

Comparison between MiSeq and HiSeq using Paired Data

Quanhui Sheng, Bojana Jovanovic, Scott Austin Beeler

February 25, 2014

Contents

1 Purpose	2
2 Dataset	2
3 Raw file quality control	2
4 Mapping	9
5 Correlation between MiSeq and HiSeq	9
6 Heatmap	10
7 PCA	10
8 Differential expression comparison	11

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 further analysis.

2 Dataset

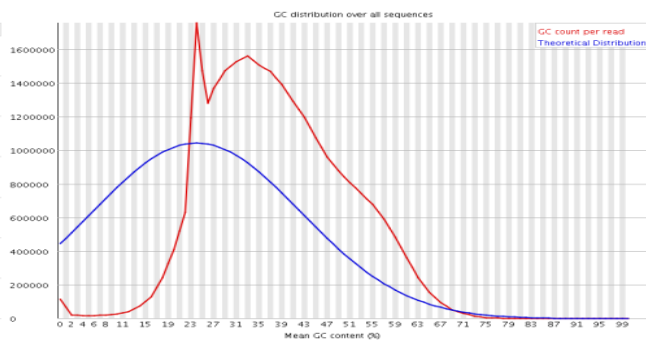
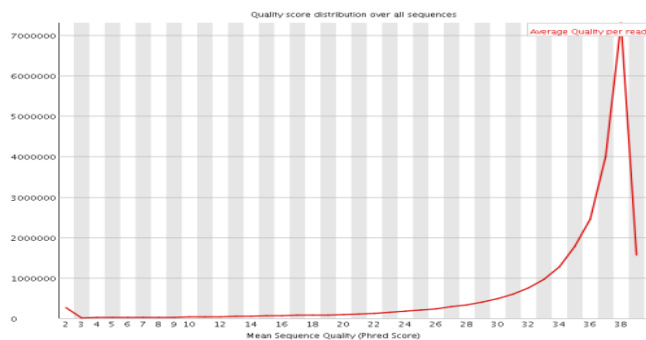
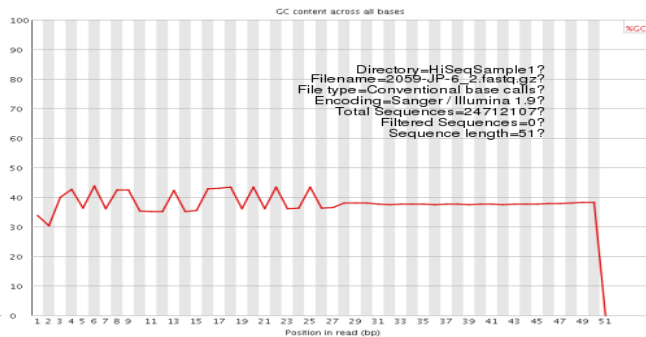
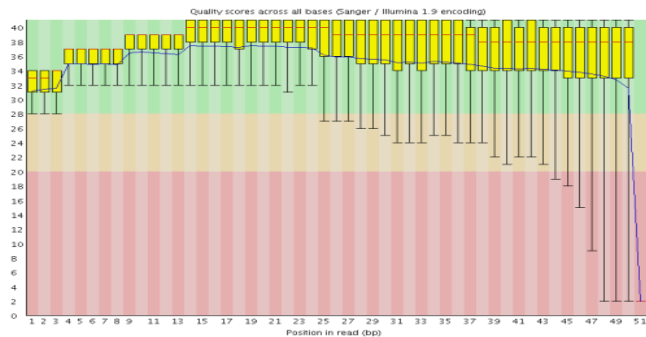
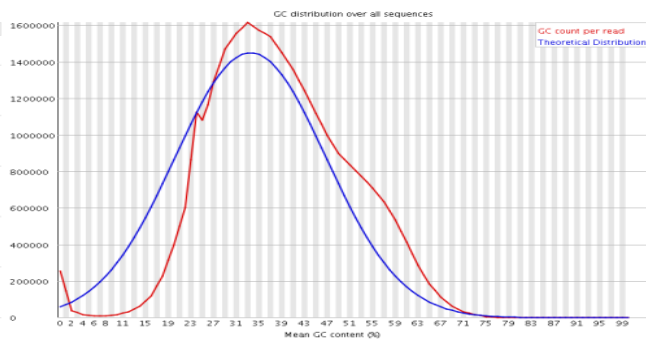
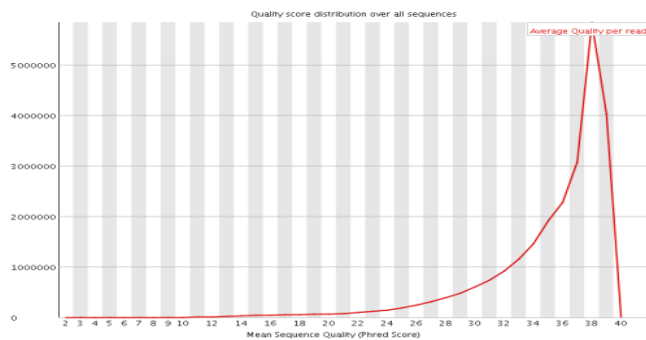
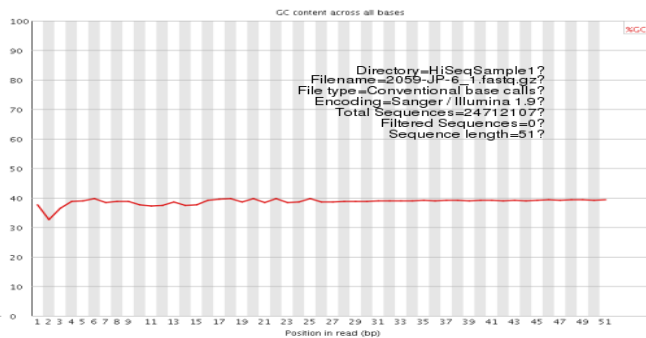
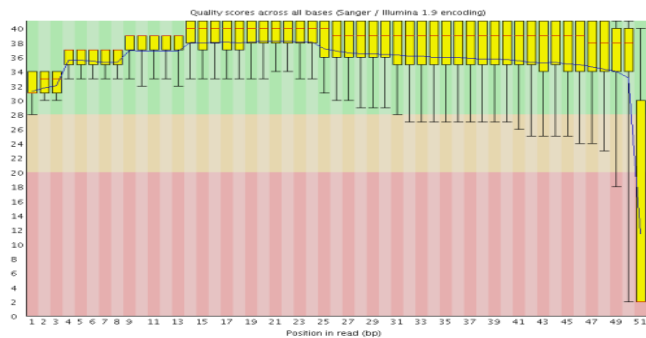
Three samples were pair-end sequenced by both MiSeq and HiSeq machines, as table 1 illustrated.

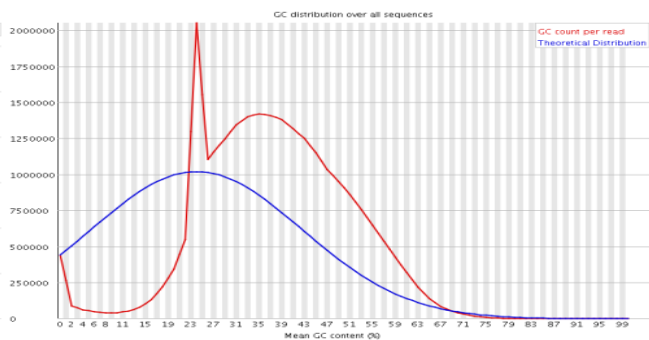
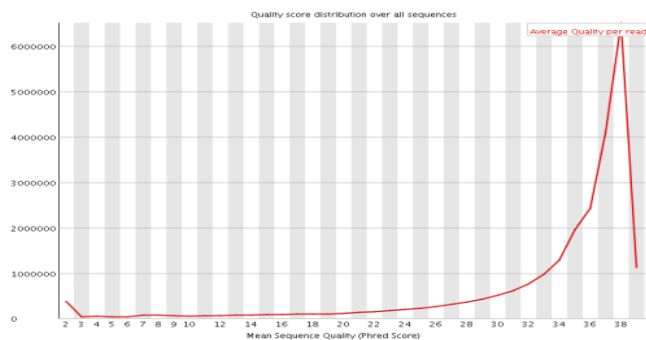
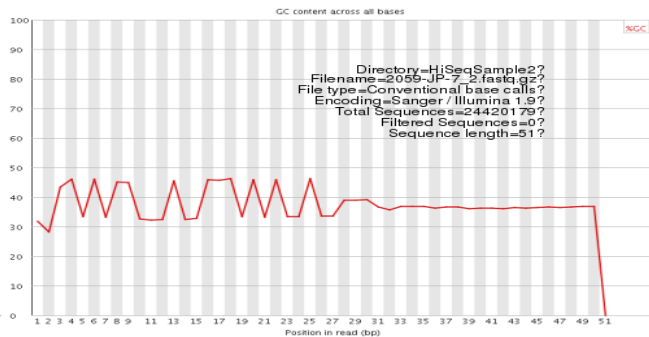
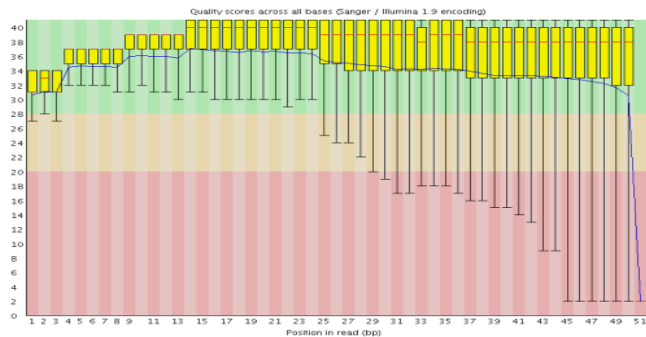
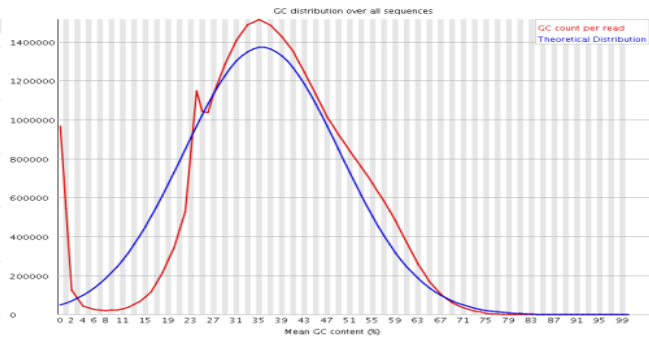
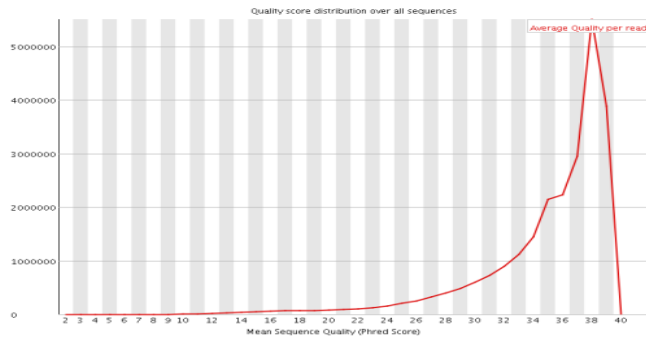
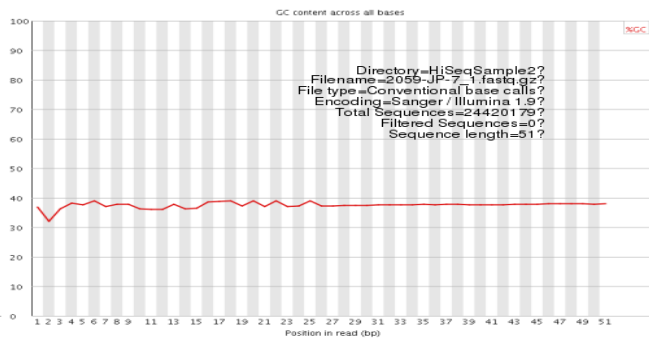
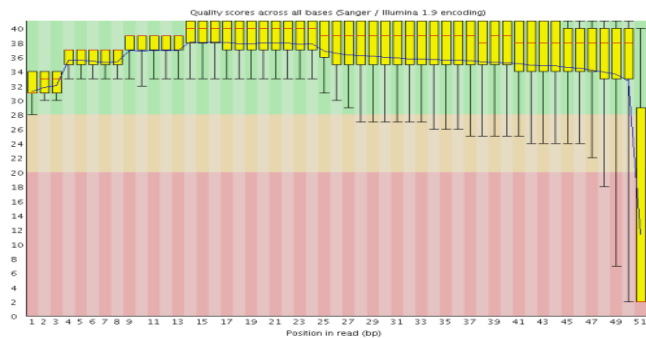
Table 1: Sample Information		
Type	Source	Alias
HiSeq	2059-JP-6	HiSeqSample1
HiSeq	2059-JP-7	HiSeqSample2
HiSeq	2059-JP-9	HiSeqSample3
MiSeq	IG-33_408637_S4	MiSeqSample1
MiSeq	IG-34_S3	MiSeqSample2
MiSeq	IG-39_408648_S5	MiSeqSample3

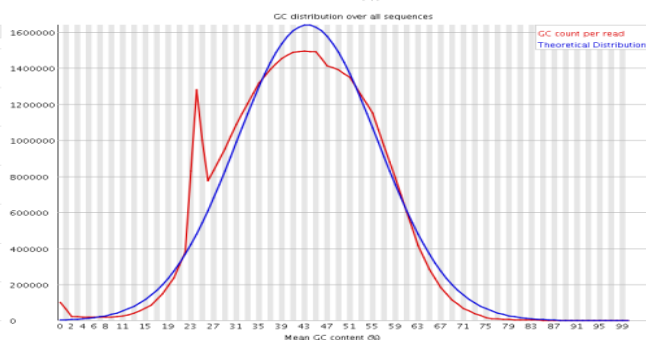
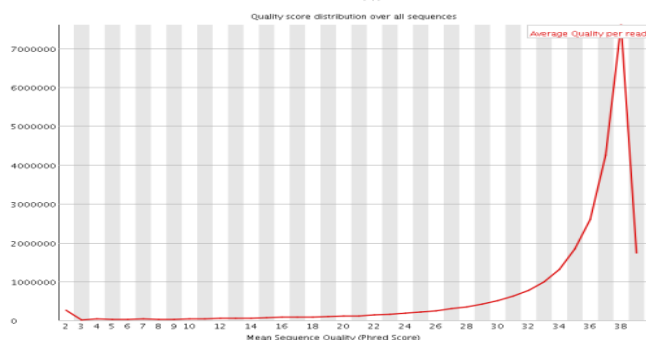
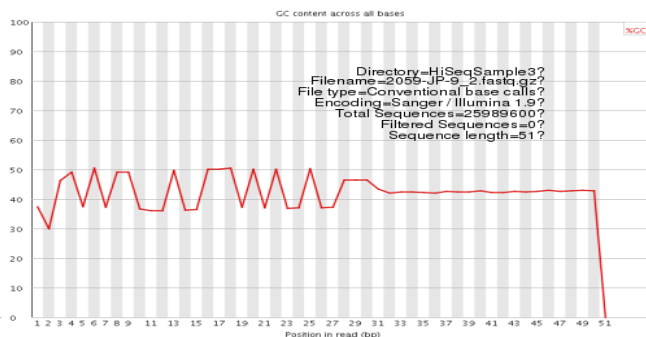
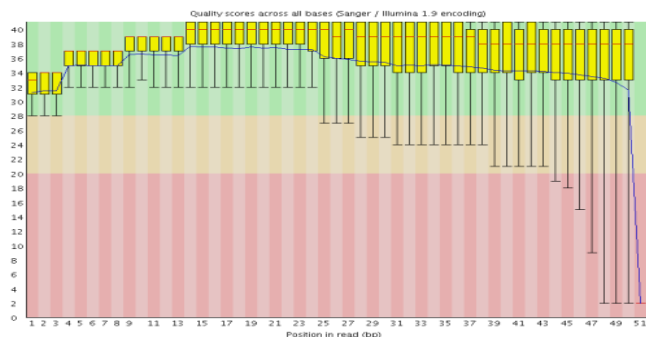
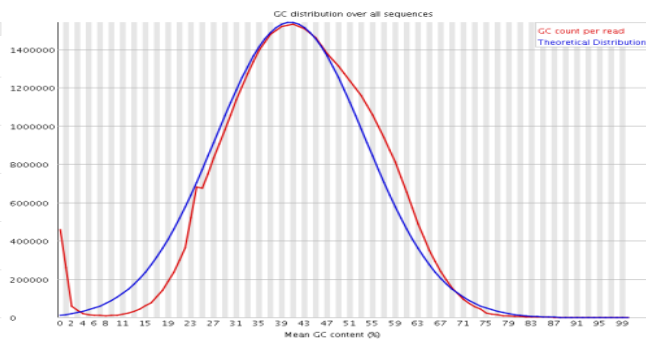
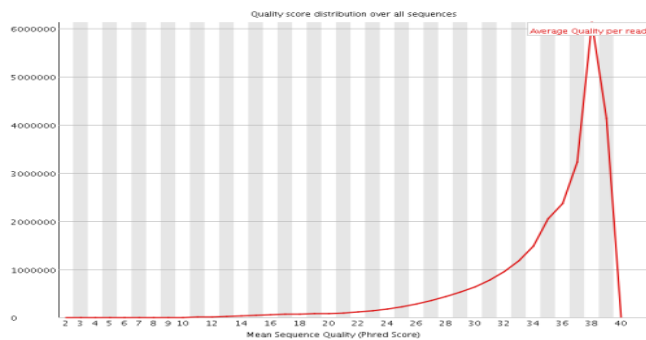
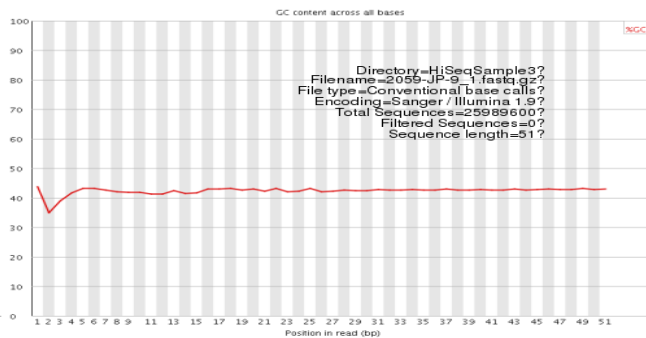
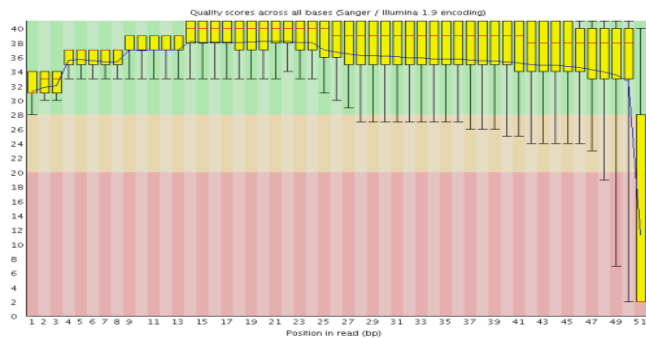
3 Raw file quality control

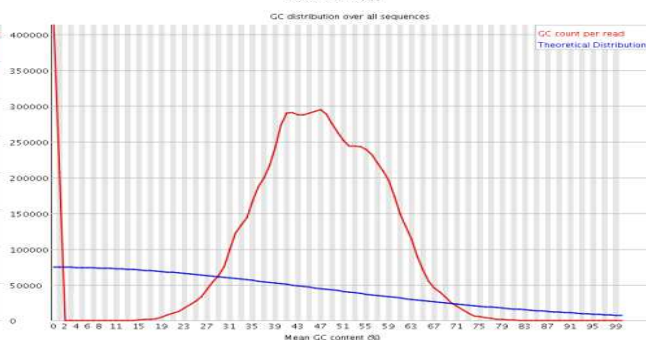
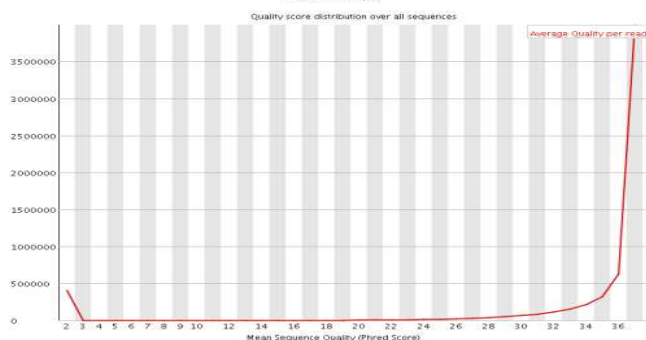
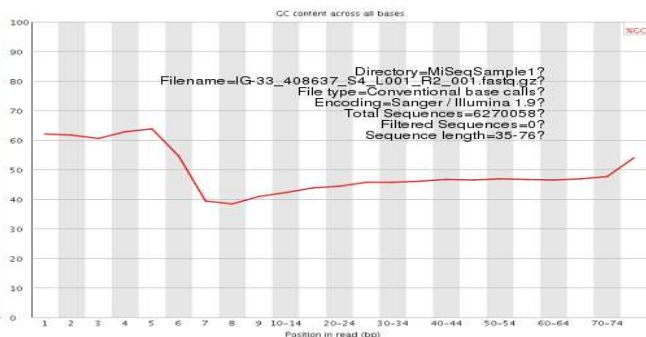
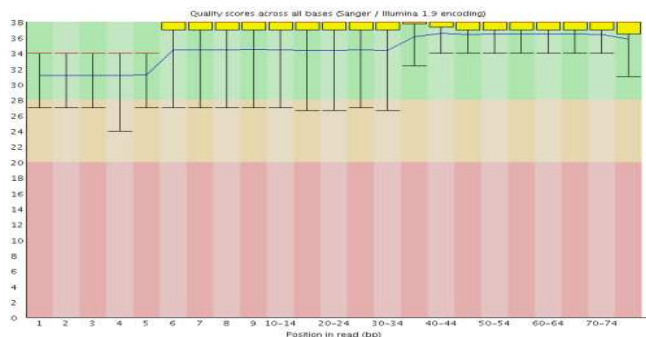
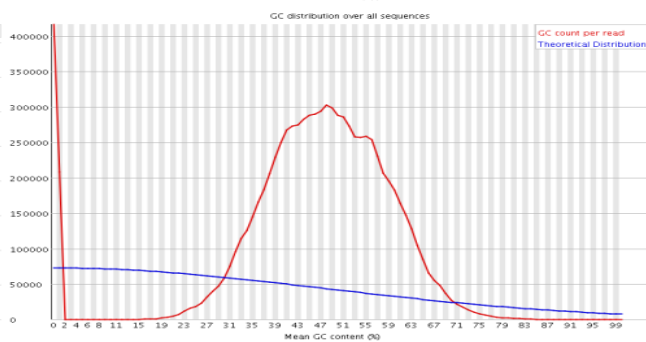
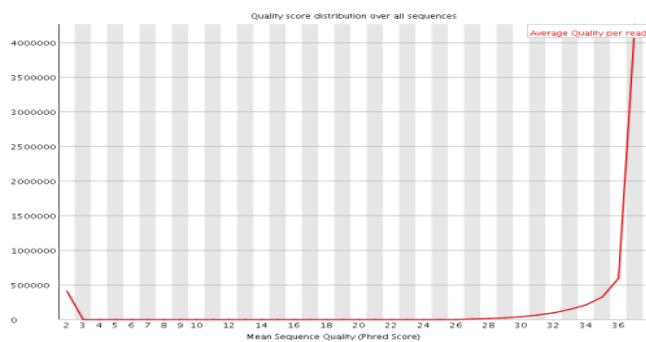
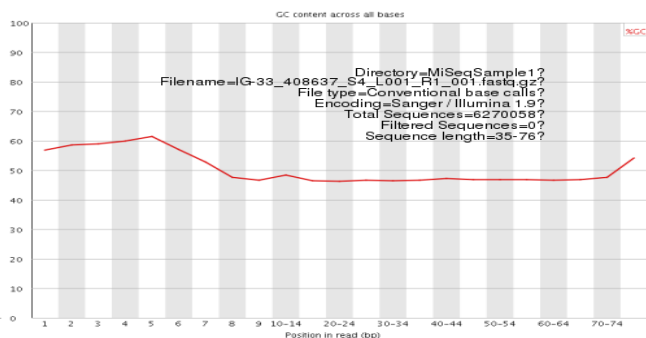
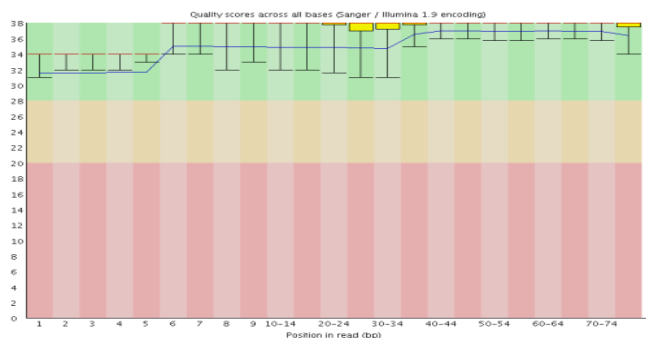
FastQC v0.10.1 was used. Two GC content bias were observed:

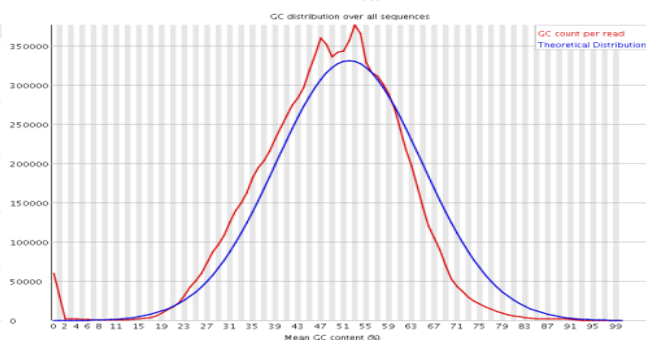
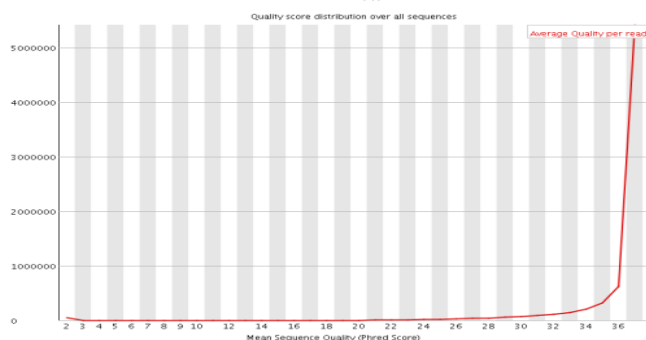
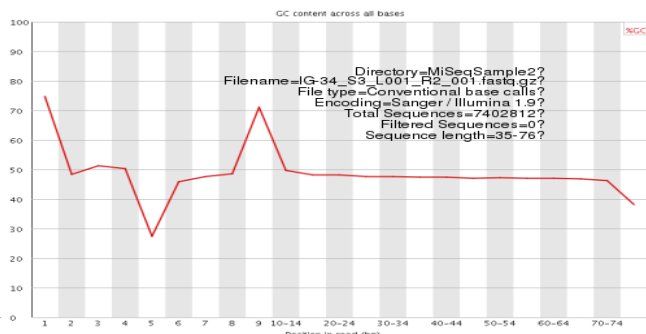
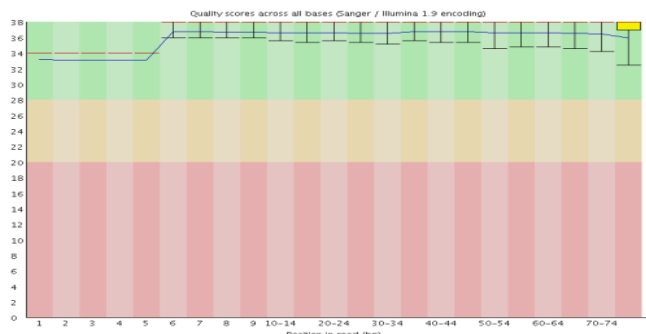
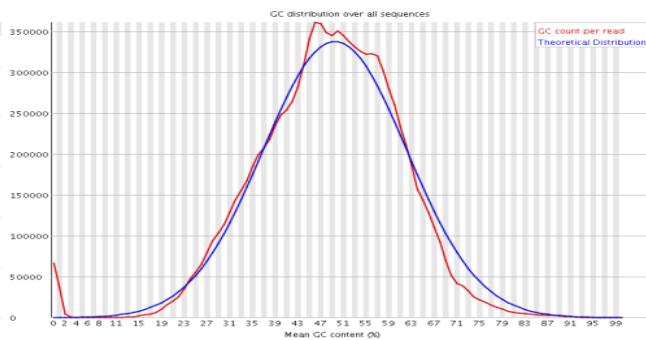
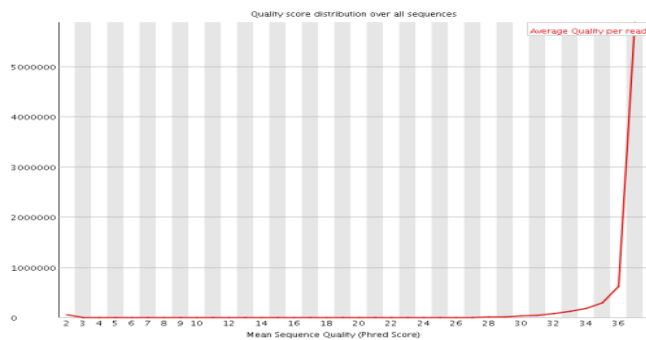
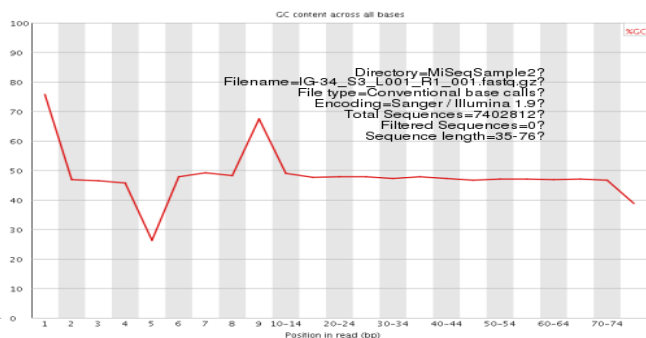
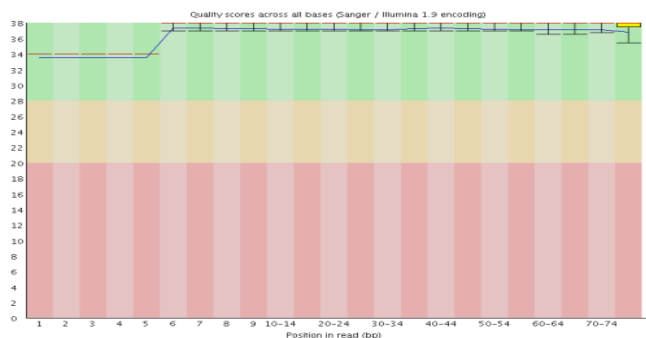
1. The GC contribution of HiSeq data were shifted to 0.4
2. There are a lot of reads in MiSeq have no GC which means they might be all 'N'.

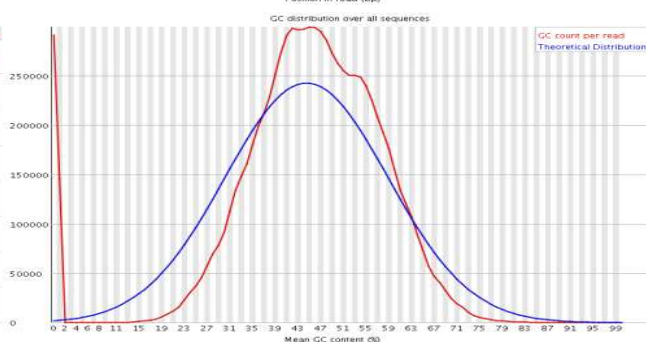
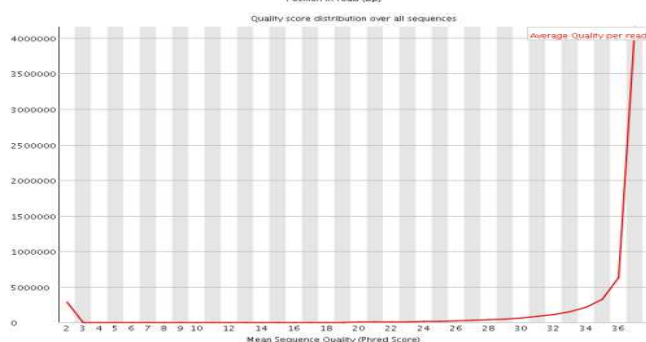
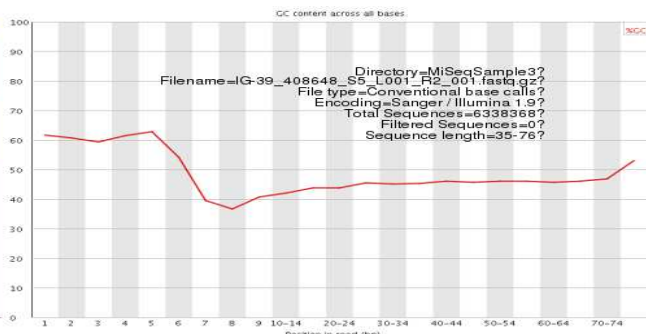
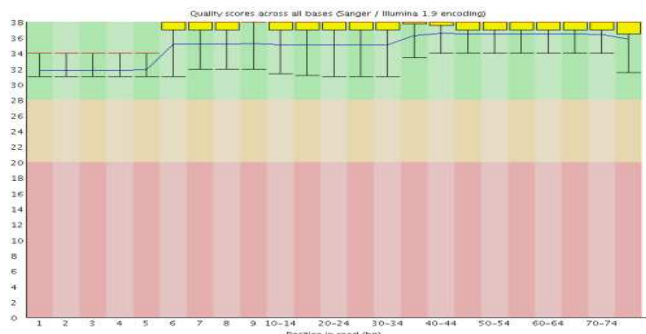
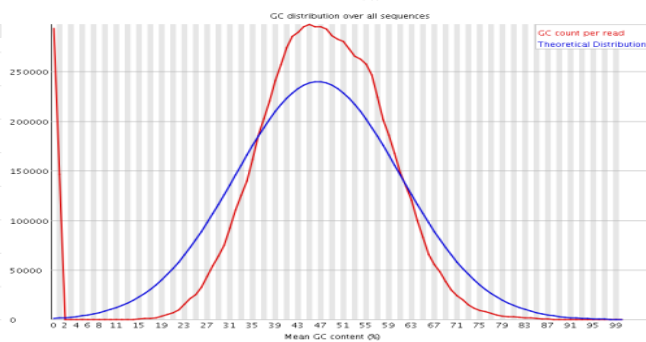
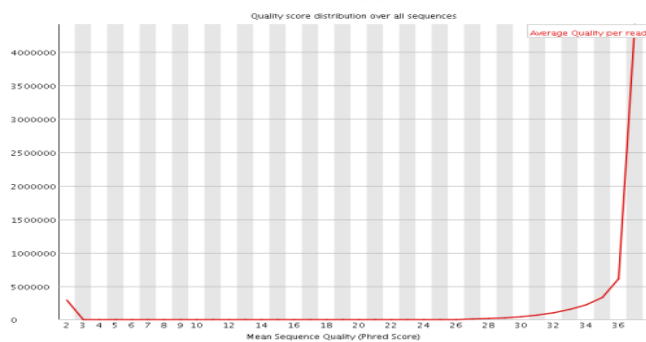
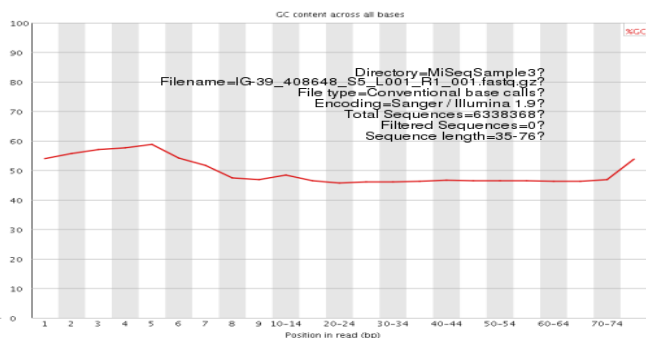
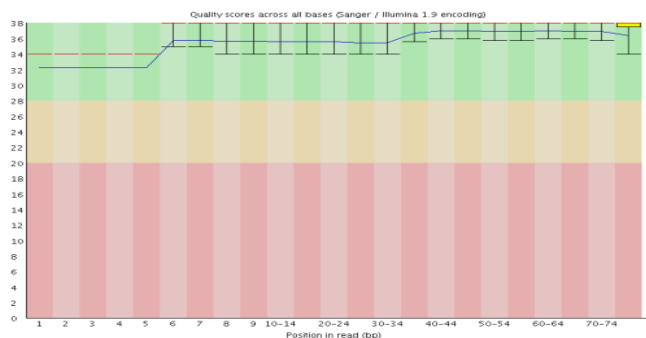












4 Mapping

TopHat v2.0.9 was used to map reads to human genome 19. HTSeq v0.5.4p3 was used to count genes.

```
tophat2 --segment-length 25 -r 0 -p 8 --keep-fastq-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

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
```

5 Correlation between MiSeq and HiSeq

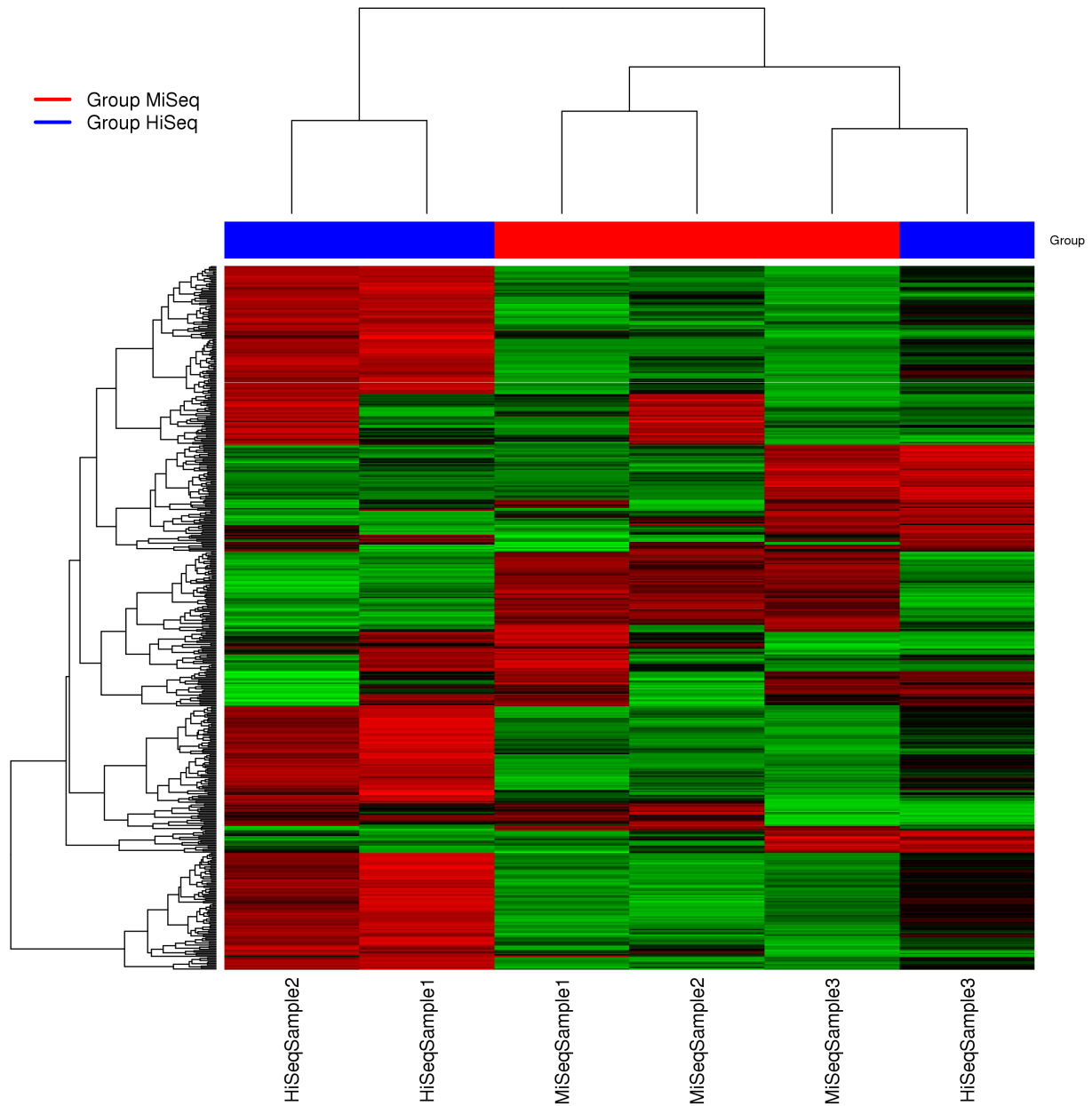
Spearman correlation was calculated between MiSeq gene count and HiSeq gene count for each sample (table 2).

Table 2: Spearman Correlation

sample	spcorr
Sample1	0.72
Sample2	0.75
Sample3	0.86

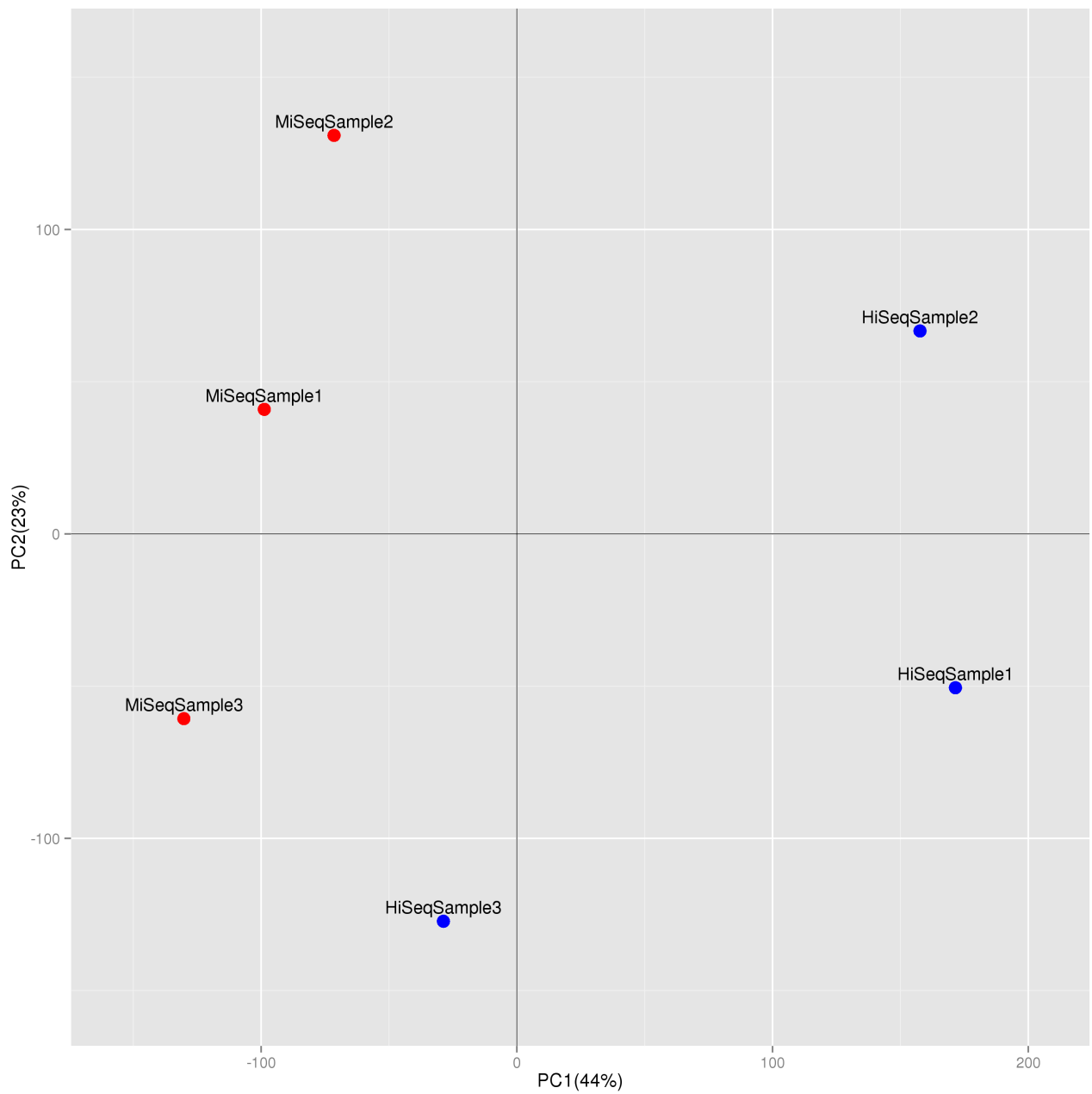
6 Heatmap

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 (0.72 and 0.75).



7 PCA

Based on PCA image, the MiSeq and HiSeq data from sample 3 were also more similar than the data from sample 1 and 2.



8 Differential expression comparison

Ideally, there should be no difference between MiSeq and HiSeq paired data. But actually, DESeq2 detected 6156 genes differential expressed with adjust pvalue less than 0.05.

Table 3: Top 50 differential expressed genes

Name	M1	M2	M3	H1	H2	H3	log2(fc)	pvalue	padj
COL6A2	944	2257	646	151	421	215	-3.26	0.00	0.00
SREBF1	445	1423	1640	58	177	374	-3.72	0.00	0.00
SEPT9	602	831	612	34	80	88	-4.41	0.00	0.00
ATN1	1427	792	626	121	100	148	-3.89	0.00	0.00
DNAJB2	216	318	348	12	23	42	-4.62	0.00	0.00
ATXN2L	391	543	538	57	88	178	-3.35	0.00	0.00
MARK2	578	754	492	114	158	186	-3.04	0.00	0.00
ACTN4	2335	2482	2192	586	657	1008	-2.72	0.00	0.00
SNRNP70	775	858	795	92	141	156	-3.68	0.00	0.00
PPP6R1	274	118	331	14	9	35	-4.69	0.00	0.00
KDM5C	545	499	867	78	106	274	-3.26	0.00	0.00
SCRIB	454	524	251	27	59	36	-4.29	0.00	0.00
COL18A1	148	615	96	12	61	16	-4.14	0.00	0.00
PLEC	1211	1587	365	98	218	51	-4.06	0.00	0.00
FASN	835	424	2844	108	41	716	-3.74	0.00	0.00
AGRN	215	450	224	17	28	19	-4.65	0.00	0.00
PTK7	425	820	736	60	132	289	-3.27	0.00	0.00
AP3D1	315	1038	487	59	187	201	-3.08	0.00	0.00
NBEAL2	221	556	314	26	44	50	-4.10	0.00	0.00
ADRBK1	340	273	473	44	33	147	-3.54	0.00	0.00
PHRF1	154	228	337	16	17	49	-4.21	0.00	0.00
LRP5	179	198	230	13	17	49	-4.10	0.00	0.00
FURIN	212	358	125	14	37	28	-4.05	0.00	0.00
SBF1	206	337	158	22	55	33	-3.67	0.00	0.00
PRKCSH	805	1117	791	193	237	364	-2.84	0.00	0.00
SF3A2	190	702	213	27	73	39	-3.82	0.00	0.00
LMNB2	409	379	500	27	19	98	-4.39	0.00	0.00
KRT5	1202	445	191	263	136	96	-2.67	0.00	0.00
CEP170B	321	223	473	21	17	111	-4.13	0.00	0.00
PRR12	166	92	175	9	4	12	-4.99	0.00	0.00
EPN1	171	212	259	22	26	42	-3.84	0.00	0.00
TAOK2	283	183	302	12	6	43	-4.83	0.00	0.00
SEMA3F	86	193	224	6	13	41	-4.26	0.00	0.00
CHD3	618	410	1598	53	75	456	-3.56	0.00	0.00
AKT1	515	768	3426	72	198	1344	-3.05	0.00	0.00
SLC2A4RG	97	298	211	17	61	75	-3.10	0.00	0.00
OGFR	139	276	141	11	40	24	-3.91	0.00	0.00
PPP1R14B	269	669	347	32	150	117	-3.24	0.00	0.00
MYO18A	305	90	220	39	11	69	-3.49	0.00	0.00
PPP1R18	391	281	180	71	74	85	-2.85	0.00	0.00
ARHGEF11	685	333	172	49	29	50	-3.97	0.00	0.00
NFIX	1051	829	50	172	209	21	-2.97	0.00	0.00
EHBP1L1	338	177	152	4	8	4	-5.85	0.00	0.00
LCLAT1	211	67	195	3306	1428	4471	3.15	0.00	0.00
C9orf69	80	240	138	10	33	40	-3.52	0.00	0.00
WIZ	222	419	135	14	60	28	-3.91	0.00	0.00
MAP2K7	196	385	116	43	90	53	-2.83	0.00	0.00
EMD	220	156	202	27	33	60	-3.33	0.00	0.00
LENG8	1059	686	643	66	115	111	-3.96	0.00	0.00
AL589743.1	98	175	107	12	27	27	-3.54	0.00	0.00