



# Ion Torrent - The First Six Months

Speed, Simplicity, Scalability, and Accuracy – 30 June 2011

For research use only. Not intended for any animal or human therapeutic or diagnostic use.

**ion torrent**  
△ ★ △ ○ × □ + ≈

by *life* technologies™

# Ion PGM™ Decodes Killer *E. coli* Outbreak in Germany

HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR TIMES TOPICS SL

**The New York Times**

## Europe

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

AFRICA AMERICAS ASIA PACIFIC **EUROPE** MIDDLE EAST



"The PGM takes the shortest time to generate genomic data."

Junjie Qin, scientist who led the *E. coli* sequencing efforts at BGI.

**ion torrent**  
by life technologies

## Virulent *E. coli* Strain Spreads in Germany and Puzzles Health Officials

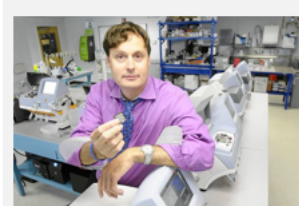
Hartford Courant  
**courant.com** | ARTICLE COLLECTIONS

You are here: Courant.com > Collections > Dna

### High-Speed Connecticut Technology Decodes Deadly *E. coli* Strain

Desktop Sequencer Developed At Ion Torrent In Guilford

June 02, 2011 | By WILLIAM WEIR, [bweir@courant.com](mailto:bweir@courant.com), The Hartford Courant



Richard Messina, Hartford Courant


The *E. coli* bacterium implicated in a deadly outbreak in Europe is a new, "super-toxic" strain never previously seen, but health officials know more about its genetic makeup because of new, high-speed technology developed by a Connecticut company.

"It really is a Frankenstein bacteria," said Jonathan Rothberg, founder of Ion Torrent in Guilford, which developed the sequencing machine that this week decoded the mutant bacterium's DNA.

"It's a hybrid of other killer bacteria, a cross between two deadly strains," Rothberg said.

Scientists at two laboratories in Germany and one in China used the company's new Ion Personal Genome Machine to sequence the genome of the bacterium in a matter

of hours. Prior to the development of the sequencing technology, it would have taken days to decode the DNA.



genomeweb In Sequence  
The Inside Read on Genome Sequencing

Home News Magazine Blogs Careers

Arrays Dx/PGx Informatics PCR Proteomics RNAi/miRNA

Home » News » In Sequence

### Ion Torrent PGM Enables Rapid Identification of New Hybrid *E. coli* Strain in European Outbreak

June 03, 2011



**ion torrent**  
△ ★ △ ○ × □ + ≈

# German *E. coli* Outbreak Strain Identified Using Ion PGM™ in 3 Days

“The biggest advantage [of the PGM] from my point of view as a public health official is that it's speedy, and speed is what is needed at the moment.”

Prof. Dr. Med Dag Harmsen,  
University Hospital Muenster, Germany

“[The PGM] takes the shortest time to generate genomic data.”

Junjie Qin, BGI, China



# Speed: *E. coli* Outbreak Isolate Sequenced Across 2 Continents within 3 Days

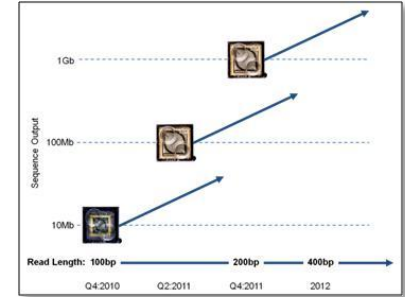
*Time to Results Matter*

**May 30, 2011**  
 O104:H4 and HUSC41 samples (reference) strain libraries prepared  
 O104:H4 amplified and sequenced  
 2 x 2 runs (Ion 314)

**May 31, 2011**  
 O104:H4 sequenced  
 3 x 2 runs (Ion 314)

**June 1, 2011**  
 Data assembled and packaged

Total of 10 runs on Ion 314 (>10Mb/run) or, single run on Ion 316 (>100Mb/run, July release) or, 1/10 of a run on a 318 (>1Gb/run, Q4 release)



June 5

June 12

June 19

**June 2, 2011**  
 Draft genome identified, submitted and released from NCBI

BGI releases assembly to public

Death count rises to 32

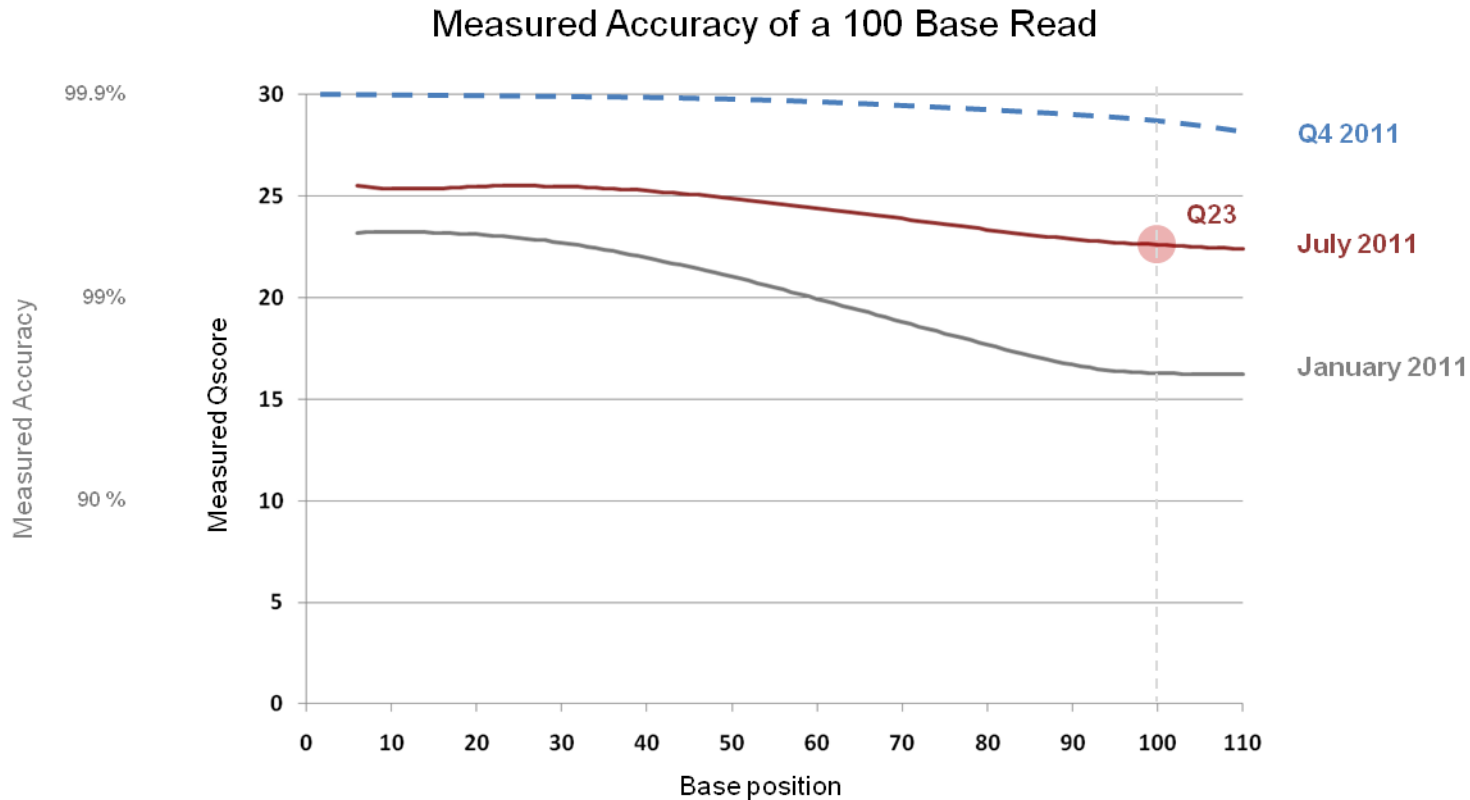
3,048 infected

**June 21, 2011**  
 MiSeq data released

**May 23, 2011**  
 CDC reports HUS increase

# Ion Semiconductor Sequencing Accuracy

## Six Months and Rapidly Improving



Q4 2011 - Long read enzyme and V1.5 software signal processing improvements minimize drop in quality Vs. base position and enable longer accurate reads

July 2011 - Data available at: <http://ioncommunity.iontorrent.com> in [Torrent Dev](#) utilizing Improvements in chemistry and V1.4 software algorithms

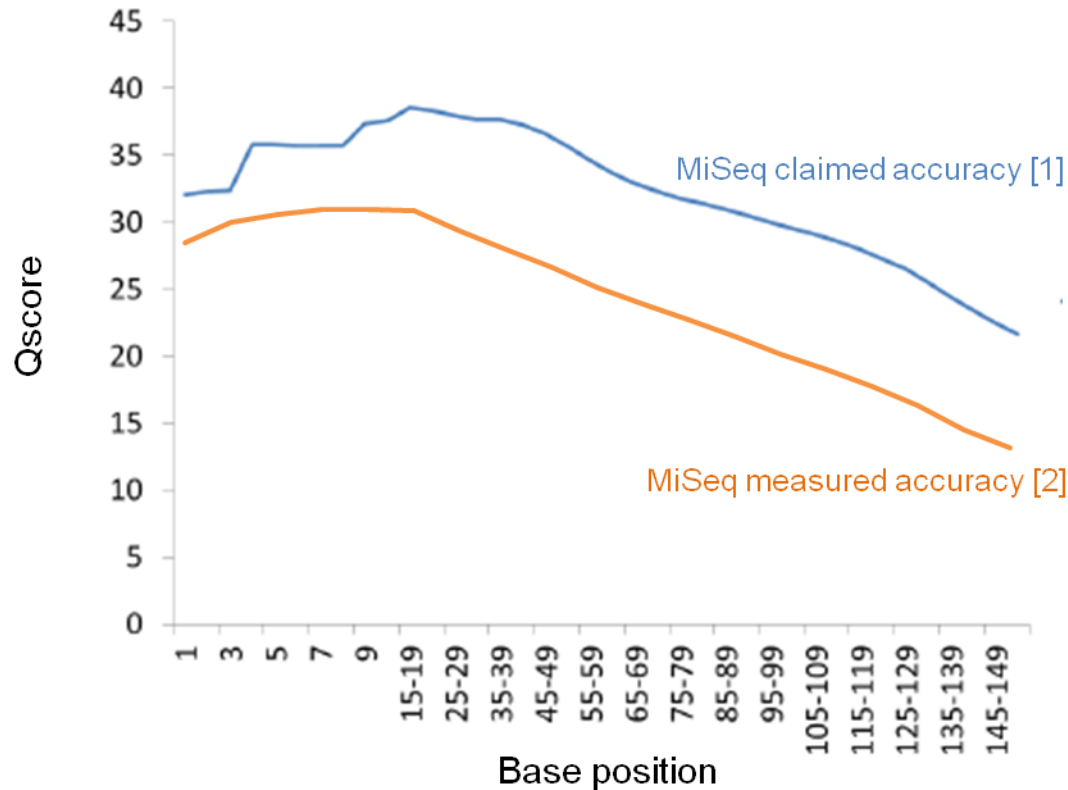
January 2011 - From best Ion internal run

All plots show measured accuracy by alignment to reference genome (*E. coli* DH10b) at each read position.

Durfee, T., et al. (2008). The complete genome sequence of *Escherichia coli* DH10B, 190(7), 2597–2606. doi:10.1128/JB.01695-07.



# MiSeq Claimed vs Measured Accuracy (*E.coli* K-12 MG1655) *Six Years and Still Inflating*

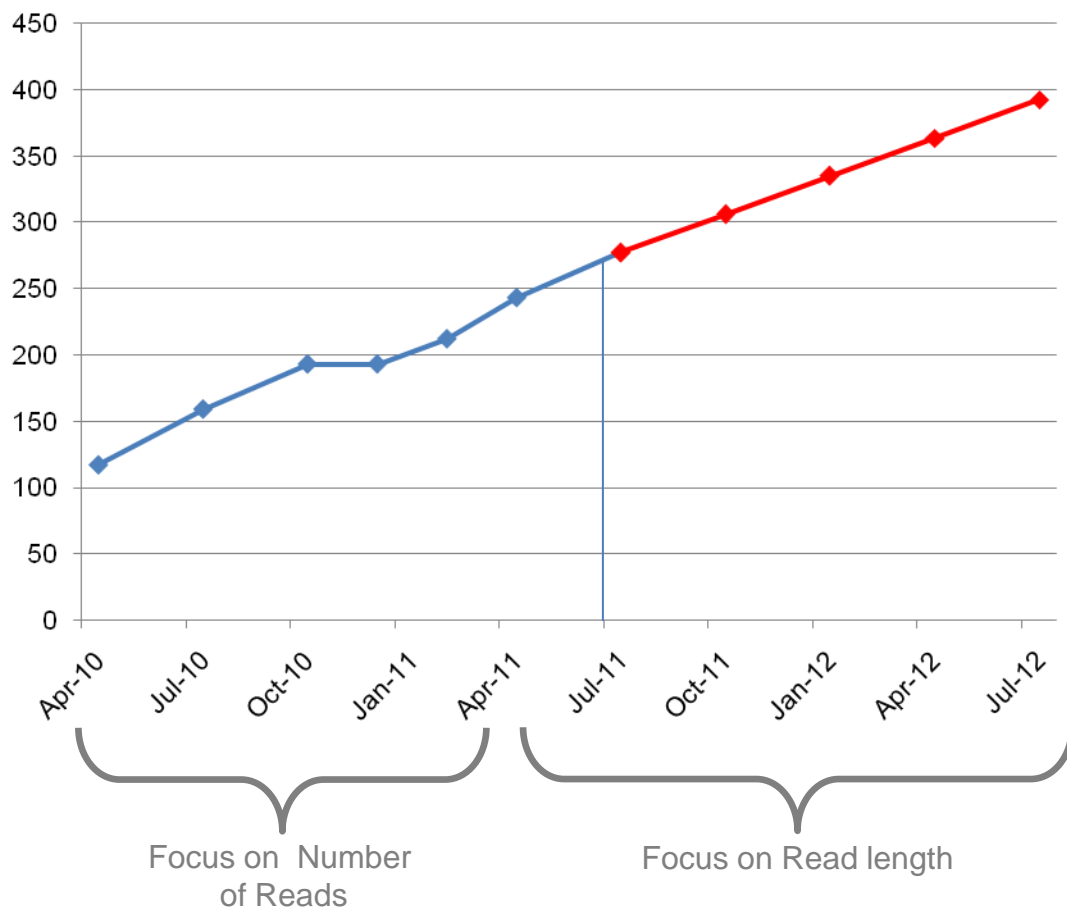


[1] Slide 12 of [http://www.illumina.com/Documents/systems/miseq/MiSeq\\_Ecoli\\_Data\\_June2011.pdf](http://www.illumina.com/Documents/systems/miseq/MiSeq_Ecoli_Data_June2011.pdf)

[2] Measured from [http://www.illumina.com/downloads/Data/SequencingRuns/MiSeq\\_Ecoli\\_MG1655\\_110527.bam](http://www.illumina.com/downloads/Data/SequencingRuns/MiSeq_Ecoli_MG1655_110527.bam) full analysis methods in appendix.

# Ion PGM™ Rapidly Improving Read Length

## Longest Perfect Reads



Rapid R&D Path to 200bp

New sequencing enzyme to give higher signal levels at longer reads and reduced in-well buffering

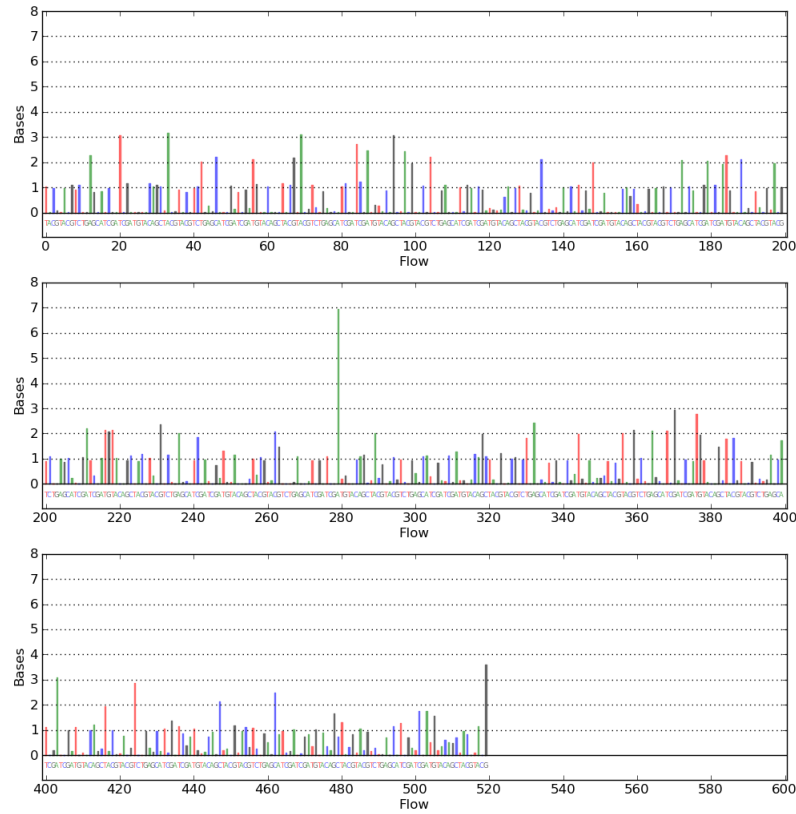
Improved software algorithms for better accuracy (correct background signal and phase errors)

Improved amplification performance to allow longer insert libraries

— Projected Read Length Improvement

# 250 bp Perfect Read

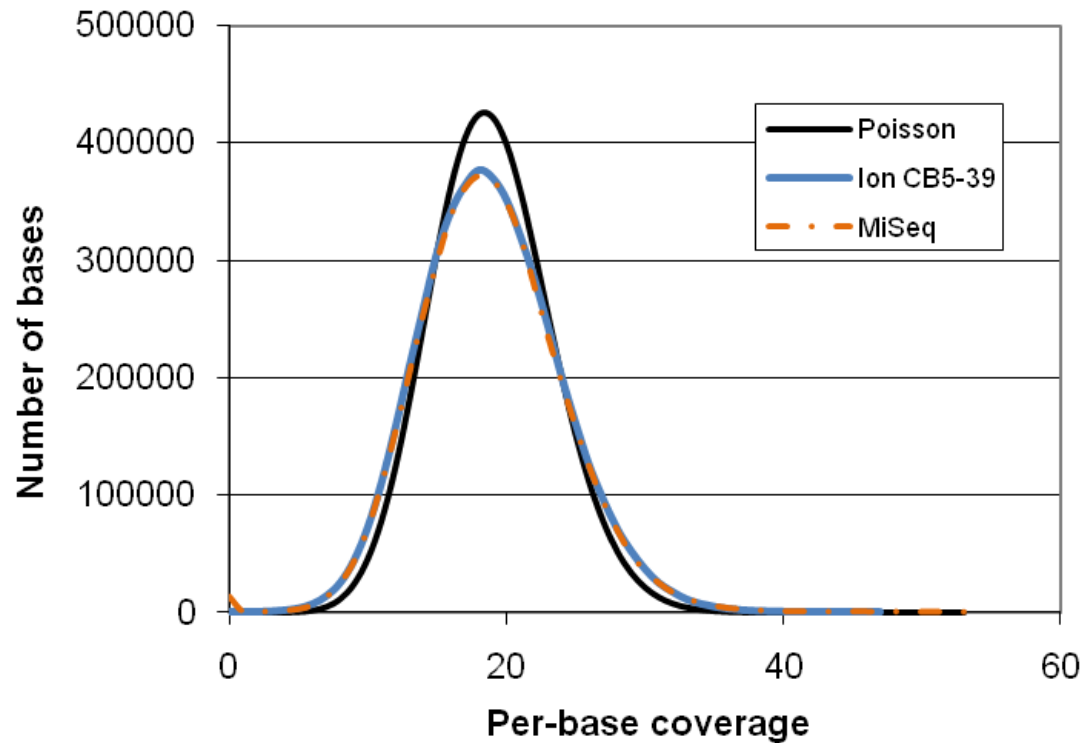
Single-Well Ionogram for (278,1510)



TCAAGACTTTGCAGCAAATCTCTTCCGTGTTCTCGGAAATGCTTTCAACGGGAAGGCTTGATGACGCACTGCCACTGTATACAGCGAGCAACAGAACAATTGCCTGAAAGTCAGCGGAATATTGGTTAGCACTGGCAATCCAGTATCGCCGATGTAAAAAACAAGCTGCAGCACAGGCTGCGCTTAATGCTTATCTTGGCAATTGGGCATTGGTGTTCCTGACAAATAAAGTCATTCTTGGCTGTCA  
 TCAAGACTTTGCAGCAAATCTCTTCCGTGTTCTCGGAAATGCTTTCAACGGGAAGGCTTGATGACGCACTGCCACTGTATACAGCGAGCAACAGAACAATTGCCTGAAAGTCAGCGGAATATTGGTTAGCACTGGCAATCCAGTATCGCCGATGTAAAAAACAAGCTGCAGCACAGGCTGCGCTTAATGCTTATCTTGGCAATTGGGCATTGGTGTTCCTGACAAATAAAGTCATTCTTGGCTGTCA

# Highly Uniform Genome Coverage

*Six Months vs Six Years*



Ion *E. coli* CB5-39 run available at: <http://ioncommunity.iontorrent.com/torrentdev>

MiSeq run downloaded from: [http://www.illumina.com/downloads/Data/SequencingRuns/MiSeq\\_Ecoli\\_MG1655\\_110527.bam](http://www.illumina.com/downloads/Data/SequencingRuns/MiSeq_Ecoli_MG1655_110527.bam) MiSeq content randomly down-sampled to same level of coverage as Ion run.

# Scientists View of Ion Technology



“Baylor researchers, while visiting Ion Torrent, recently broke the company's internal record for the highest output on the 316 chip, .....Since then, Baylor has had runs over 200 megabases at its own center, too, Muzny said, .... it tested it on six microbial genomes with different GC content, using the 314 chip, and found that its performance was "very comparable to other platforms." The error rate at that time was about 1.2 percent.”

Donna Muzny, Director of Sequencing Operations. Baylor Human Genome Sequencing Center



“The Broad Institute, ....., has also switched over to the 316 chip, getting "very good yields" from it, with some runs exceeding 250 megabases.... The institute has also beaten Ion Torrent's own best runs, he said.....The quality of the PGM data is "good enough," he said, with an error rate of between 1 and 2 percent, depending on how it is measured”

Chad Nusbaum, co-director of the Broad's genome sequencing and analysis program.



“...reflecting a substitution error rate of 0.092%...” [=99.9% accuracy]

“...reflecting an indel error rate of 0.726%...” [=99.3% accuracy]

Dan Koboldt - <http://www.massgenomics.org/2011/06/first-look-data-from-iontorrents-316-chip.html>



“The 316 chip generated 7x more data than the 314 chip (175Mb compared to an average of 24Mb on our 314 chips.) for an average coverage of 35X for E. coli – right in the sweet spot for *de novo* assembly, or mapping and variant calling.”

Justin Johnson- <http://www.edgebio.com/blog/>



# Appendix: Analysis Method

## *MiSeq Measured Accuracy*

### Step 1. Download the bam file from Illumina website - <http://www.illumina.com/systems/miseq/ecoli.ilmn>

As of 6/28 the posted BAM file has the following characteristics

File name: MiSeq\_Ecoli\_MG1655\_110527.bam

File size in bytes: 1,295,172,173

Md5 checksum: b0a12bdad62e7162d132bc0ae267def6

Number of alignments: 11,648,093

### Step 2. Reverse the CIGAR strings on reverse strand hits and regenerate the bam file. The CIGAR strings on the reverse strand hits are reversed in the bam file provided by Illumina.

Command: samtools view -h input.bam | fix.pl | samtools fillmd -bS - reference.fasta > fixed.bam

fix.pl

```
#!/usr/bin/perl
while(<>)
{split/\t/;
if (!/^@/ && $_[1]&0x10)
{
@md=$_[5]=~/(\d+\D+)/g;
$_[5]=join " ",reverse @md;
$_=join"t",@_;
}
print;
}
```

reference.fasta can be obtained here: [http://www.ncbi.nlm.nih.gov/nucore/NC\\_000913?](http://www.ncbi.nlm.nih.gov/nucore/NC_000913?)

### Step 3. Download GATK from the Broad. We recommend downloading the latest version from the git repository and building it.

[http://www.broadinstitute.org/gsa/wiki/index.php/Downloading\\_the\\_GATK](http://www.broadinstitute.org/gsa/wiki/index.php/Downloading_the_GATK)

### Step 4. Run GATK CountCovariates to generate the recalibration data.

```
java -Xmx3g -jar GenomeAnalysisTK.jar -I INFO -R reference.fasta -I input.bam -T CountCovariates -cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov PositionCovariate -recalFile output.csv --run_without_dbsnp_potentially_ruining_quality
```

The `--run_without_dbsnp_potentially_ruining_quality` is used because the K12 data is not expected to have many variants.

The jar file is part of the GATK pipeline.

The recalibration is being performed to calculate the empirical QV values from the number of base calls and the number of errors in the aligned reads in order to compare against the reported QVs.

### Step 5. Run Analyzecovariates to generate per position empirical versus reported QV values found in the \*.PositionCovariate.dat file.

```
java -Xmx3g -jar AnalyzeCovariates.jar -recalFile output.csv -outputDir outdir -resources Sting/R
```

The resources argument points to the R scripts distributed with GATK.

Complete instructions for base quality recalibration are also present on the GATK wiki:

[http://www.broadinstitute.org/gsa/wiki/index.php/Base\\_quality\\_score\\_recalibration](http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration)

# Appendix: Analysis Method

## *Highly Uniform Genome Coverage*

Standard coverage histogram.

1. Ion Torrent reads were mapped using TMAP with default parameters against the reference genome of *E. coli* K12 substr. DH10B.
2. A number of reads covering each base, i.e. per-base coverage, is counted for all bases in the genome.
3. MiSeq data were randomly deprecated to the same mean coverage as ION data.
4. Numbers of bases were plotted against per-base coverage.

# ion torrent



by *life* technologies™