Bead based: Agilent SureSelect



PCR amplification: Raindance Technologies

- Raindance microfluidic PCR
  - Up to 4,000 PCR primer pairs, in uniplex reactions