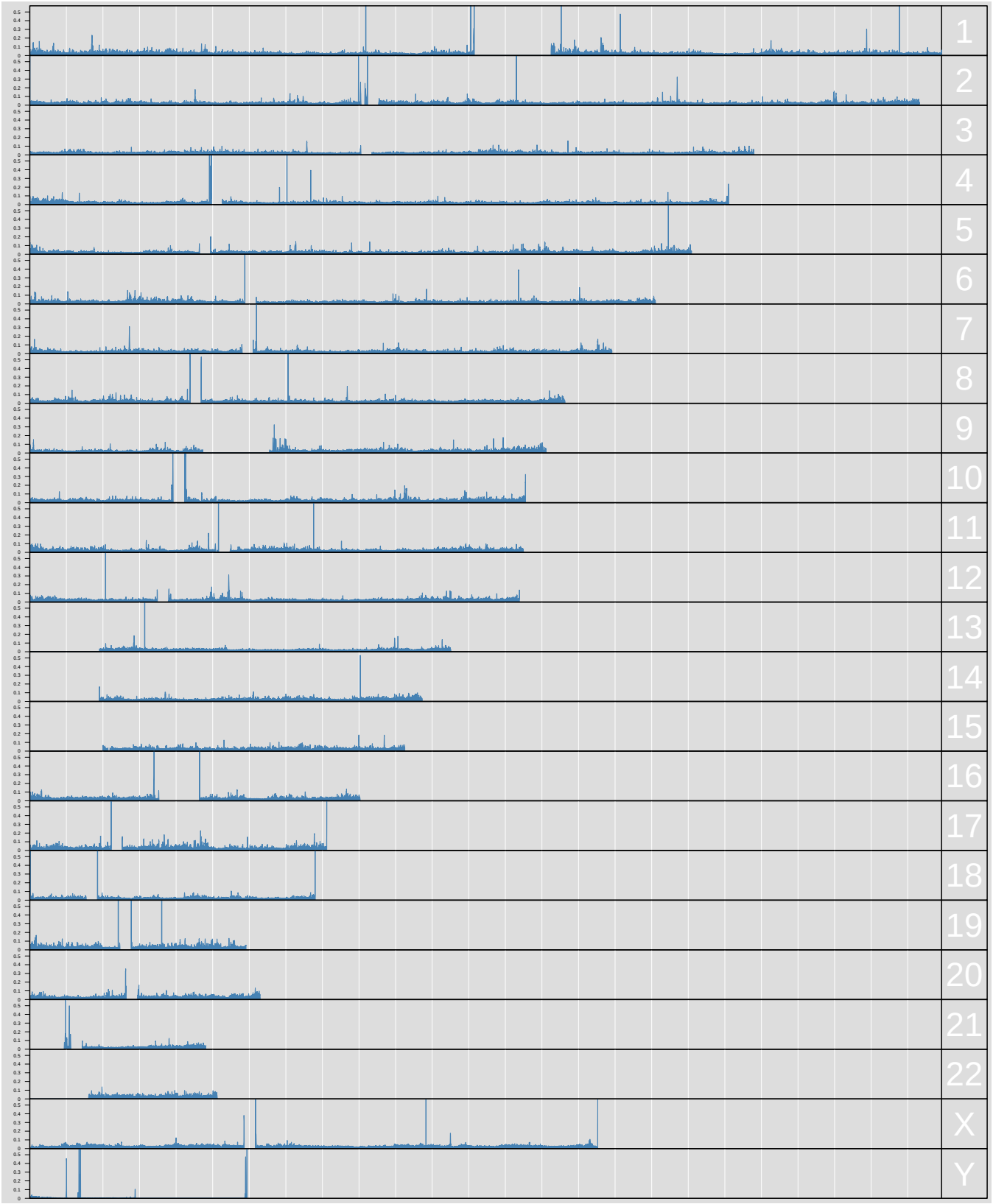
ENCODE CD14+Monocyte Histone ChIP-seq: 671k27

Prepared by Zhe Zhang

April 30, 2013

Contents


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

1 Introduction

Project:

Sample name: 671k27

Genome name: hg19

1.1 BAM file

Size: 1.46 GB

Created: 2013-04-29 11:46:49

Modified: 2013-04-29 07:33:57

Location: > home > zhangz > hts > projects > ks001 > 2013-04_BGI_ChIPseq > bams > novoalign_671k27_-aligned.bam

1.2 Summary statisitics

Number of chromosomes	24
Total reference size (bp)	3,095,677,412
Total effective size (bp)	2,897,316,137
Total entries	27,158,306
Total mapped reads	2,063,686
Total upmapped reads	0
Total mappings	27,158,306
Total mapping locations	1,853,567
Base N%	0.0031
(G+C)%	46.91
Mapped to forward strand%	50.01
Duplicated mapping reads%	13.4
Best sequencing quality	41
Average sequencing quality	38.77
Maximum mapping length (bp)	49
Minimum mapping length (bp)	25
Average mapping length (bp)	48.5
Best mapping quality	70
Average mapping quality	45.49
Highest sequencing depth	9,965
Average sequencing depth	0.034
Mapped reads per kilobase	0.71

Table 1: **Summary statistics**
Effective size: chromosome length without assembly gaps.
Sequencing quality score: assigned by the resequencing machine to indicate base calling confidence.
Mapping quality score: assigned by the alignment program to indicating mapping confidence.
Mapping location: strand-specific chromosomal location mapped to by the first base of one or more reads.
Duplicated mapping: the first base of multiple reads mapped to the same strand and chromosomal location.

2 Read count and sequencing coverage

This section summarizes the sequencing depth of reference chromosomes. Sequencing depth equals how many times a nucleotide base was sequenced.

2.1 Depth categories

Depth	Count	Percentage
Depth=0	2,776,387,677	97.03
Depth>=1	84,944,929	2.97
Depth>=5	225,580	0.01
Depth>=10	100,653	0.00
Depth>=20	43,408	0.00
Depth>=30	26,048	0.00
Depth>=50	12,256	0.00
Depth>=100	5,637	0.00
Depth>=1000	1,642	0.00
Depth=9965	1	0.00

Table 2: **Depth by cutoffs.** Number and percentage of genomic locations (single bases) having the same or higher sequencing depth than given values.

2.2 Depth by chromosome

Table 3: Sequencing depth by chromosome

Chromosome	Chromosome_length	E↑ecitive_size	Total_reads	Unique_mapping	Average_depth	Maximum_depth	Maximum_location
chr1	249,250,621	225,280,621	183,108	155,787	0.04	5,721	91,092,890
chr2	243,199,373	238,207,373	181,440	153,048	0.04	5,224	128,285,576
chr3	198,022,430	194,797,140	115,158	111,578	0.03	165	52,466,861
chr4	191,154,276	187,661,676	125,170	107,384	0.03	9,965	66,939,006
chr5	180,915,260	177,695,260	112,242	105,348	0.03	3,322	171,332,108
chr6	171,115,067	167,395,067	108,718	103,598	0.03	1,278	130,184,088
chr7	159,138,663	155,353,663	101,563	97,765	0.03	116	58,669,050
chr8	146,364,022	142,888,922	112,740	88,910	0.04	7,817	67,487,295
chr9	141,213,431	120,143,431	82,376	79,488	0.03	89	47,511,230
chr10	135,534,747	131,314,747	101,824	94,676	0.04	271	39,020,270
chr11	135,006,516	131,129,516	94,504	86,910	0.03	2,688	73,980,570
chr12	133,851,895	130,481,895	90,459	81,721	0.03	4,334	20,544,443
chr13	115,169,878	95,589,878	58,679	53,277	0.03	1,329	12,398,369
chr14	107,349,540	88,289,540	58,875	55,083	0.03	1,829	71,341,410
chr15	102,531,392	81,694,769	57,735	55,825	0.03	44	8,264,635
chr16	90,354,753	78,884,753	70,250	62,564	0.04	3,002	33,853,738
chr17	81,195,210	77,795,210	66,052	62,787	0.04	443	19,256,013
chr18	78,077,248	74,657,248	50,666	46,281	0.03	808	74,656,467
chr19	59,128,983	55,808,983	64,474	49,008	0.05	5,787	23,974,122
chr20	63,025,520	59,505,520	46,611	44,671	0.04	89	42,919,877
chr21	48,129,895	35,108,702	36,208	24,921	0.05	4,920	316,273
chr22	51,304,566	34,894,566	28,892	27,808	0.04	54	661,661
chrX	155,270,560	151,100,560	104,959	96,017	0.03	3,752	104,487,419
chrY	59,373,566	25,653,566	10,983	9,112	0.02	383	25,653,527

2.3 Depth by genomic feature

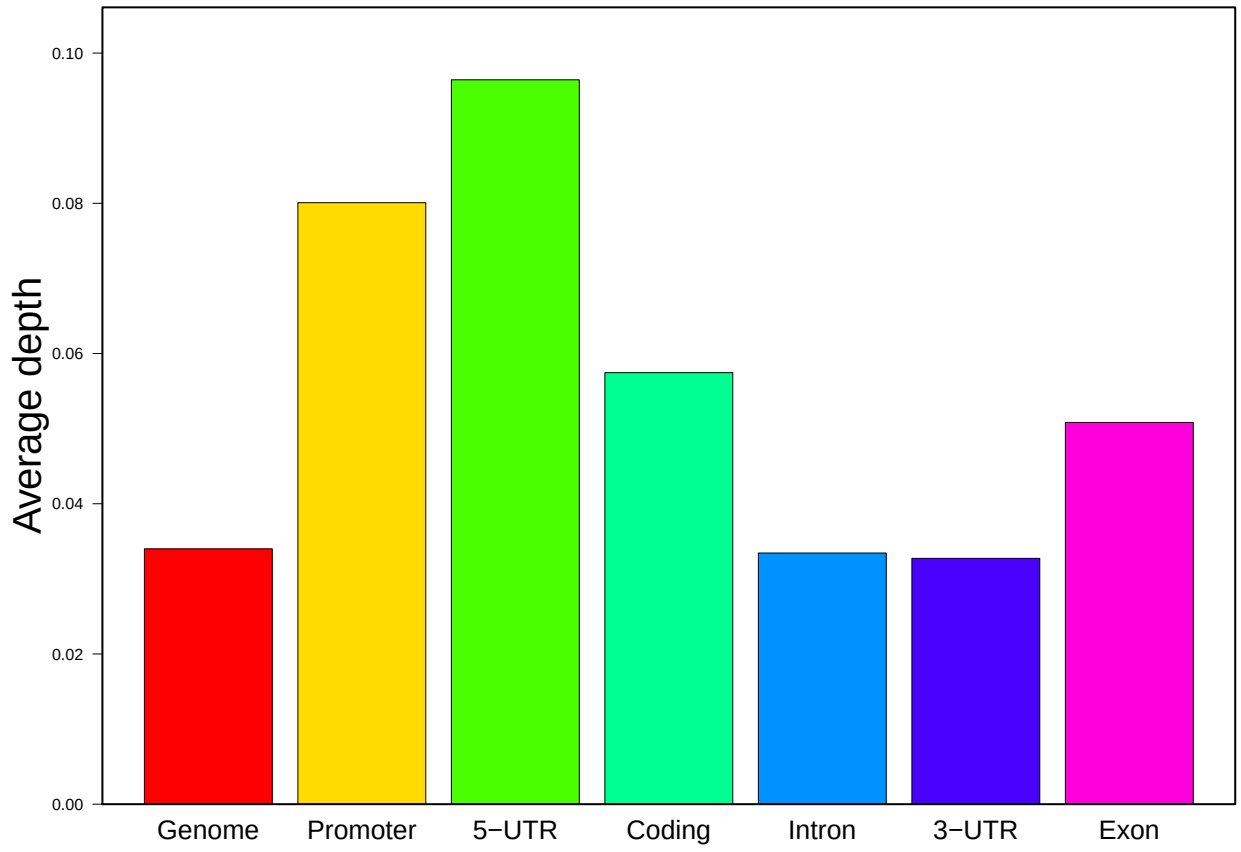


Figure 1: **Average depth of genomic features.** Genomic features are regions annotated based on previous knowledge, such as the RefSeq gene track downloaded from UCSC genome browser. Many applications of high-throughput sequencing technologies, such as exome sequencing and RNA-seq, expect higher depth at exons.

3 Sequencing quality

This section summarizes the sequencing quality scores assigned by the sequencer to single bases in each sequencing read and stored in the **<QUAL>** field of BAM files.

Quality score summary:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.00   37.00   40.00   38.77   41.00   41.00
```

3.1 Quality score categories

Score	Count	Percentage
Score=0	0	0.00
Score>=5	6,115,009	99.88
Score>=10	6,109,807	99.80
Score>=13	6,107,202	99.75
Score>=20	6,093,994	99.54
Score>=30	6,025,542	98.42
Score>=40	3,472,376	56.72
Score=41	2,719,313	44.42

Table 4: **Score categories.** The number and percentage of base calls having the quality socre equal to or higher than given values.)

3.2 Overall score distribution

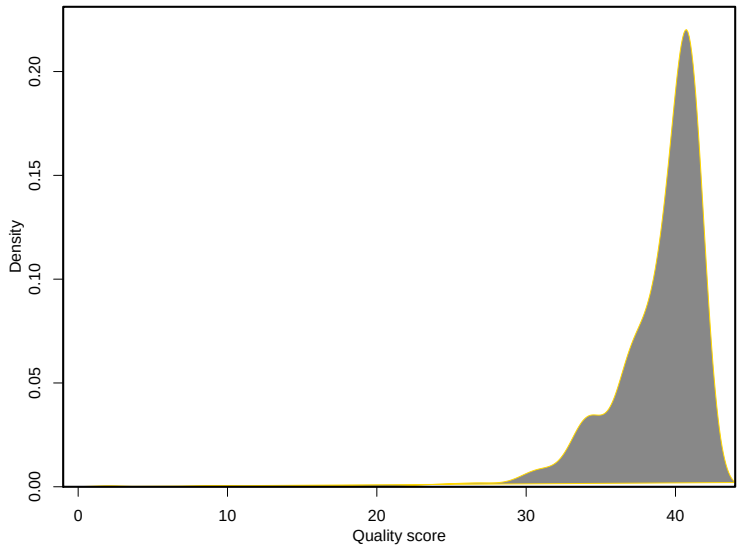


Figure 2: **Score distribution.** This distribution is based on all bases of randomly selected sequencing reads, so position-specific sequencing quality is not considered (see below). The quality scores are calculated by subtracting 33 from the integers corresponding to the ASCII characters in **<QUAL>**. If the convention of Sanger sequencing was applied to generate the ASCII characters, they are equal to -10*Log10(p value), where p value is the likelihood of incorrect base call.

3.3 Position-specific score distribution

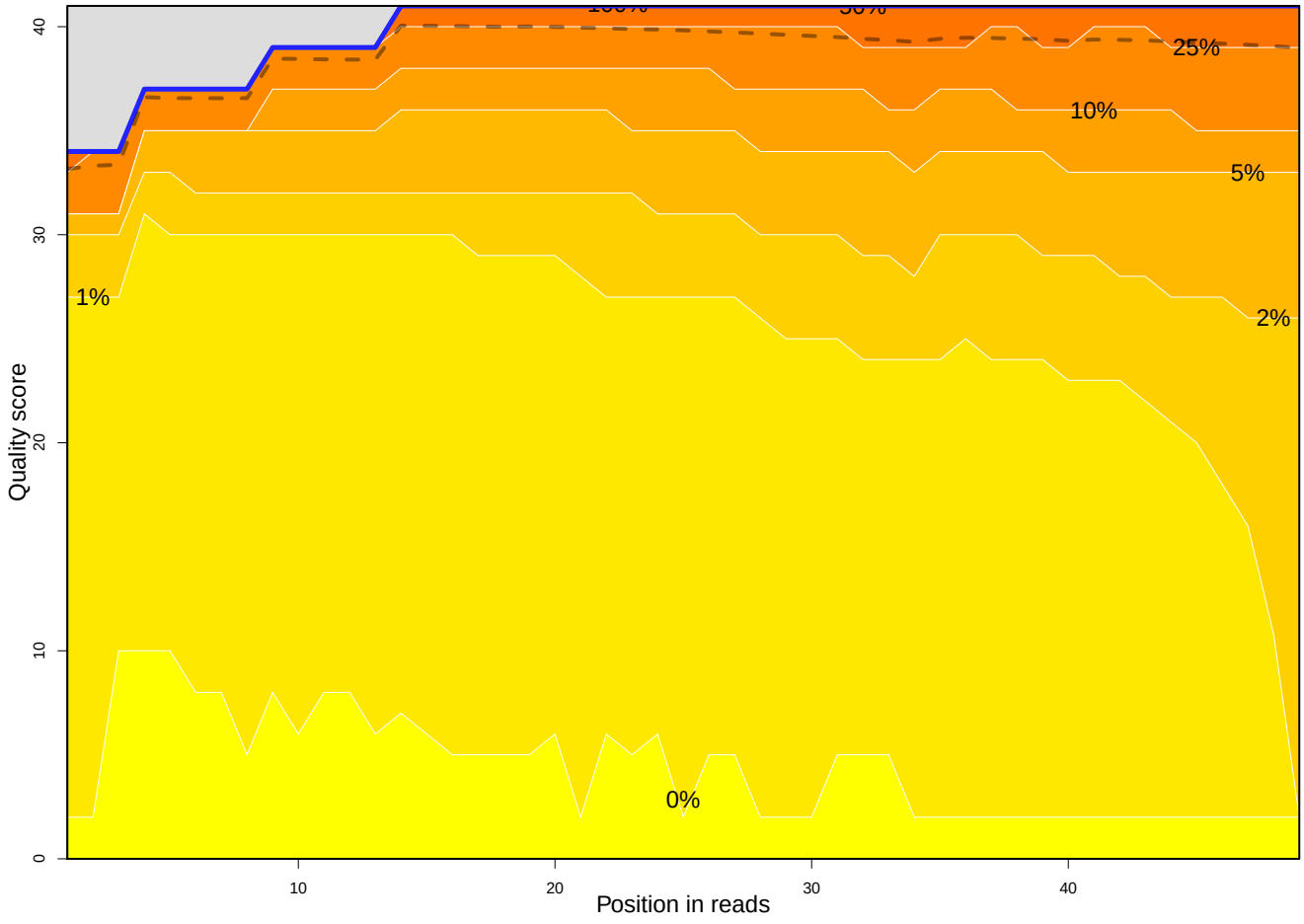


Figure 3: **Position-specific sequencing scores.** This plot shows quality scores at different positions within reads. The dashed lines represents the means of quality scores at different positions; whereas the heat gradient corresponds to percentiles.

4 Mapping to reference

This section summarizes the mapping of sequencing reads to reference chromosomes.

4.1 Mapping length

Mapping length corresponds to the **<QWIDTH>** field in BAM files, which is the number of bases in a read mapped to reference. Hard clipping reduces mapping length while soft clipping does not.

Mapping length summary:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
25.0    49.0    49.0    48.5    49.0    49.0
```

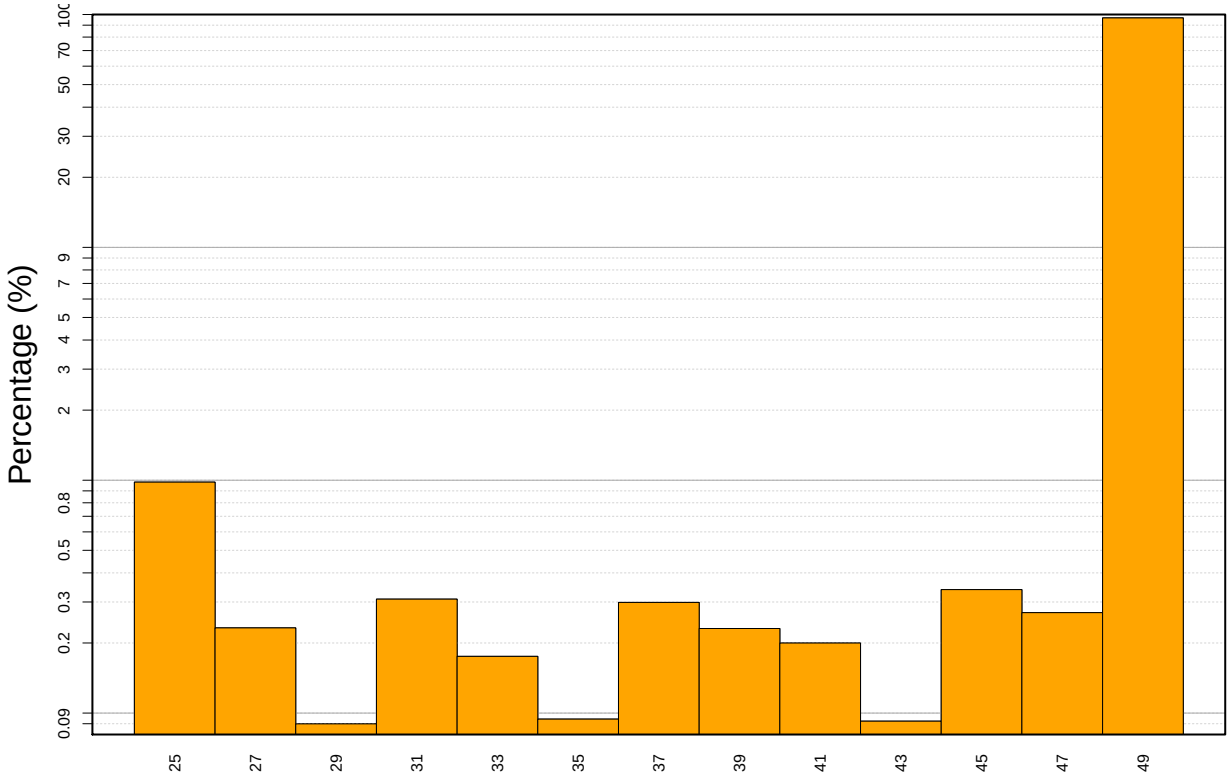


Figure 4: **Frequency of mapping lengths.**

4.2 Mapping flag

Mapping flag is stored in the **<FLAG>** field of BAM files. It uses a series of bitwise codes to represent different combinations of mapping results:

Bitwise	Description
0X1	template having multiple segments in sequencing
0X2	each segment properly aligned according to the aligner
0X4	segment unmapped
0X8	next segment in the template unmapped
0X10	SEQ being reverse complemented
0X20	SEQ of the next segment in the template being reversed
0X40	the first segment in the template
0X80	the last segment in the template
0X100	secondary alignment
0X200	not passing quality controls
0X400	PCR or optical duplicate

4.2.1 Mapping flag categories

Code	Count	Percentage
0X1	0	0.00
0X2	0	0.00
0X4	0	0.00
0X8	0	0.00
0X10	13,576,225	49.99
0X20	0	0.00
0X40	0	0.00
0X80	0	0.00
0X100	25,094,620	92.40
0X200	0	0.00
0X400	0	0.00

Table 5: **Mapping flag categories.** The total number and percentage of reads flagged by each category.

4.2.2 Flag value breakdown

Table 6: The breakdown of values into 'ag categories.

Value	Count	Percentage	0X1	0X2	0X4	0X8	0X10	0X20	0X40	0X80	0X100	0X200	0X400
0	1,031,106	3.8	-	-	-	-	-	-	-	-	-	-	-
16	1,032,580	3.8	-	-	-	-	X	-	-	-	-	-	-
256	12,550,975	46.2	-	-	-	-	-	-	-	-	X	-	-
272	12,543,645	46.2	-	-	-	-	X	-	-	-	X	-	-

4.3 Mapping score

Mapping scores are assigned by the alignment program to indicate the likelihood of false alignment and stored in the **<MAPQ>** field of BAM files. Higher score usually means longer alignment, less mismatch, and/or higher uniqueness.

Mapping score summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00   70.00   45.49   70.00   70.00
```

4.3.1 Mapping score categories

Score	Count	Percentage
mapq=0	33,367	26.66
mapq>=1	91,794	73.34
mapq>=2	89,353	71.39
mapq>=3	88,212	70.48
mapq>=4	86,275	68.93
mapq>=5	86,238	68.90
mapq>=10	85,976	68.69
mapq>=20	84,324	67.37
mapq>=30	80,873	64.62
mapq>=40	80,100	64.00
mapq=70	76,440	61.07

Table 7: **Mapping score categories.** The total number and percentage of reads having mapping scores equal to or higher than given values.

4.3.2 Overall score distribution

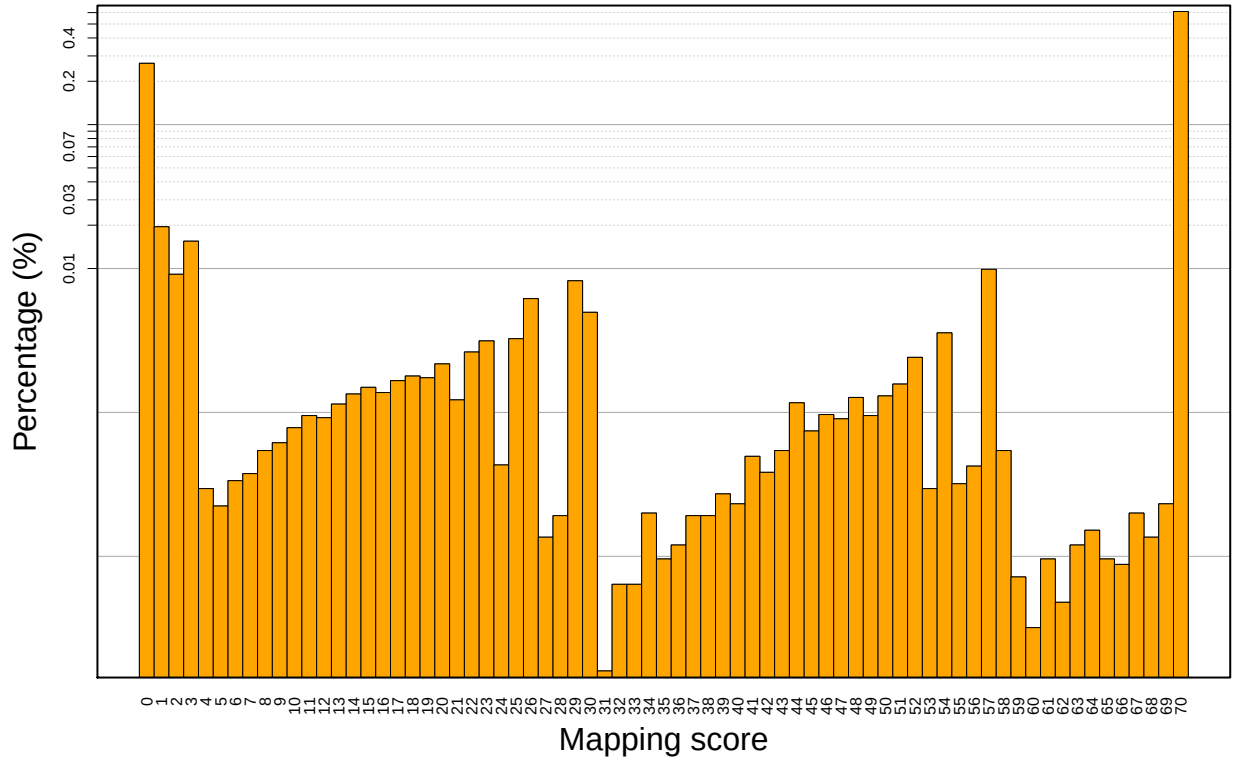


Figure 5: **Mapping score distribution.** By definition, mapping quality equals to -10*Log10(p value), where p value is the likelihood of incorrect mapping; however, its calculation depends on individual programs.

4.4 Mismatch (CIGAR)

SAM uses the **<CIGAR>** field to compactly represent alignments. CIGAR characters are used in concert with lengths to describe various types of matching, mismatching, clipping, padding and splicing events within an alignment.

4.4.1 Mismatch categories

Bitwise	Description
M	alignment match (can be a sequence match or mismatch)
I	insertion to the reference
D	deletion from the reference
N	skipped region from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch

Category	Count	Percentage
M	2,063,686	100.00
I	18,302	0.89
D	8,832	0.43
S	158,755	7.69
H	68,351	3.31

Table 8: **Mismatch categories** The total number and percentage of reads having specific types of mismatches.

4.4.2 Gapped alignment

Not reads having gapped alignment.

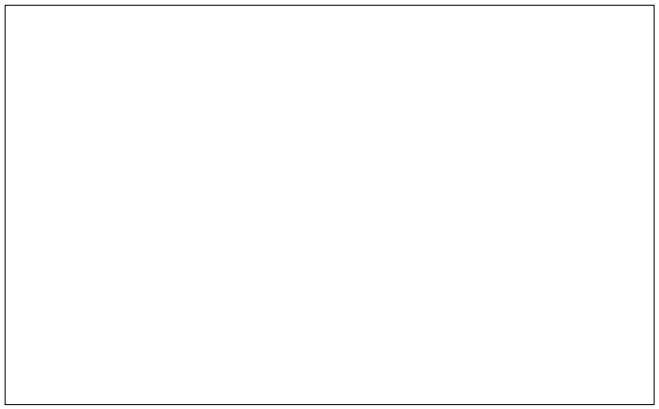


Figure 6: **Distribution of gap size.** If the alignment program tried to align subsequence of the same read to remote locations, **<CIGAR>** will provide the size of gapped regions

4.5 Duplicated mapping

Duplicated mapping refers to multiple reads having their first base mapped to the same strand and location. Duplication level is the number of reads sharing the same duplicated mapping. It is an indicator of the effect of PCR artifact, but also depends on local and overall sequencing depth.

4.5.1 Duplication level categories

The average number of duplicated reads at each mapping location is 1.113.

Level	Location_count	Read_count	Percentage
1	1,787,091	1,787,091	86.597
2	56,725	113,450	5.497
3	4,185	12,555	0.608
4	1,281	5,124	0.248
5	698	3,490	0.169
6	472	2,832	0.137
7	281	1,967	0.095
8	248	1,984	0.096
9	174	1,566	0.076
10	151	1,510	0.073
>10	2,261	132,117	6.402

Table 9: **Duplication level categories.** Numbers of mapping locations and reads having the duplication levels of the given values.

4.5.2 Overall duplication distribution

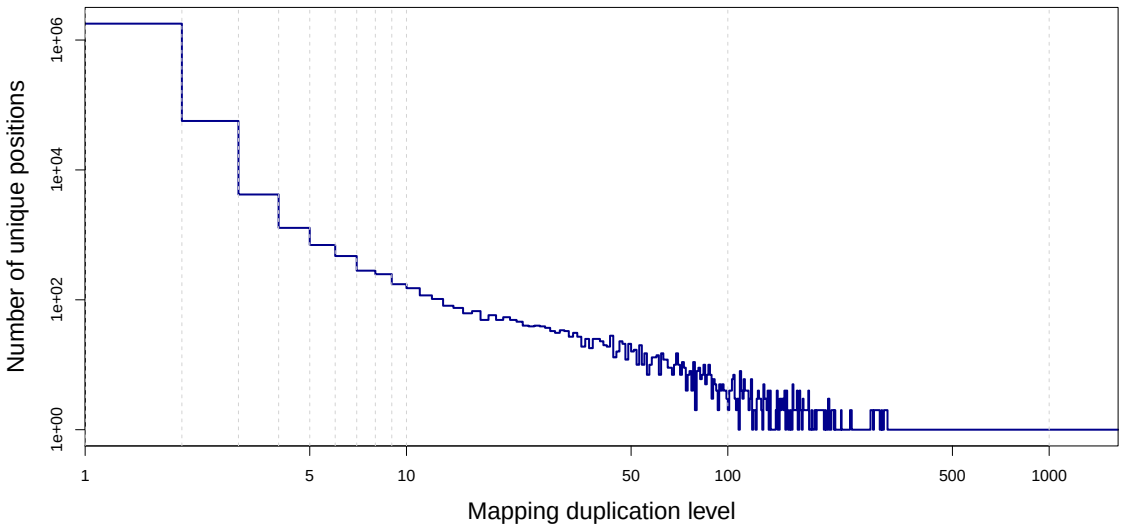


Figure 7: textbfDistribution of duplication levels. The x-axis indicates the number of reads sharing the same mapping location of their 5'-end and the y-axis is the total occurance of each level. Only reads mapped to the forward strand and the first 10 million reads of each chromosome was used to reduce computation.

4.6 Paired reads

No information about paired-end reads is available in this BAM file.

4.6.1 Read count summary

Not applicable.

Category	Count	Percent
Total paired-end reads	0.00	0.00

Table 10: **Paired-end reads.** Read counts in this table are based on the "flag" field in BAM file. Properly mapping paired-end reads are reads mapped to the opposite strand of the same chromosome.

4.6.2 Insertion size of paired reads

Not applicable.

Figure 8: **Distribution of insertion size.** Insertion size is the distance between the mapping locations of the 5'-end of paired reads. It represents the size of DNA fragment to be sequenced.

5 Base frequency

This section summarizes the frequency of nucleic acid bases within sequencing reads in order to identify sequencing bias.

5.1 Base N frequency

	Total	N	Percentage
Base	6,122,337	191	0.0031
Read	125,161	179	0.1430

Table 11: **N base frequency.** The Ns in the reads are assigned by the sequencing machine to suggest that the base cannot be determined due to low quality or other reasons. This table shows the number and percentage of Ns and reads including any Ns. Ns are then excluded from the following analyses of base frequency.

5.2 Expected vs. observed frequency

	A	C	G	T	GC
Expected(%)	29.51	20.47	20.48	29.55	40.94
Observed(%)	26.41	23.51	23.40	26.69	46.91
Observed/Expected(%)	89.49	114.87	114.26	90.32	114.57

Table 12: **Expected vs. observed base frequency.** The expected base frequency is based on the whole reference genome and the observed frequency is the base frequency in sequencing reads. Their ratio reflects the sequencing bias of nucleic acid bases.

5.3 GC content

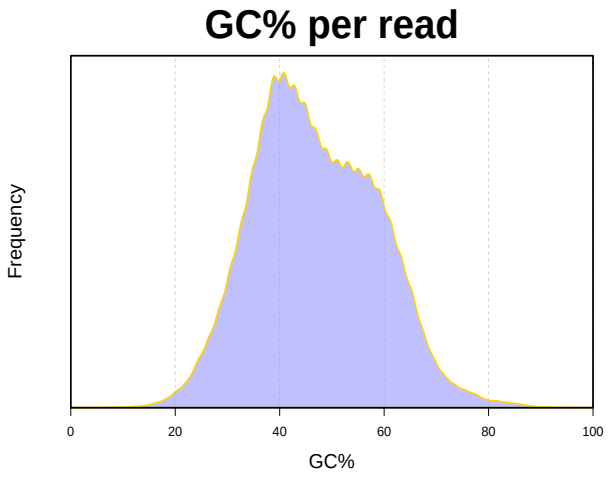


Figure 9: **GC content.** Percentage of C/G bases within each read.

5.4 Position-specific base frequency

Position-specific frequency of bases indicates whether there is a sequencing bias at both ends of the reads. The bias can be introduced via a variety of sources, such as DNA fragmentation and primer contamination.

5.4.1 Single base

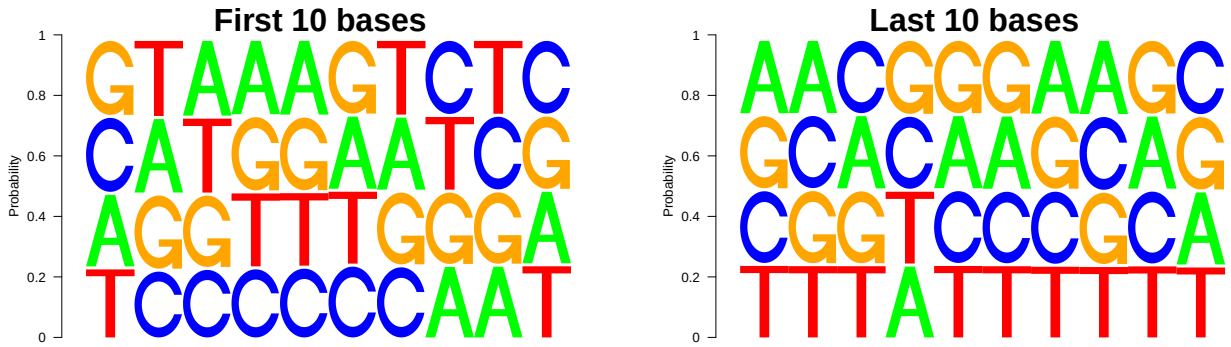


Figure 10: **Single base frequency at both ends.**The base frequency of the first and last 10 bases (the rightmost is the last base) of reads. The frequency was normalized by the overall base frequency with sequencing reads, so this summary indicates the preference of sequencing to start with a given nucleic acid base.

5.4.2 First two bases

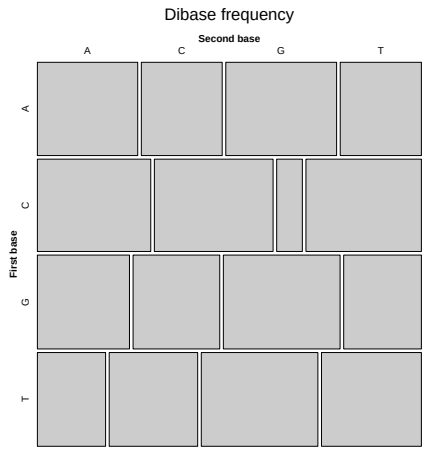


Figure 11: **First two base combination.**This plot summarizes the frequency of the two-base combinations at the 5'-end of reads. The size of the blocks represent their relative frequency after adjusted by their expected frequency based on the position-specific frequency of the first two bases.

5.4.3 5-mer frequency

The frequency of 5-mer at both ends of reads.

Table 13: Lowest frequency

5-mer	Expected_count	Observed_count	Observed/Expected
|CGCGT	101.50	1	0.0099
|CGACG	103.42	2	0.0193
|CGTCG	103.45	3	0.0290
|TCGGT	100.79	3	0.0298
CGACG|	102.08	4	0.0392
TACGA|	131.82	4	0.0303
|TCGAC	95.07	5	0.0526
ACGCG|	100.44	5	0.0498
|TCGTA	115.20	6	0.0521
|GTACG	120.59	6	0.0498

Table 14: Highest frequency

5-mer	Expected_count	Observed_count	Observed/Expected
|TTTTT	175.40	608	3.47
AAAAA|	167.20	565	3.38
TTTTT|	155.53	510	3.28
|AAAAA	176.87	458	2.59
|CCCAG	90.94	427	4.70
CTGGG|	99.25	391	3.94
|CTGGG	109.91	381	3.47
|ATTTT	182.77	358	1.96
|AGAGA	138.04	358	2.59
AAAAT|	164.00	354	2.16

Table 15: Highest relative enrichment

5-mer	Expected_count	Observed_count	Observed/Expected
|CCCAG	90.94	427	4.70
CTGGG|	99.25	391	3.94
|CTGGG	109.91	381	3.47
|TTTTT	175.40	608	3.47
AAAAA|	167.20	565	3.38
GAGGC|	103.94	350	3.37
CCTCC|	101.43	341	3.36
|CCAGG	100.18	332	3.31
TTTTT|	155.53	510	3.28
CCCAG|	102.49	328	3.20
GGAGG|	104.02	321	3.09
CCAGG|	103.22	318	3.08
|GGAGG	110.16	336	3.05
|CCTCC	89.09	270	3.03
GCCTC|	104.95	317	3.02
GCTGG|	101.64	306	3.01

6 ChIP-seq

This section of the report summarizes information related to a ChIP-seq experiment.

6.1 Strand-strand correlation

Since sequencing usually starts from the 5-prime end of DNA fragments, reads mapped to the forward and reverse strands were skewed to the left and right respectively. While we expect a positive correlation between the two strands if reads were enriched around ChIP-ed regions, the forward strand needs to be shifted towards the right, or vice versa, to achieve the maximal strand correlation. The assoication between correlation coefficients and numbers of bases to shift indicates the distribution of DNA fragment sizes.

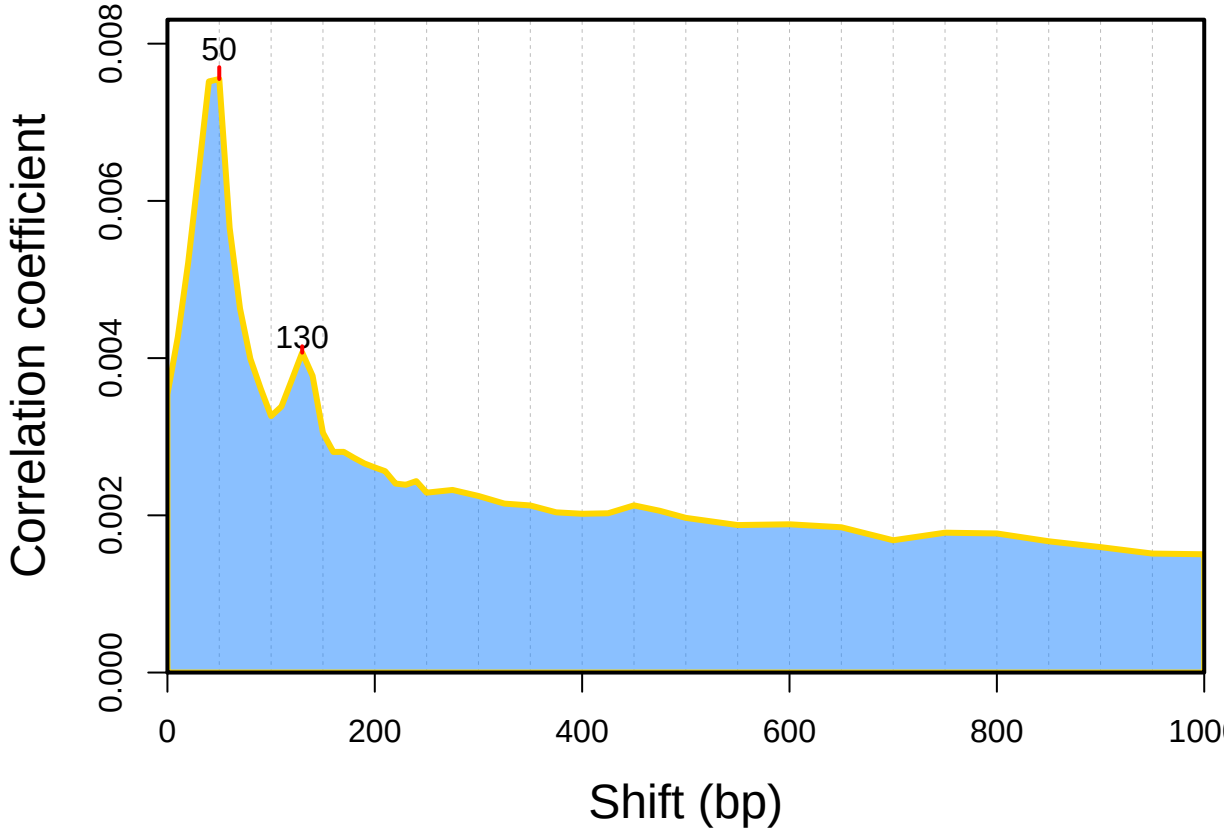


Figure 12: **Strand-strand correlation vs. shifting.**The x-axis is the number of bases to shift the forward strand towards the right. Correlation was calculated between the 5-prime end of mapping locations after removing duplicated mappings.

6.2 Peaks

This section of the report is a quick summary of peaked regions without using specifically designed peak calling program. All reads were extended to base pairs at the 3-prime end. A peak is defined here as a continuous region with at least 1X depth.

Height	Count	Average_width
>=10	976	1,154.23
>=25	321	1,927.70
>=50	154	2,452.29
>=100	69	2,402.51
>=200	39	2,468.41
>=500	27	979.15
>=1,000	24	1,001.21
>=5,000	6	1,229.33
>=10,000	3	592.00
=14,985	1	632.00

Table 16: **Peak summary.** Numbers of peaks with given depth and their average width.

6.2.1 Peak height

Peak height is the maximal sequencing depth within a peak.

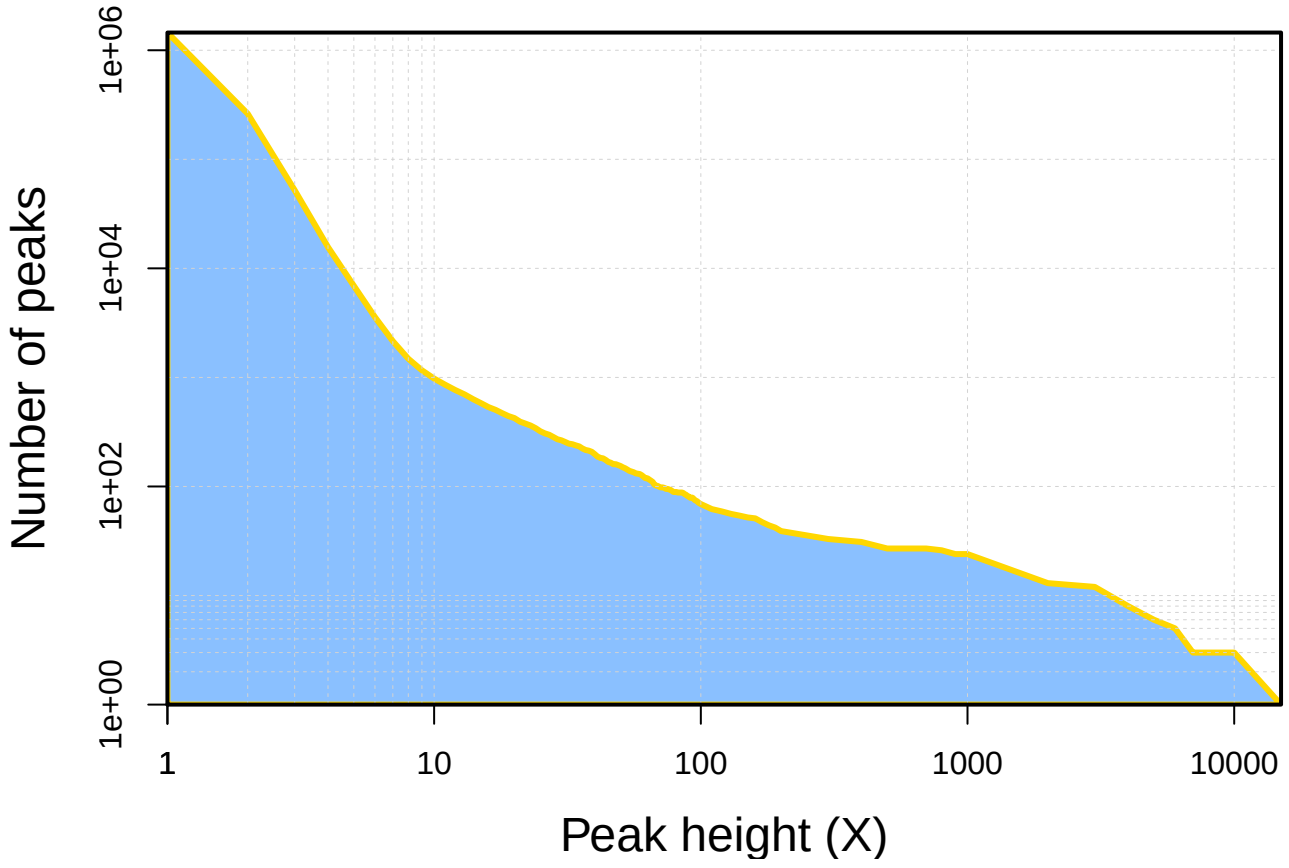


Figure 13: **Peak height distribution.**

6.2.2 Peak width

Peak width is the size of a continuous region with a minimum of 1X depth.

Peak width summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  200.0   200.0   200.0   220.8   200.0 30110.0
```

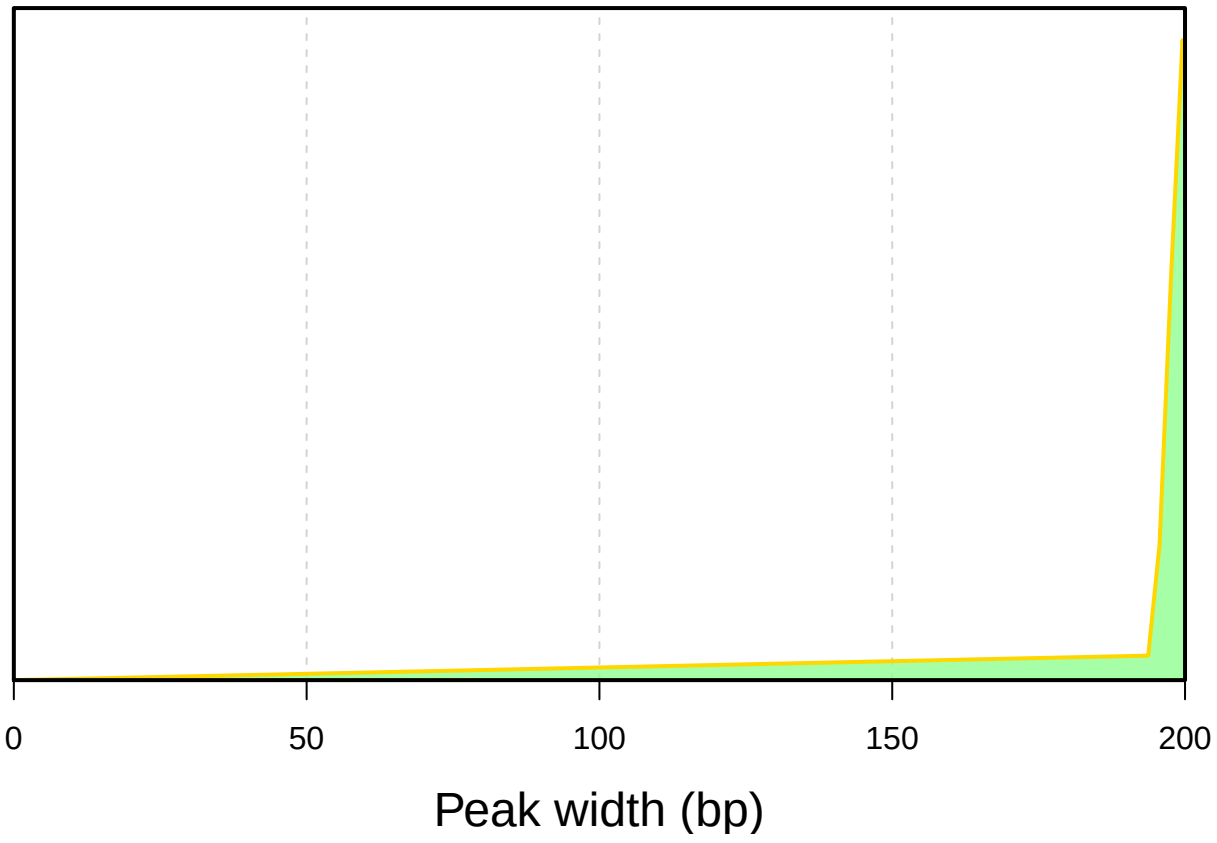


Figure 14: **Peak width distribution.**

6.2.3 Peak frequency by genomic feature

Table 17: Number of peaks mapped to genomic features.

Feature	Promoter	5-UTR	Coding	Intron	3-UTR	Exon
Height>=10	38	34	104	248	7	124
Height>=25	7	1	15	71	3	24
Height>=50	5	0	5	38	1	12
Height>=100	4	0	1	24	1	6
Height>=200	2	0	0	11	0	2
Height>=500	2	0	0	11	0	2
Height>=1000	2	0	0	11	0	2
Height>=5000	1	0	0	3	0	1
Height>=10000	0	0	0	2	0	0
Height>=14985	0	0	0	1	0	0

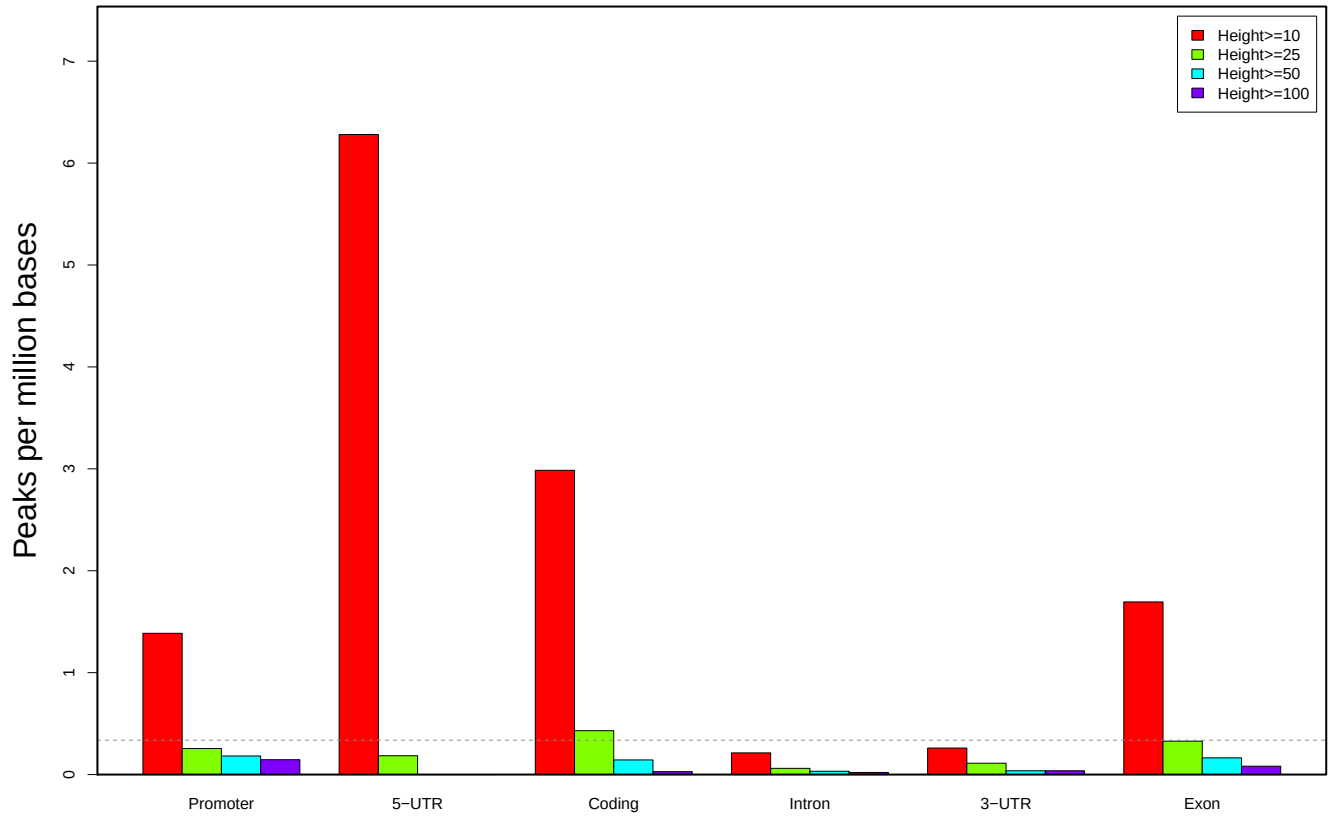


Figure 15: **Peak frequency within genomic features.**The dashed line is the overall frequency of peaks with depth no less than 10 within the whole genome.

6.2.4 Top peaks

Table 18: Top 20 peaks with the highest height.

Chromosome	Start	End	Width	Height
chr8	70,602,155	70,602,786	632	14,985
chr1	91,852,607	91,853,314	708	11,185
chr4	70,296,491	70,296,926	436	11,130
chr2	133,011,216	133,013,962	2,747	6,719
chr21	9,825,283	9,827,730	2,448	6,514
chr19	24,183,920	24,184,324	405	5,787
chrX	108,297,196	108,298,002	807	4,813
chr12	20,704,189	20,704,656	468	4,758
chr19	36,066,333	36,066,921	589	3,590
chr5	174,541,594	174,542,319	726	3,387
chr11	77,597,130	77,597,846	717	3,385
chr16	33,962,396	33,966,603	4,208	3,016
chr13	31,418,041	31,418,658	618	2,677
chr1	145,277,150	145,277,647	498	1,912
chr14	90,341,021	90,341,601	581	1,831
chr1	120,543,552	120,544,208	657	1,811
chr2	9,841	10,465	625	1,497
chr1	121,482,686	121,485,607	2,922	1,441
chr2	133,036,355	133,036,887	533	1,429
chr1	237,766,173	237,766,699	527	1,388

6.3 TSS

This section of the report summarizes sequencing depth around transcription start sites (TSS).

6.3.1 Strand-specific depth around TSS

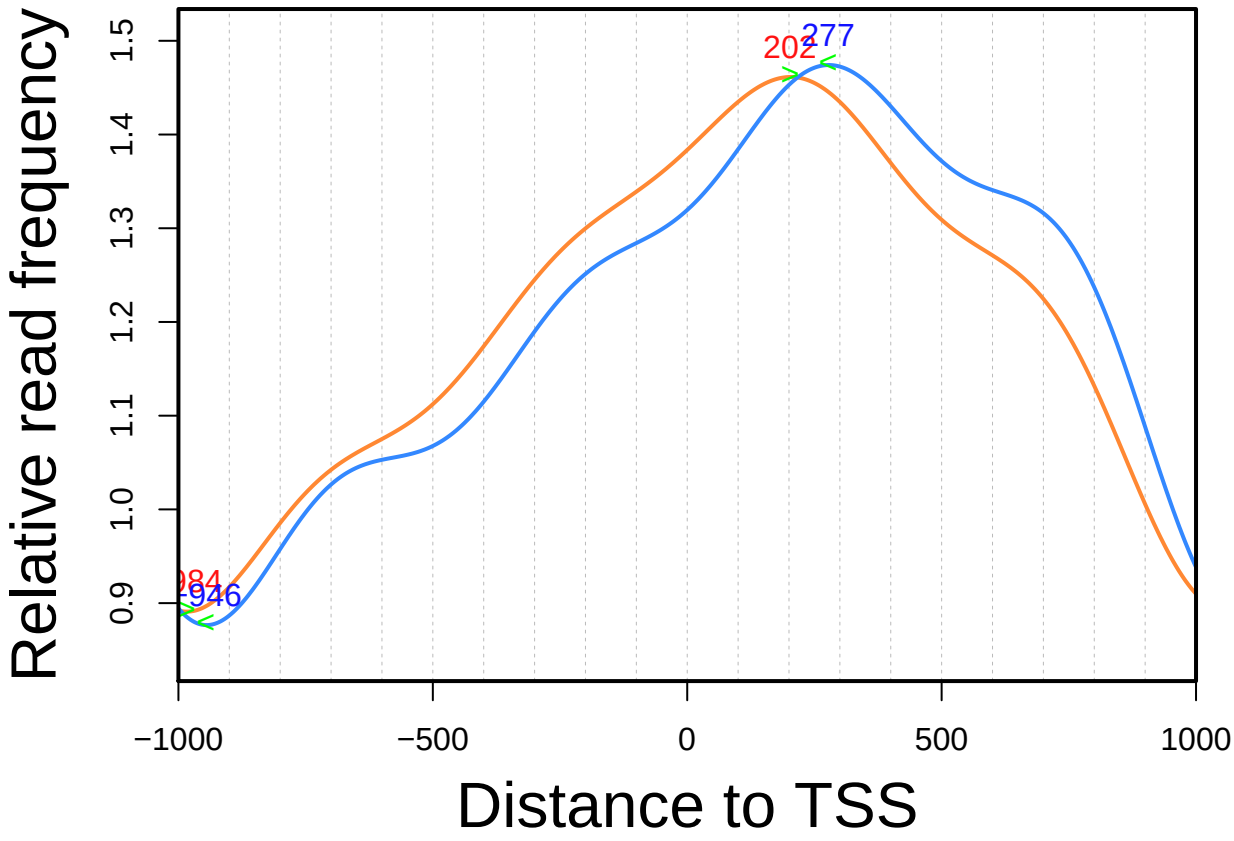


Figure 16: **Read frequency around TSS.**This plot shows the frequency of reads whose 5-prime end was mapped around TSS of RefSeq genes. The read counts were normalized by the global average after duplicated mapping was not removed.

6.3.2 Read counts around individual TSSs

Read counts around TSS of individual genes.

Read count summary:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   3.184   3.000 1080.000
```

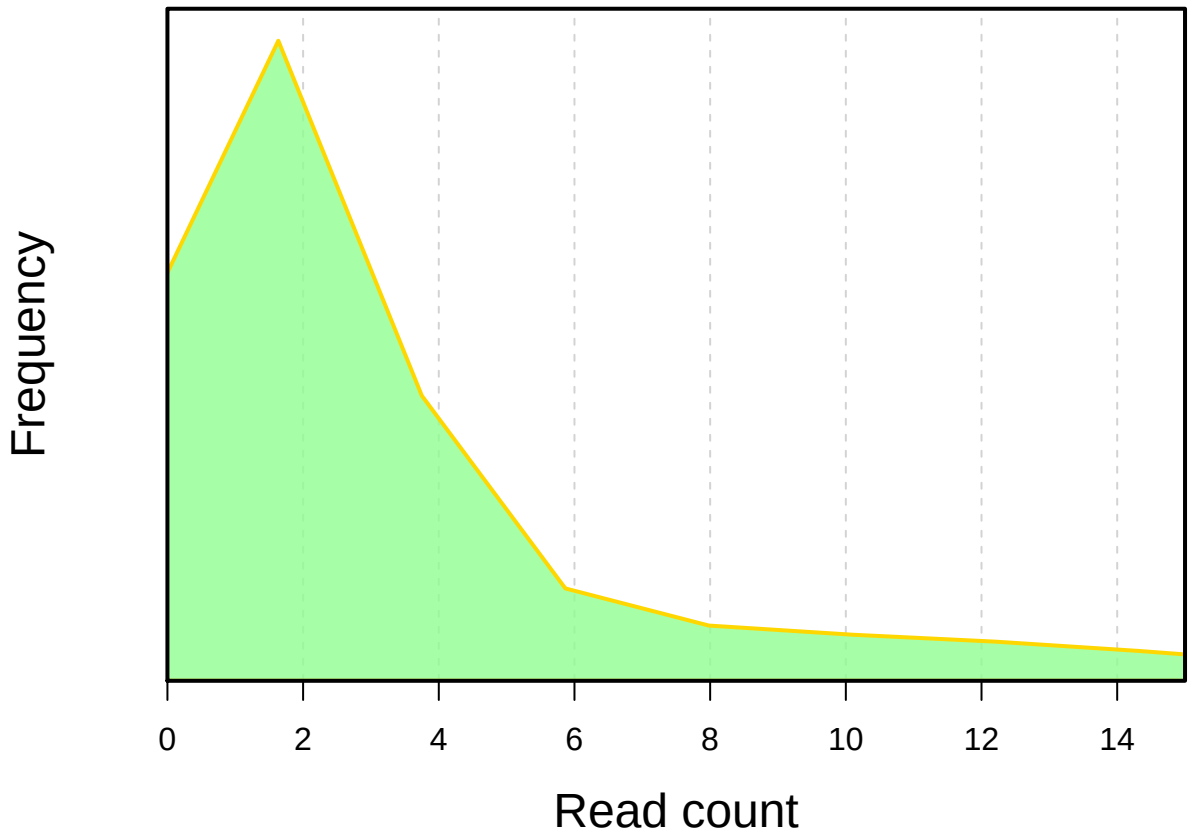


Figure 17: **Read count distribution of genes.**This plot shows the distribution of read counts within the [-1kb, 1kb] region of RefSeq TSSs. Duplicated mapping was excluded.

Table 19: Top 21 genes with the highest read counts around TSS

RefSeq_ID	Sense	Antisense	Total
NR_037458	532	548	1,080
NR_037421	390	376	766
NR_033770	293	329	622
NR_040095	211	200	411
NR_030386	106	120	226
NR_038368	100	108	208
NR_027020	53	35	88
NR_031608	53	34	87
NM_022133	21	18	39
NM_152836	21	18	39
NM_152837	21	18	39
NM_005604	20	14	34
NR_037416	16	15	31
NM_022359	13	18	31
NM_182704	19	10	29
NM_006262	15	14	29
NM_003013	16	13	29
NM_001042544	10	18	28
NR_047517	17	11	28
NM_031912	14	14	28
NM_181519	14	14	28

7 Alerts

— GC contents of reads are more than 110% (114.6%) of expected percentage.
— There are 16 5-mers overrepresented at either end of reads.