Comparison between MiSeq and HiSeq using Paired Data

Quanhu Sheng, Bojana Jovanovic, Scott Austin Beeler

March 4, 2014

Contents

1 Purpose

The MiSeq and HiSeq sequencing machines have been widely used in next generation sequencing projects. Here, we want to validate the possibility to combine the data from those two machines together for futher analysis.

2 Material and method

2.1 Dataset

Three samples were pair-end sequenced by both MiSeq and HiSeq machines, as table 1 illustrated.

Table 1: Sample Information

Type	Source	Alias
HiSeq	2059-JP-6	HiSeqSample1
HiSeq	2059-JP-7	HiSeqSample2
HiSeq	2059-JP-9	HiSeqSample3
MiSeq	IG-33_408637_S4	MiSeqSample1
MiSeq	IG-34_S3	MiSeqSample2
MiSeq	IG-39_408648_S5	MiSeqSample3

2.2 Raw file quality control

FastQC v0.10.1 was used to evaluate the quality of FastQ files.

```
fastqc -t 2 -o HiSeqSample1 2059-JP-6_1.fastq.gz 2059-JP-6_2.fastq.gz
```

2.3 Terminal 'N' trimming

The terminal 'N' in data was trimmed by in-house software cqstools.

```
mono-sgen CQS.Tools.exe fastq_trimmer -n -z -i 2059-JP-6_1.fastq.gz -o 2059-JP-6_1_trim.fastq.gz
```

2.4 Mapping

TopHat v2.0.9 was used to map reads to human genome 19.

```
tophat2 –segment-length 25 -r 0 -p 8 –keep-fasta-order –no-coverage-search \
    –rg-id HiSeqSample1 –rg-sample HiSeqSample1 –rg-library HiSeqSample1 \
    -G Homo_sapiens.GRCh37.73.gtf –transcriptome-index=hg19_GRCh37_73 \
    -o . bowtie2_index/hg19 2059-JP-6_1.fastq.gz 2059-JP-6_2.fastq.gz

mv accepted_hits.bam HiSeqSample1.bam

mv accepted_hits.bam.bai HiSeqSample1.bam.bai
```

2.5 Counting genes

HTSeq v0.5.4p3 was used to count genes.

```
samtools sort -n -@ 8 HiSeqSample1.bam HiSeqSample1.sortedname

samtools view HiSeqSample1.sortedname.bam | htseq-count -q -m intersection-nonempty \
    -s no -i gene_id - Homo_sapiens.GRCh37.73.gtf >HiSeqSample1.count
```

2.6 FPKM of genes

Cuffdiff v2.1.1 were used to estimate gene FPKM value. cqstools was used to extract FPKM values.

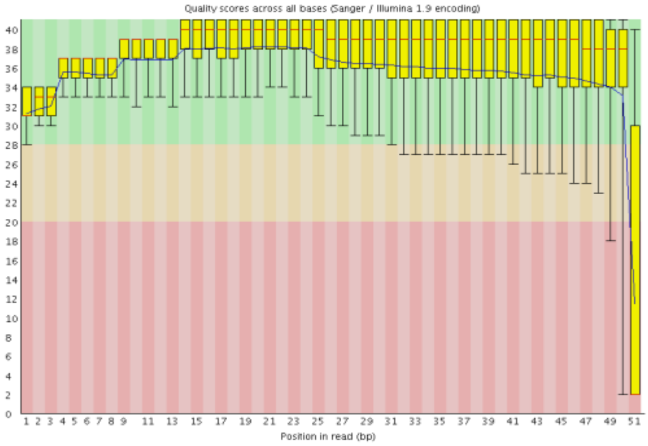
```
cuffdiff -p 8 -u -N -o . -L MiSeq,HiSeq -b hg19_chr.fa Homo_sapiens.GRCh37.73.gtf \
      MiSeqSample1.bam,MiSeqSample2.bam,MiSeqSample3.bam \
      HiSeqSample1.bam,HiSeqSample2.bam,HiSeqSample3.bam

cqstools cuffdiff_count -i genes.read_group_tracking -s gene_exp.diff \
      -m 20140218_bojana_MiSeq_HiSeq_group_sample.map -o untrim_cuffdiff
```

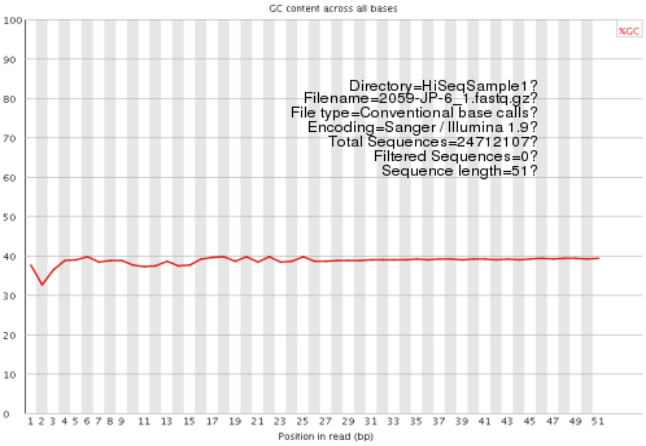
3 Result

3.1 Raw file quality control

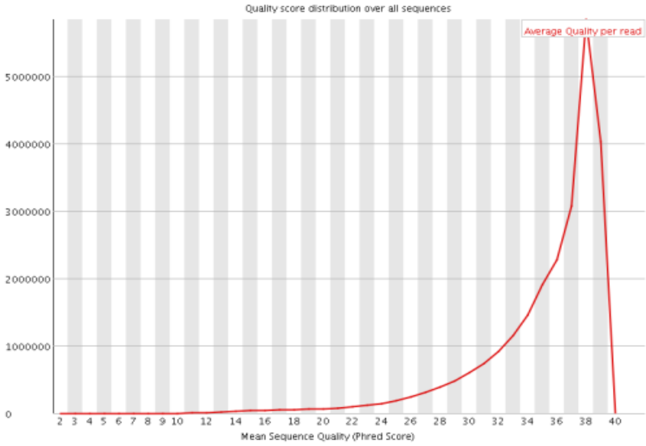
No adapter was found. There are three questionable observations:

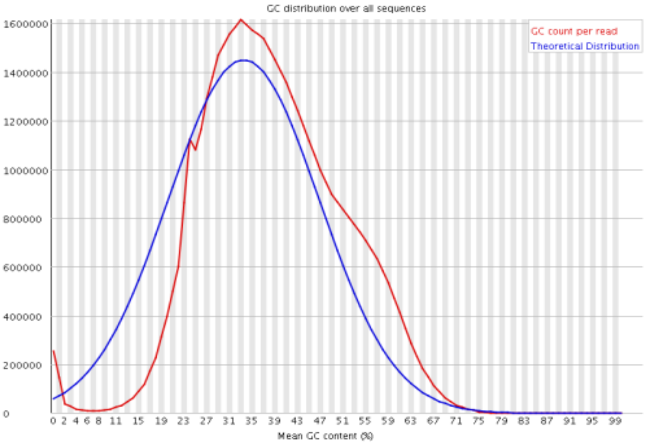
1. The GC contribution of HiSeq data was shifted to 0.4.
2. There is a 'N' in 3' of each HiSeq reads.
3. There are a lot of reads in MiSeq have no GC, which was caused by the reads whose sequence was all 'N'.

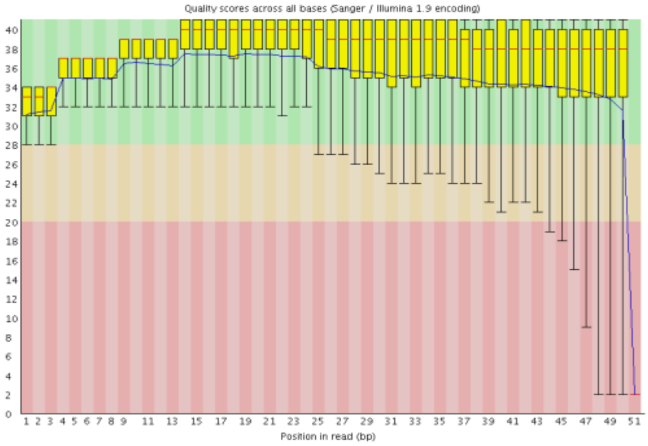


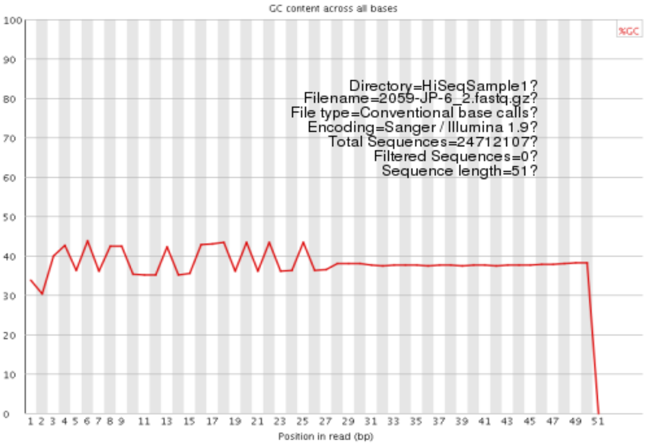


Directory=HiSeqSample1?
Filename=2059-JP-6_1.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=24712107?
Filtered Sequences=0?
Sequence length=51?

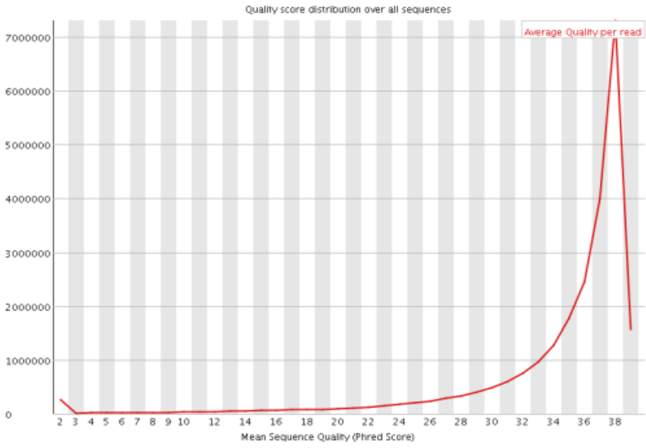


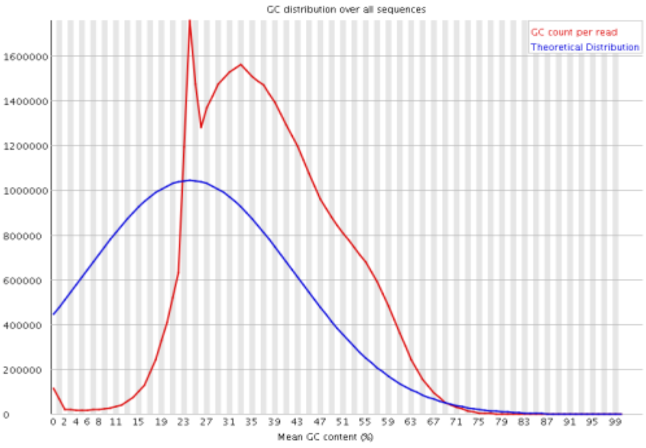


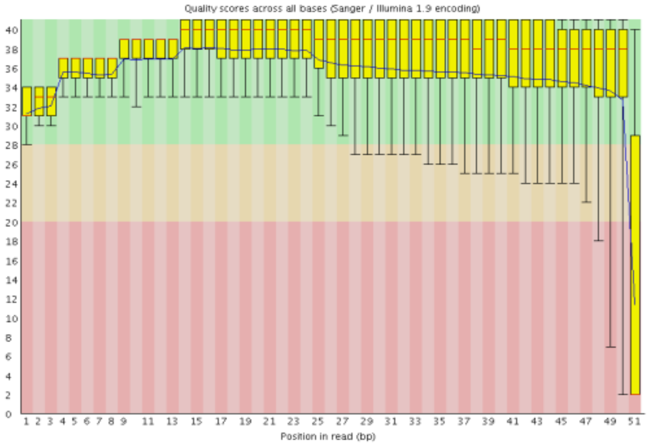


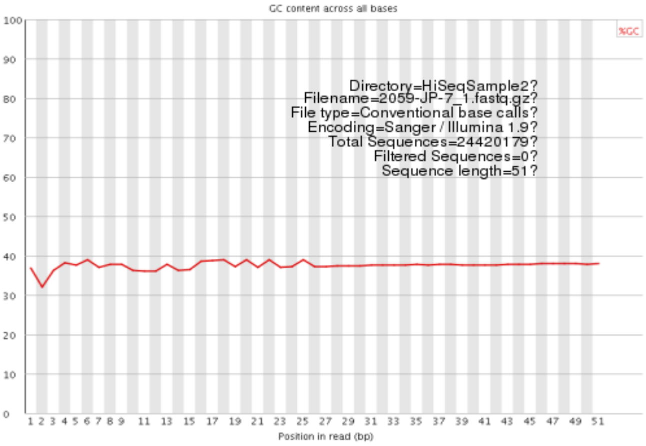


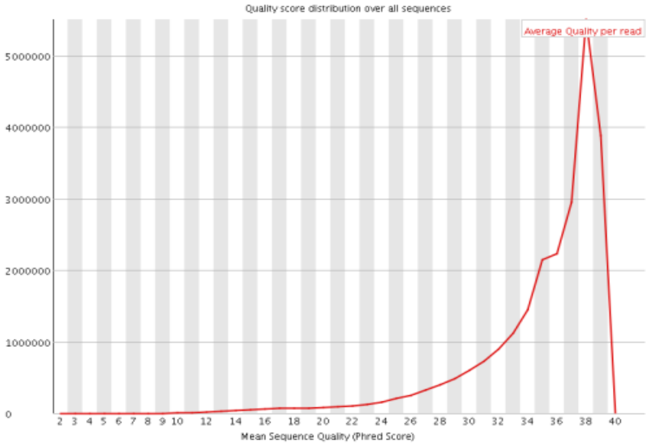
Directory=HiSeqSample1?
Filename=2059-JP-6_2.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=24712107?
Filtered Sequences=0?
Sequence length=51?

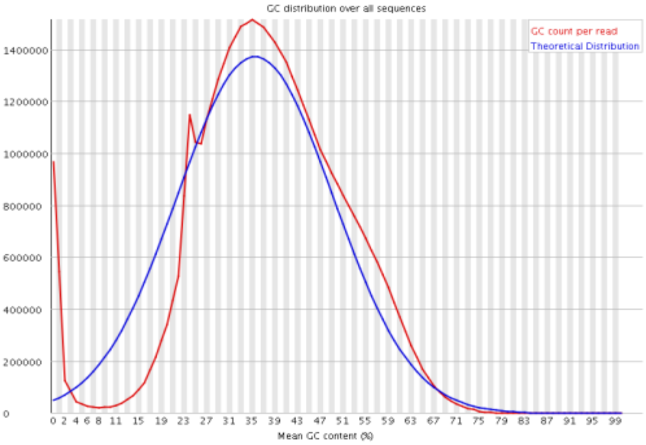


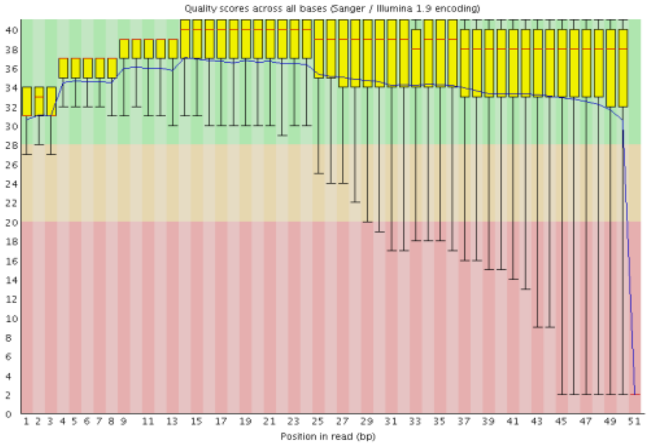


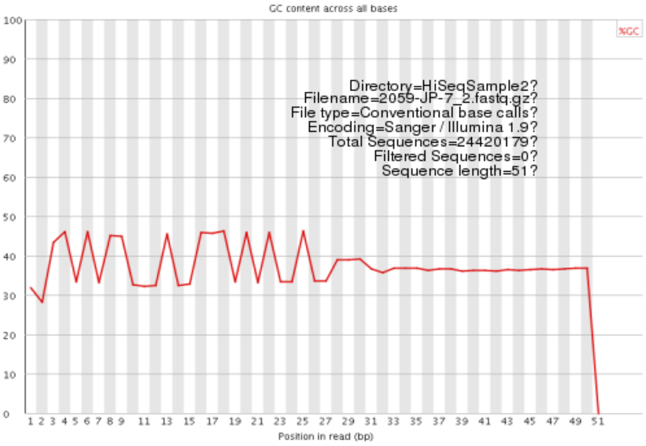

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51
Position in read (bp)

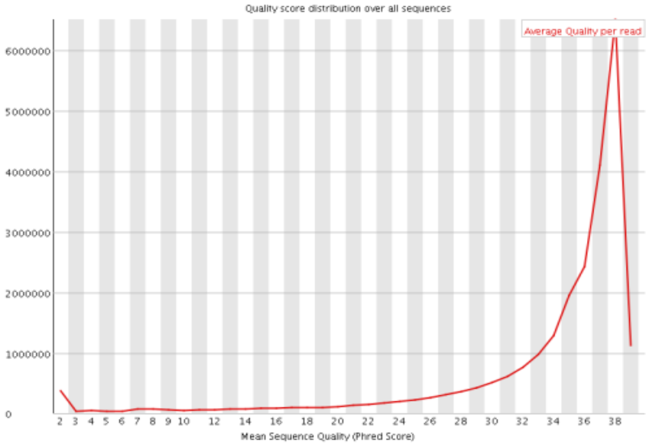

GC content across all bases
%GC
0 10 20 30 40 50 60 70 80 90 100
Directory=HiSeqSample2?
Filename=2059-JP-7_1.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=24420179?
Filtered Sequences=0?
Sequence length=51?
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51
Position in read (bp)

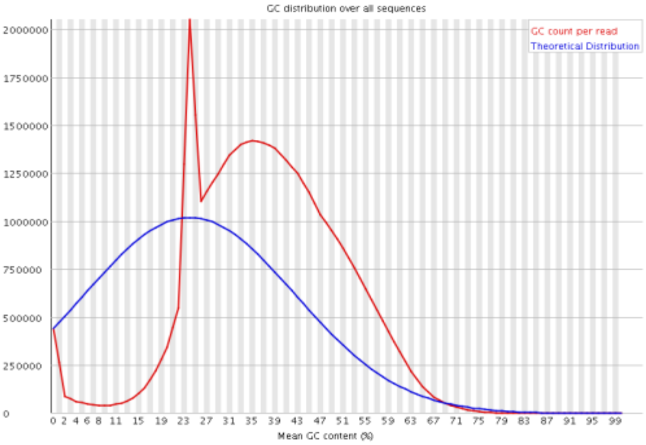

Quality score distribution over all sequences
Average Quality per read
0 1000000 2000000 3000000 4000000 5000000
2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Mean Sequence Quality (Phred Score)

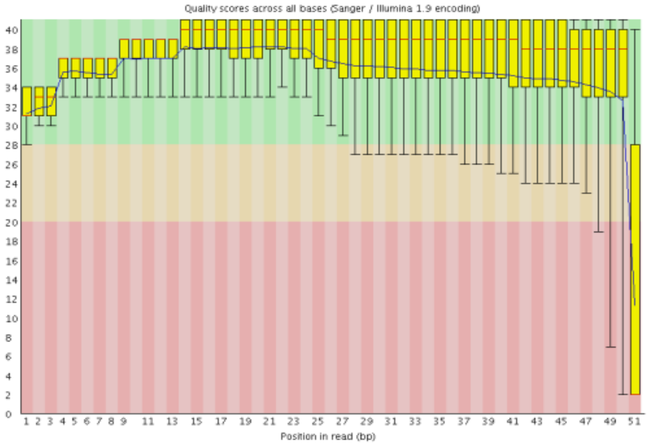

GC distribution over all sequences
GC count per read
Theoretical Distribution
0 200000 400000 600000 800000 1000000 1200000 1400000
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

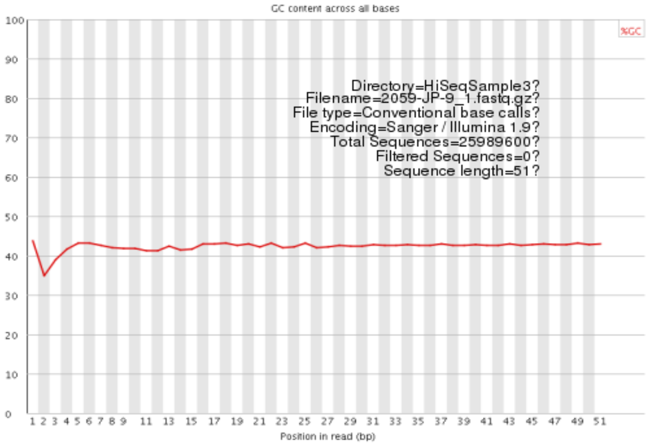

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51
Position in read (bp)


GC content across all bases
%GC
0 10 20 30 40 50 60 70 80 90 100
Directory=HiSeqSample2?
Filename=2059-JP-7_2.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=24420179?
Filtered Sequences=0?
Sequence length=51?
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51
Position in read (bp)

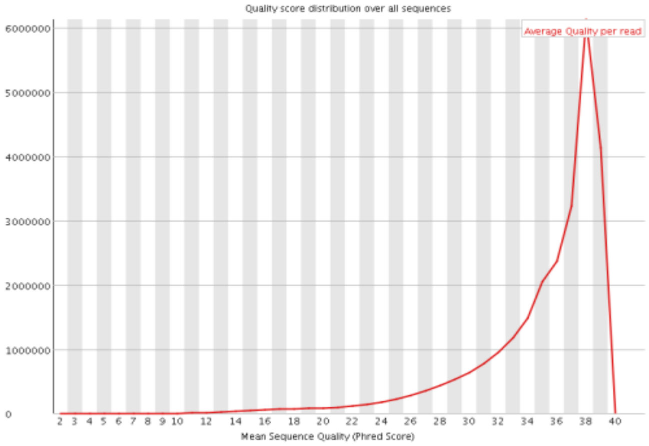

Quality score distribution over all sequences
Average Quality per read
0 1000000 2000000 3000000 4000000 5000000 6000000
2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
Mean Sequence Quality (Phred Score)

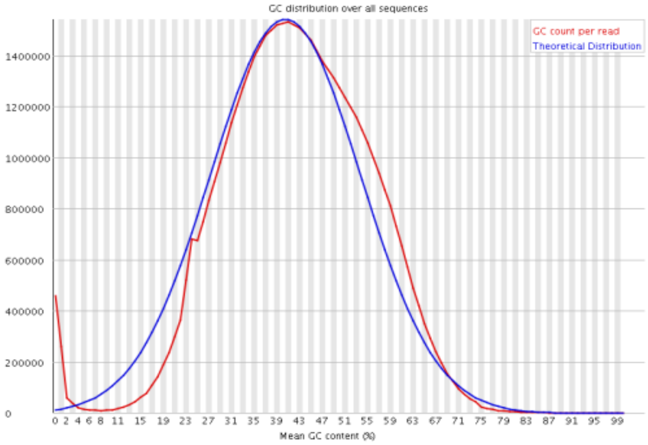

GC distribution over all sequences
GC count per read
Theoretical Distribution
0 250000 500000 750000 1000000 1250000 1500000 1750000 2000000
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

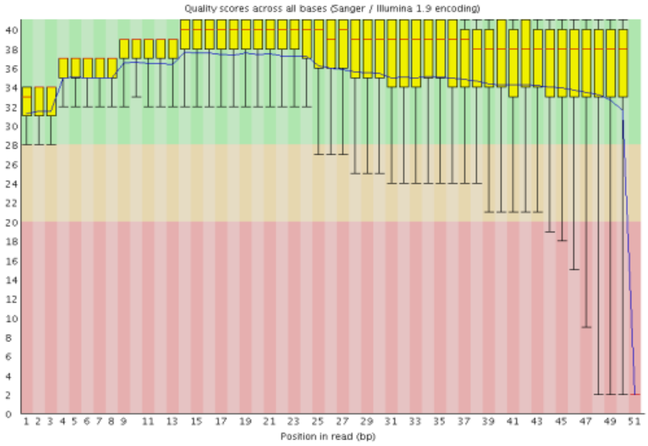


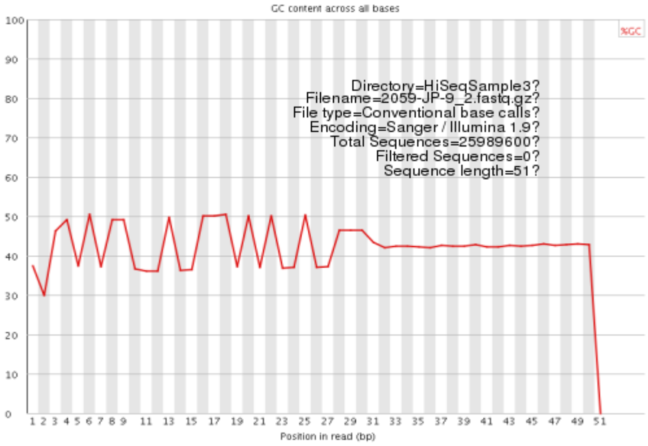


Directory=HiSeqSample3?
Filename=2059-JP-9_1.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=25989600?
Filtered Sequences=0?
Sequence length=51?

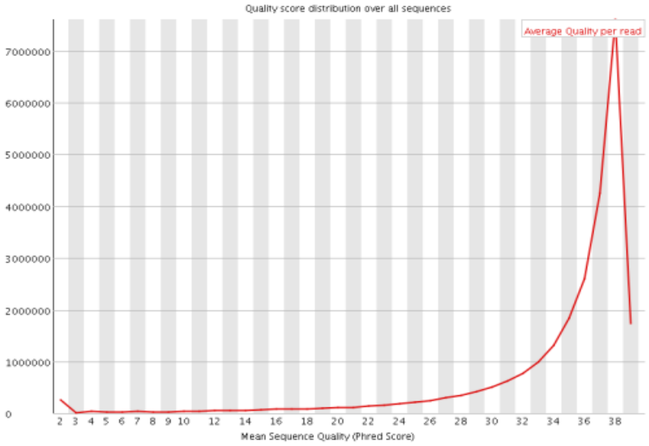


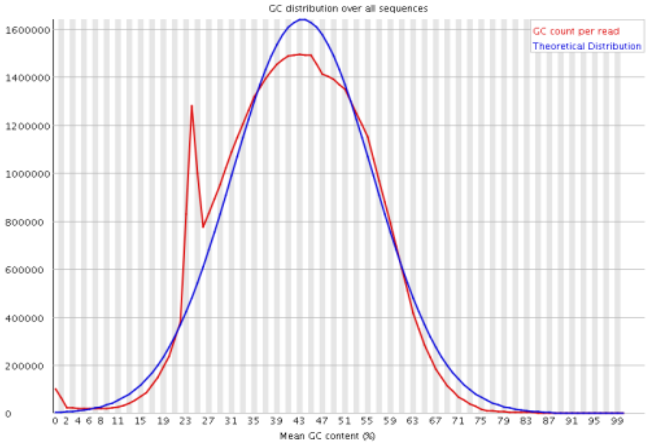


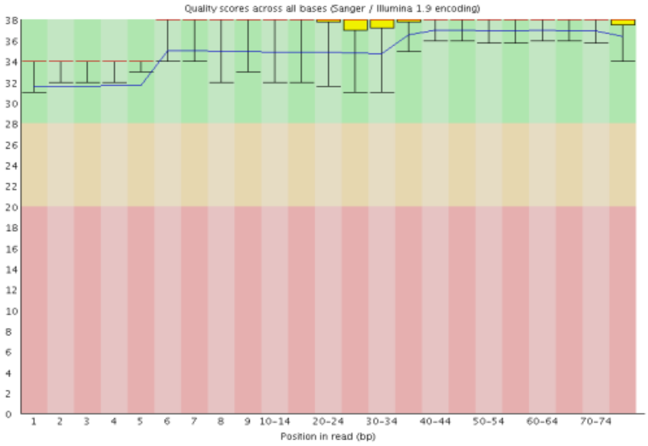




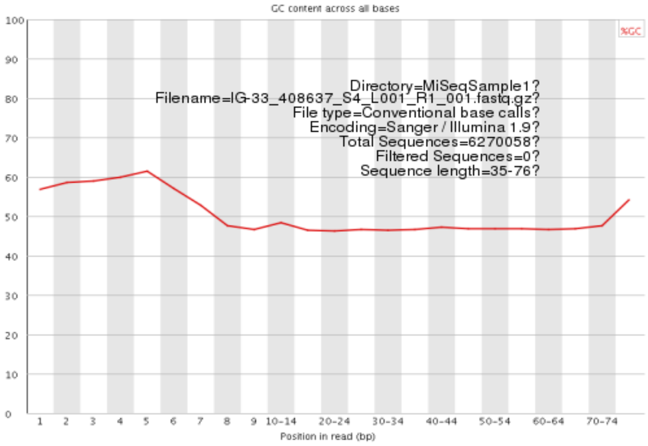
Directory=HiSeqSample3?
Filename=2059-JP-9_2.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=25989600?
Filtered Sequences=0?
Sequence length=51?





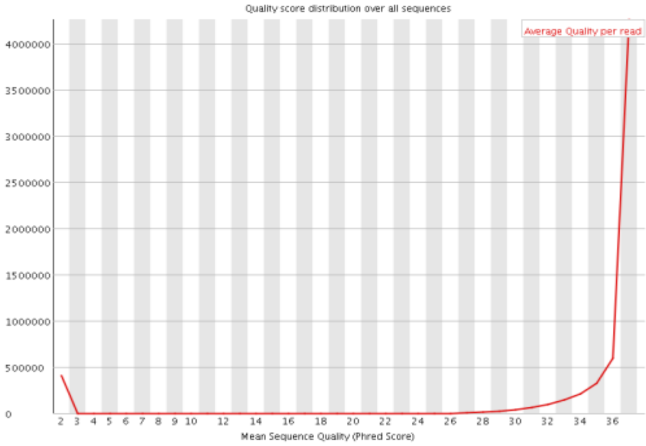


Quality scores across all bases (Sanger / Illumina 1.9 encoding)

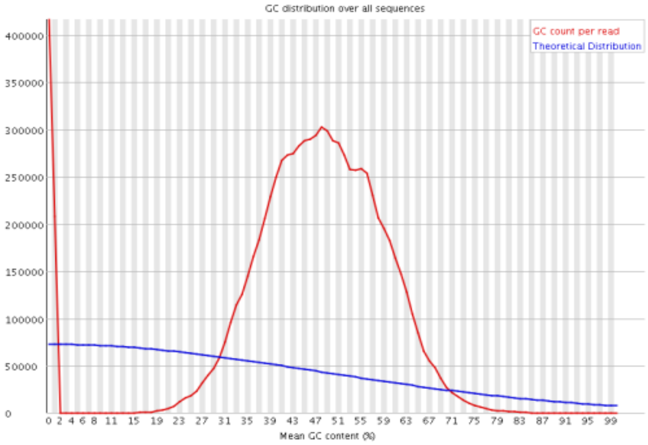


GC content across all bases

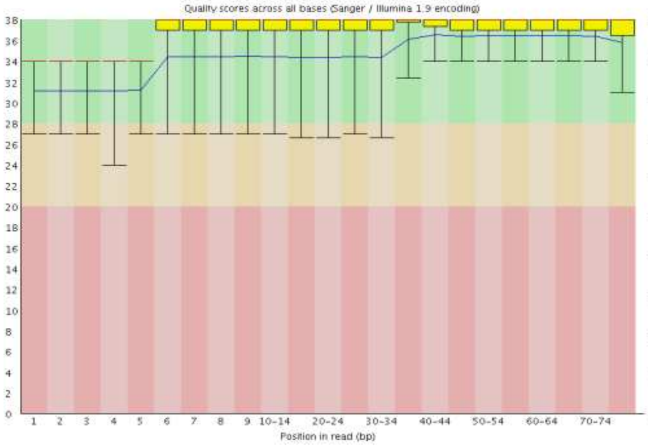
Directory=MiSeqSample1?
Filename=IG-33_408637_S4_L001_R1_001.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=6270058?
Filtered Sequences=0?
Sequence length=35-76?



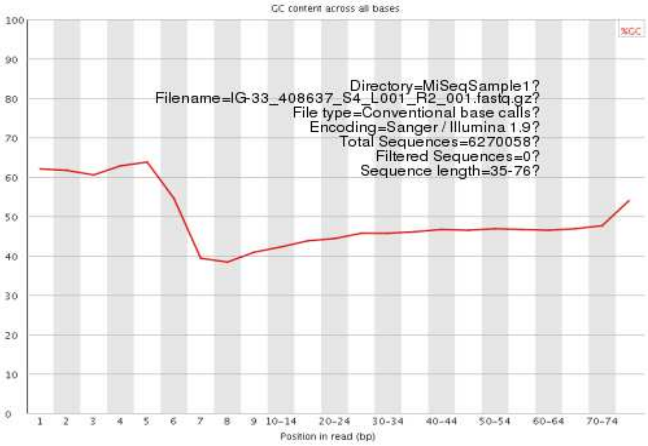
Quality score distribution over all sequences



GC distribution over all sequences

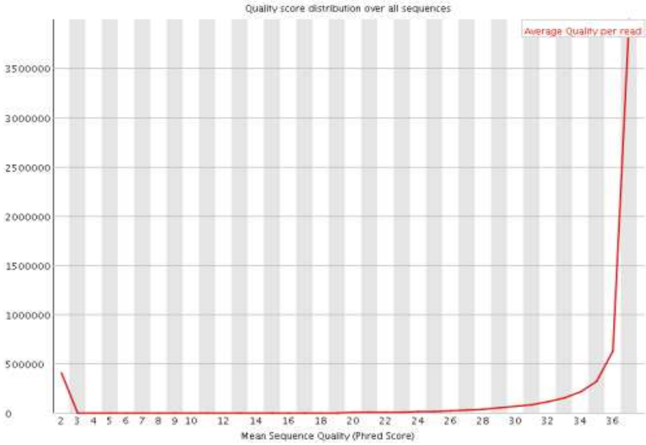


Quality scores across all bases (Sanger / Illumina 1.9 encoding)

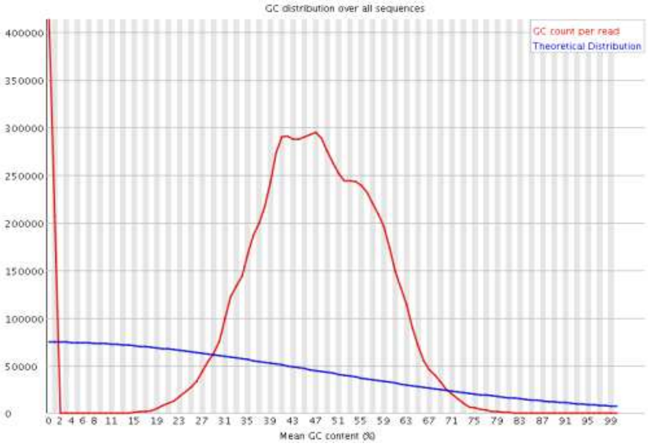


GC content across all bases

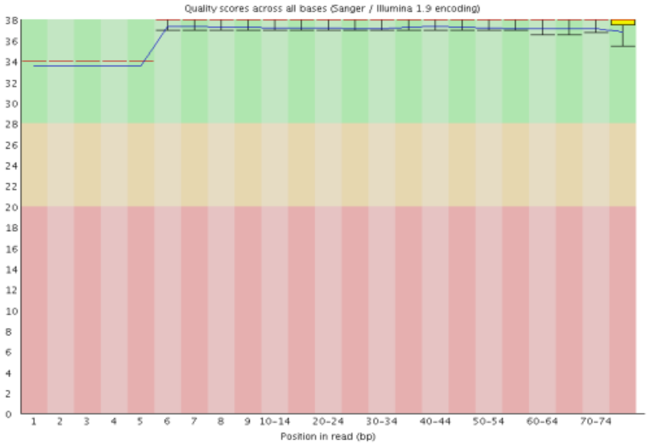
Directory=MiSeqSample1?
Filename=IG-33_408637_S4_L001_R2_001.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=6270058?
Filtered Sequences=0?
Sequence length=35-76?

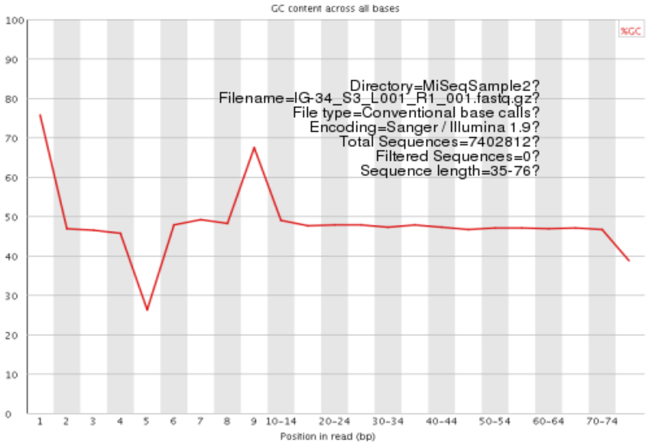


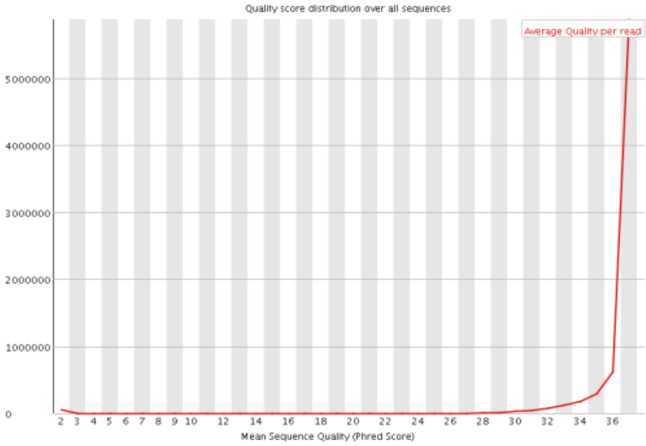
Quality score distribution over all sequences

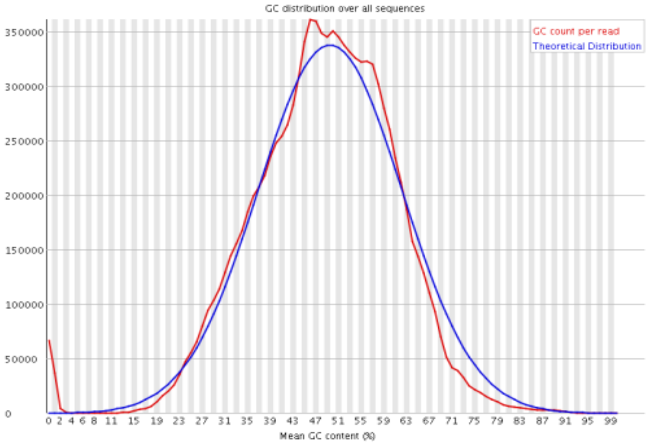


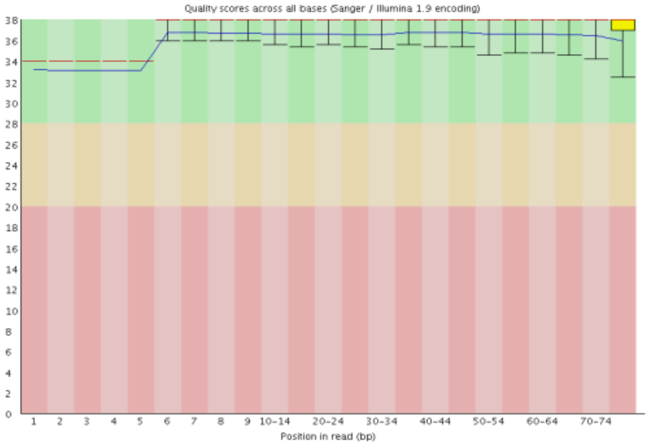
GC distribution over all sequences

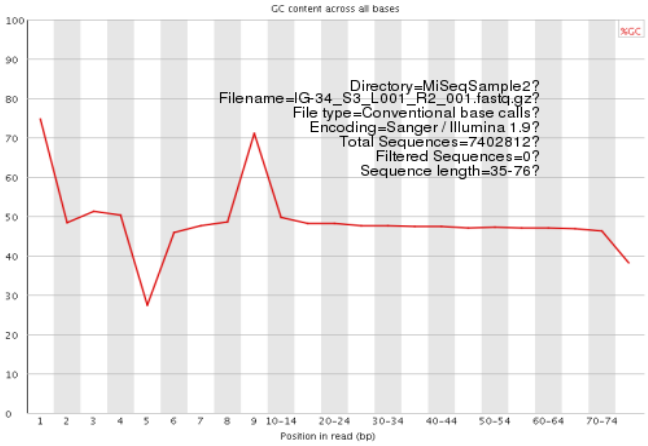

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74
Position in read (bp)

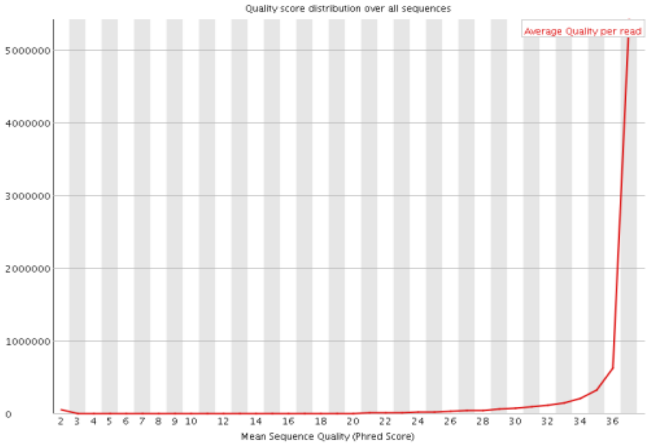

GC content across all bases
%GC
Directory=MiSeqSample2?
Filename=IG-34_S3_L001_R1_001.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=7402812?
Filtered Sequences=0?
Sequence length=35-76?
0 10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74
Position in read (bp)

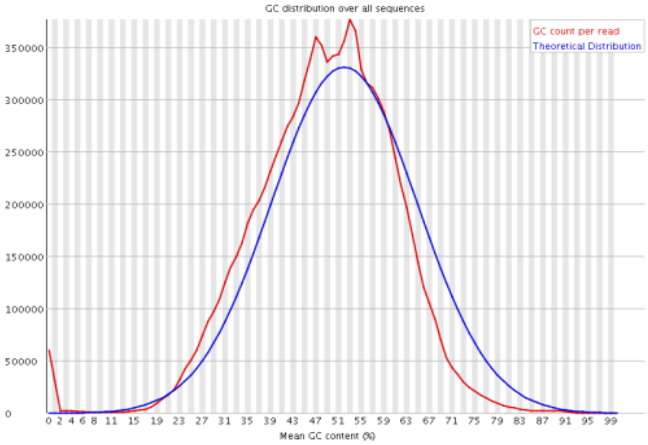

Quality score distribution over all sequences
Average Quality per read
0 1000000 2000000 3000000 4000000 5000000
2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Mean Sequence Quality (Phred Score)

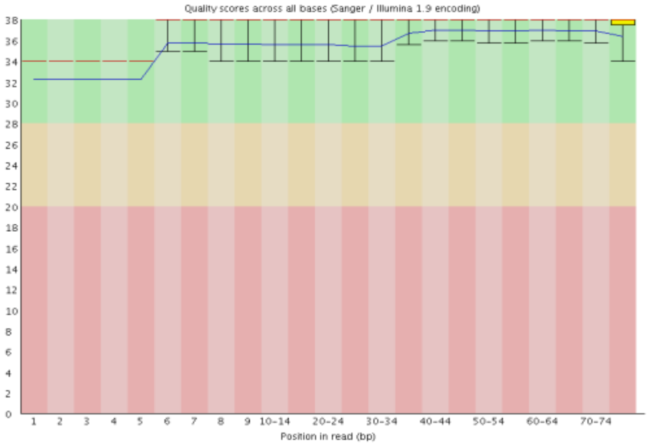

GC distribution over all sequences
GC count per read
Theoretical Distribution
0 50000 100000 150000 200000 250000 300000 350000
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

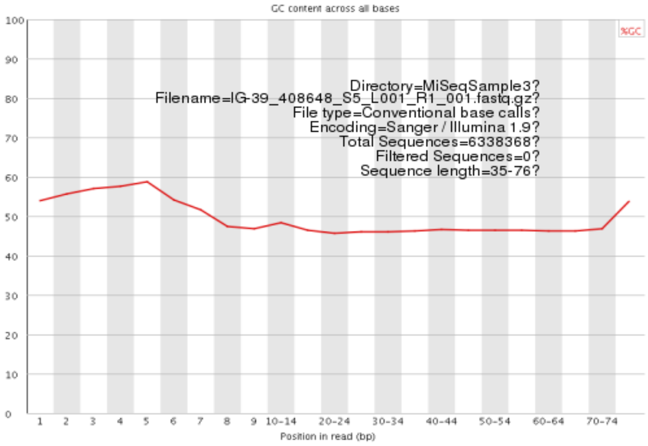

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74
Position in read (bp)

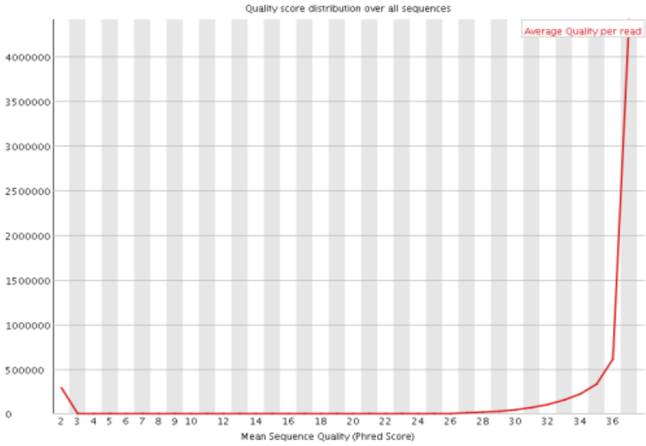

GC content across all bases
%GC
Directory=MiSeqSample2?
Filename=IG-34_S3_L001_R2_001.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=7402812?
Filtered Sequences=0?
Sequence length=35-76?
0 10 20 30 40 50 60 70 80 90 100
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74
Position in read (bp)

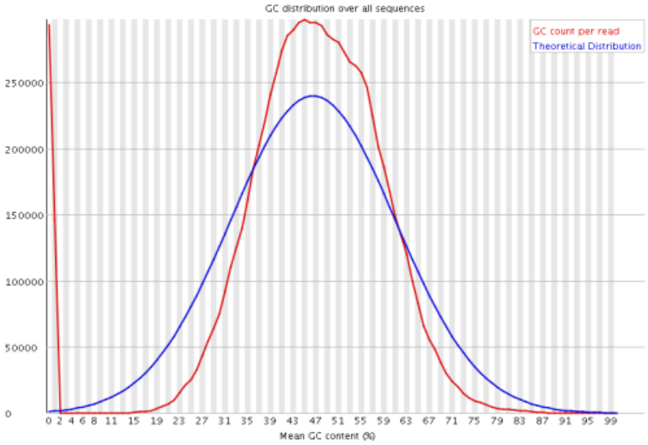

Quality score distribution over all sequences
Average Quality per read
0 1000000 2000000 3000000 4000000 5000000
2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Mean Sequence Quality (Phred Score)

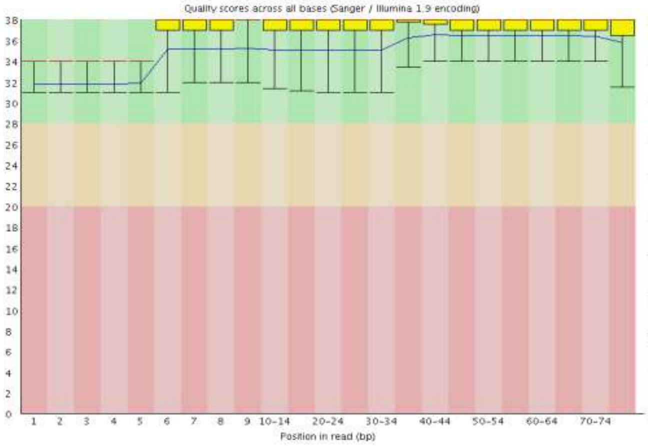

GC distribution over all sequences
GC count per read
Theoretical Distribution
0 50000 100000 150000 200000 250000 300000 350000
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

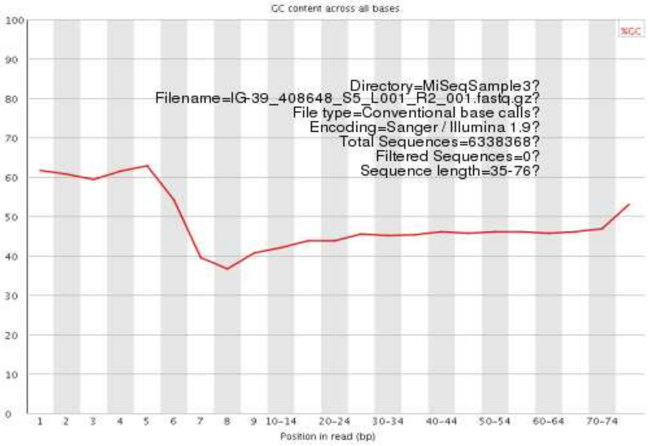

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

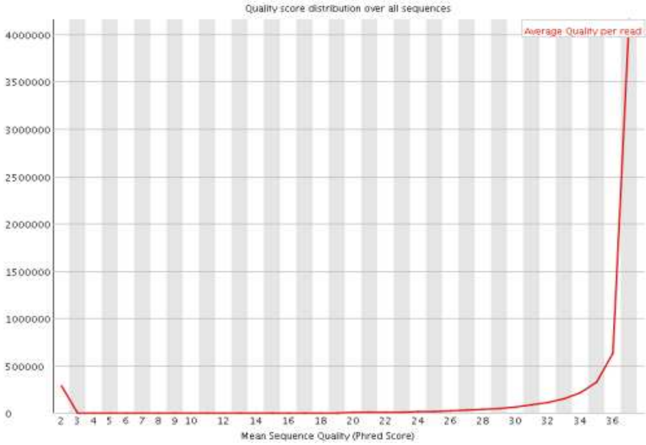

GC content across all bases
%GC
Directory=MiSeqSample3?
Filename=IG-39_408648_S5_L001_R1_001.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=6338368?
Filtered Sequences=0?
Sequence length=35-76?
Position in read (bp)

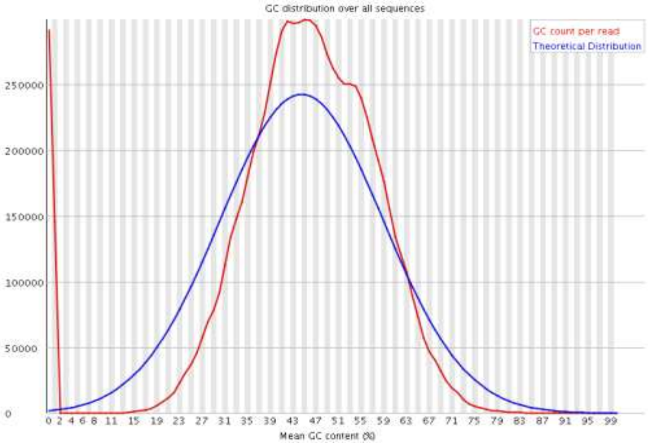

Quality score distribution over all sequences
Average Quality per read
Mean Sequence Quality (Phred Score)


GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


GC content across all bases
%GC
Directory=MiSeqSample3?
Filename=IG-39_408648_S5_L001_R2_001.fastq.gz?
File type=Conventional base calls?
Encoding=Sanger / Illumina 1.9?
Total Sequences=6338368?
Filtered Sequences=0?
Sequence length=35-76?
Position in read (bp)


Quality score distribution over all sequences
Average Quality per read
Mean Sequence Quality (Phred Score)


GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)

3.2 Mapping quality

We used cqstools to trim the terminal 'N' of the reads. Both pretrim reads (table 2) and posttrim reads (table 3) were mapped to genome and summerized.

Table 2: Mapping summary of untrimmed reads

	LeftReads	LMapped	LRate	RightReads	RMapped	RRate	AlignedPairs	ADiscordant
HiSeq1	24712107	21215897	85.9%	24712107	19631934	79.4%	19103150	6.3%
HiSeq2	24420179	19812406	81.1%	24420179	17402588	71.3%	16974487	6.3%
HiSeq3	25989600	22146933	85.2%	25989600	19273082	74.2%	18718125	6.4%
MiSeq1	6270058	4915197	78.4%	6270058	4746878	75.7%	4444182	6.0%
MiSeq2	7402812	6548949	88.5%	7402812	6414037	86.6%	6307380	2.5%
MiSeq3	6338368	5162507	81.4%	6338368	4872895	76.9%	4604894	4.2%

Table 3: Mapping summary of trimmed reads

	LeftReads	LMapped	LRate	RightReads	RMapped	RRate	AlignedPairs	ADiscordant
HiSeq1	24712107	22637599	91.6%	24712107	21039704	85.1%	20492712	8.5%
HiSeq2	24420179	21253887	87.0%	24420179	18784680	76.9%	18344411	8.7%
HiSeq3	25989600	23550896	90.6%	25989600	20593996	79.2%	19994182	7.8%
MiSeq1	5854210	4909295	83.9%	5856888	4731464	80.8%	4101953	98.2%
MiSeq2	7345681	6544854	89.1%	7346956	6405153	87.2%	5787518	98.8%
MiSeq3	6045508	5157791	85.3%	6046582	4859000	80.4%	4271930	98.9%

Trimming terminal 'N' increased the mapping sensitivity of HiSeq data but introduced a lot of discordant mapped pairs in MiSeq data. So, we chose the posttrim HiSeq and pretrim MiSeq mapping result in the following analysis.

3.3 Correlation of count between MiSeq and HiSeq

Spearman correlation was calculated between MiSeq and HiSeq gene count for each sample (table 4).

Table 4: Spearman Correlation of Count between MiSeq and HiSeq

sample	spcorr
Sample1	0.71
Sample2	0.74
Sample3	0.86

3.4 Correlation of FPKM between MiSeq and HiSeq

Spearman correlation was calculated between MiSeq and HiSeq gene FPKM value for each sample (table 5).

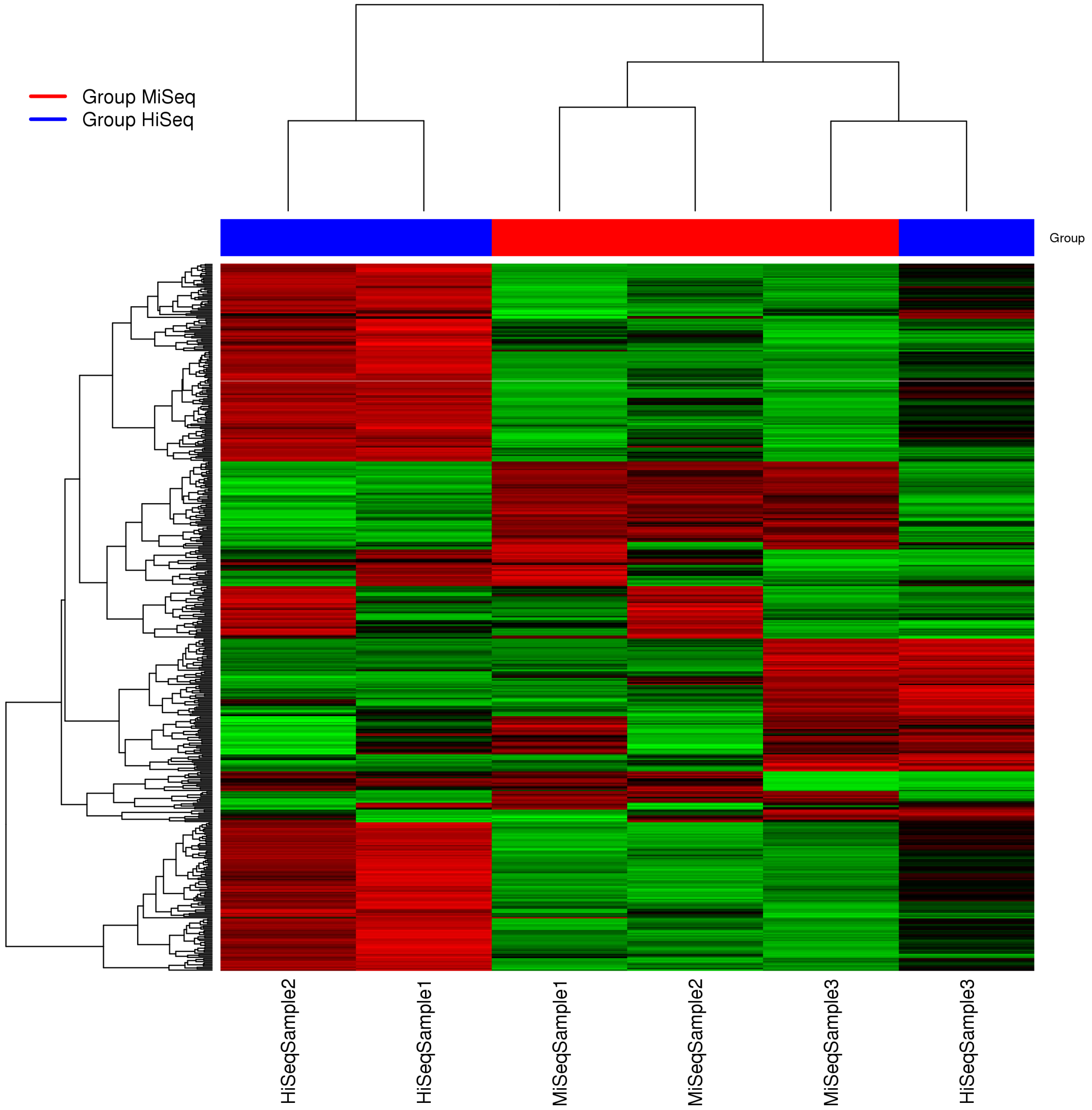
Table 5: Spearman Correlation of FPKM between MiSeq and HiSeq

sample	spcorr
Sample1	0.62
Sample2	0.64
Sample3	0.77

3.5 Heatmap

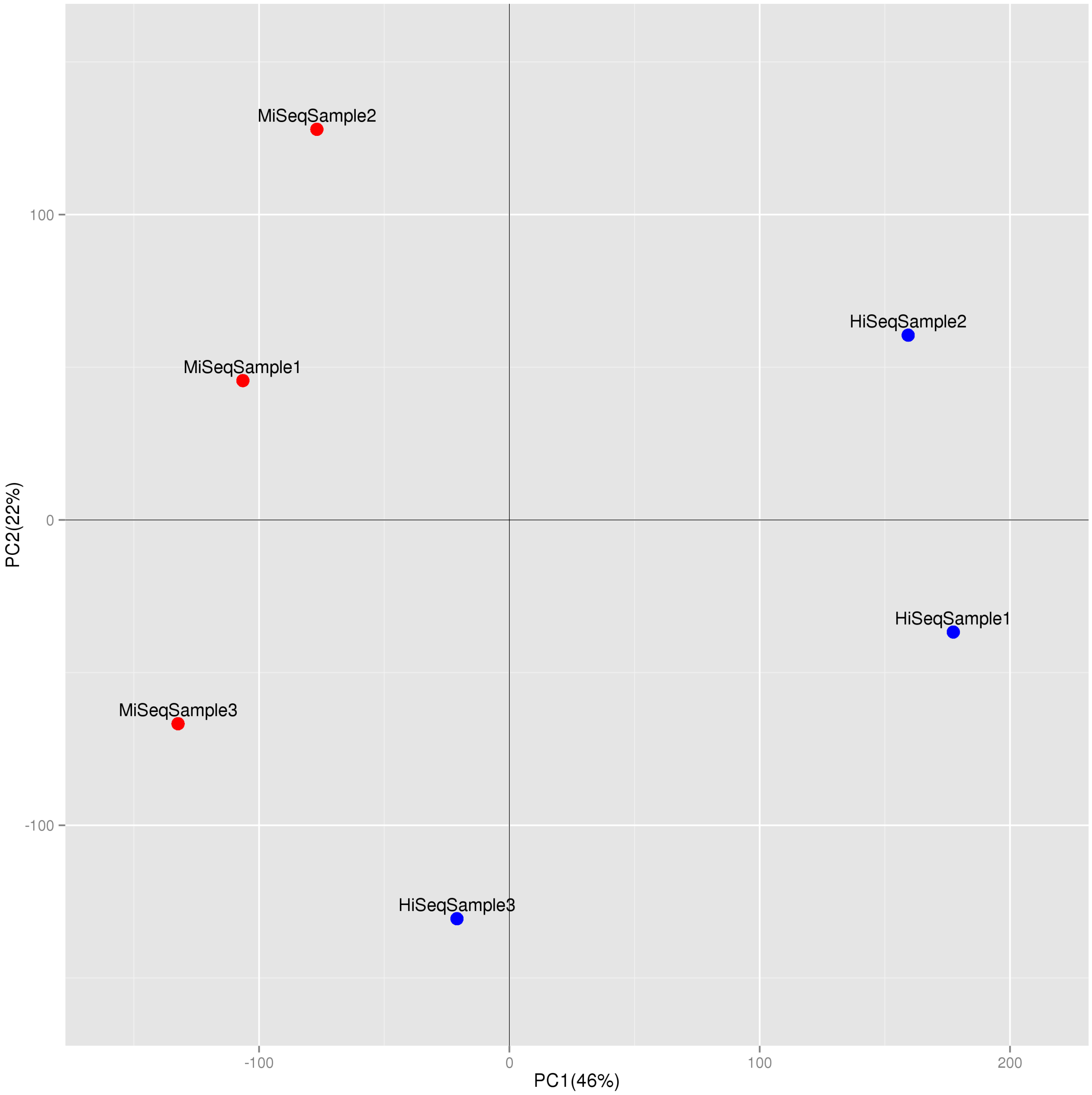
Variance stabilizing transformed (VST) value of each gene was calculated by DESeq2 from gene count. The 500 genes with largest interquartile range (IQR) of VST value were selected for drawimg heatmap.

Based on heatmap, the MiSeq and HiSeq data from sample 3 were clustered together as we expected. The spearman correlation coefficient of sample 3 (0.86) between MiSeq and HiSeq was also higher than sample 1 and 2 (0.71 and 0.74).



3.6 PCA

The VST values of all genes were used in PCA analysis. Based on PCA image, the MiSeq and HiSeq data from sample 3 were also more similar than the data from sample 1 and 2.



3.7 Differential expression comparison

Ideally, there should be no difference between MiSeq and HiSeq paired data. But actually, DESeq2 detected 6787 genes differential expressed with adjust pvalue less than 0.05.

Table 6: Top 50 differential expressed genes

Name	M1	M2	M3	H1	H2	H3	log2(fc)	pvalue	padj
SEPT9	602	831	612	38	83	94	-4.42	0.00	0.00
COL6A2	944	2257	646	151	425	216	-3.34	0.00	0.00
SREBF1	445	1423	1640	60	181	388	-3.77	0.00	0.00
ATN1	1427	792	626	123	100	149	-3.97	0.00	0.00
DNAJB2	216	318	348	12	23	43	-4.70	0.00	0.00
SNRNP70	775	858	795	92	145	160	-3.75	0.00	0.00
MARK2	578	754	492	116	159	190	-3.11	0.00	0.00
ATXN2L	391	543	538	60	89	181	-3.41	0.00	0.00
ACTN4	2335	2482	2192	626	695	1047	-2.74	0.00	0.00
PLEC	1211	1587	365	98	221	52	-4.15	0.00	0.00
SCRIB	454	524	251	27	61	37	-4.35	0.00	0.00
NBEAL2	221	556	314	25	44	50	-4.22	0.00	0.00
COL18A1	148	615	96	15	61	19	-4.07	0.00	0.00
ADRBK1	340	273	473	52	40	150	-3.47	0.00	0.00
FASN	835	424	2844	117	43	730	-3.76	0.00	0.00
PTK7	425	820	736	86	157	301	-3.11	0.00	0.00
KDM5C	545	499	867	87	123	280	-3.21	0.00	0.00
PHRF1	154	228	337	16	17	49	-4.31	0.00	0.00
SBF1	206	337	158	22	56	33	-3.75	0.00	0.00
AP3D1	315	1038	487	59	188	205	-3.16	0.00	0.00
SF3A2	190	702	213	28	76	40	-3.88	0.00	0.00
EPN1	171	212	259	22	27	44	-3.90	0.00	0.00
LRP5	179	198	230	13	17	49	-4.19	0.00	0.00
PRR12	166	92	175	9	4	12	-5.10	0.00	0.00
FURIN	212	358	125	14	38	29	-4.11	0.00	0.00
TAOK2	283	183	302	12	7	44	-4.88	0.00	0.00
CEP170B	321	223	473	21	18	111	-4.20	0.00	0.00
LMNB2	409	379	500	27	19	99	-4.48	0.00	0.00
KRT5	1202	445	191	266	138	100	-2.73	0.00	0.00
AKT1	515	768	3426	80	211	1365	-3.06	0.00	0.00
SLC2A4RG	97	298	211	17	61	76	-3.19	0.00	0.00
FBRSL1	139	244	272	19	23	43	-4.03	0.00	0.00
MAPK8IP3	297	283	158	31	20	15	-4.44	0.00	0.00
ZBTB7A	302	376	201	76	91	81	-2.90	0.00	0.00
EHBP1L1	338	177	152	6	9	4	-5.78	0.00	0.00
USF2	419	327	376	90	60	104	-3.27	0.00	0.00
SEMA3F	86	193	224	6	13	42	-4.33	0.00	0.00
NFIX	1051	829	50	182	214	21	-3.03	0.00	0.00
PPP1R14B	269	669	347	33	150	121	-3.30	0.00	0.00
PRKCSH	805	1117	791	205	232	375	-2.89	0.00	0.00

4 Discussion

Due to unknown reason, there are huge difference between the sequencing data from MiSeq and HiSeq platforms. Combining those data together to do analysis may introduce more false positives.