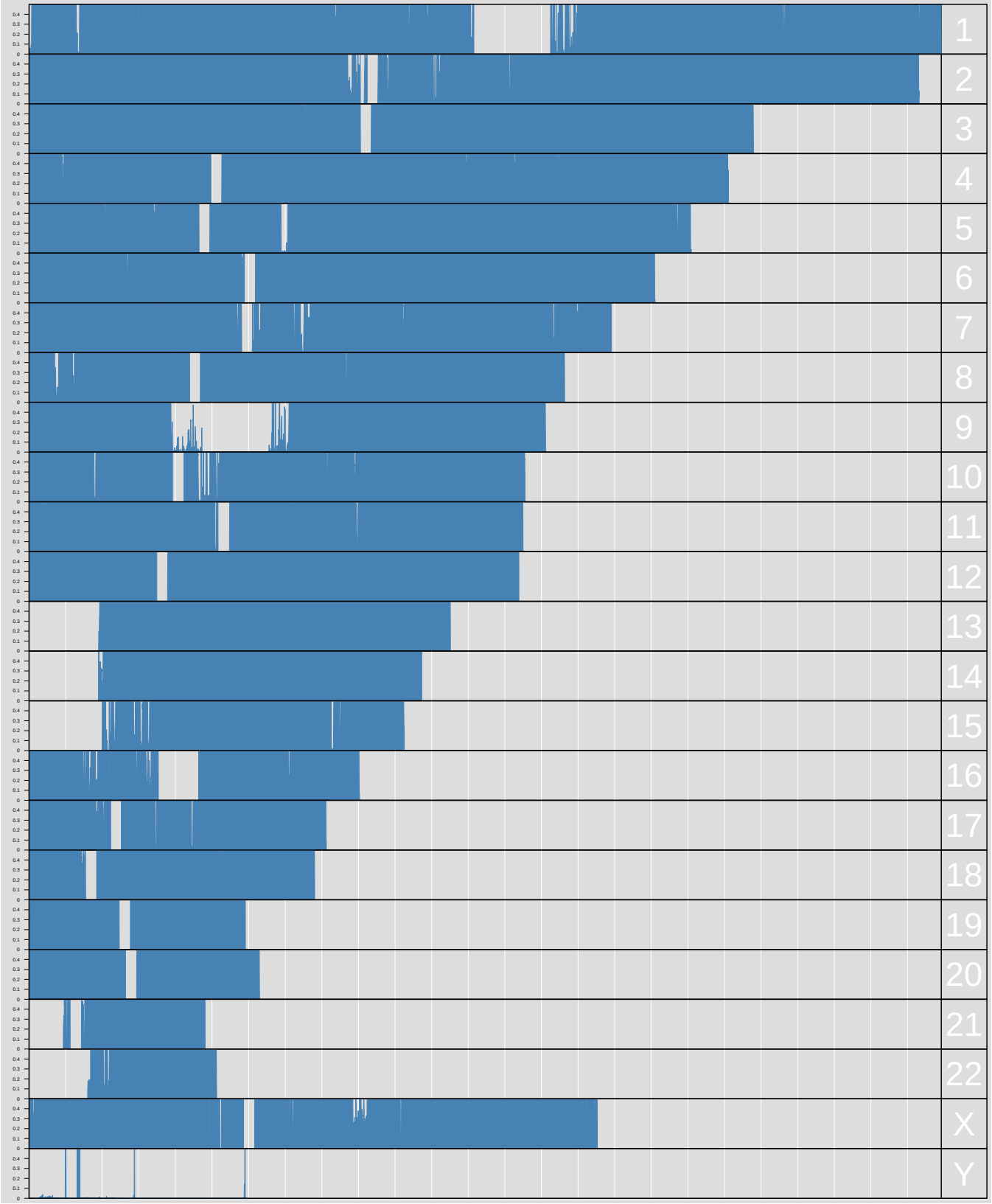
IRF1 decoy, ChIPseq: ND430_M-H3K4me3

Prepared by Zhe Zhang

August 19, 2014

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: ND430_M-H3K4me3

Genome name: hg19

1.1 BAM file

Size: 2.41 GB

Created: 2014-08-11 13:44:23

Modified: 2014-08-11 13:44:23

Location: > nas > is1 > SyncIQ > seqrepo > raw > Sullivan > CPF1406002 > ND430-M-H3K4me3_-sorted.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	41,434,758
Total mapped reads	41,434,758
Total upmapped reads	0
Total mappings	41,434,758
Total mapping locations	36,955,392
Base N%	0.0026
(G+C)%	43.53
Mapped to forward strand%	49.98
Duplicated mapping reads%	19.18
Best sequencing quality	41
Average sequencing quality	29.89
Maximum mapping length (bp)	50
Minimum mapping length (bp)	50
Average mapping length (bp)	50
Best mapping quality	70
Average mapping quality	66.13
Highest sequencing depth	25,052
Average sequencing depth	0.72
Mapped reads per kilobase	14.3

Table 1: **Summary statistics**
Effective size: chromosome length without assembly gaps.
Sequencing quality score: assigned by the re-sequencing 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	1,596,809,733	55.81
Depth>=1	1,264,522,873	44.19
Depth>=5	21,546,461	0.75
Depth>=10	2,037,706	0.07
Depth>=20	254,799	0.01
Depth>=30	156,999	0.01
Depth>=50	107,781	0.00
Depth>=100	63,952	0.00
Depth>=1000	6,315	0.00
Depth>=10000	426	0.00
Depth=25052	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	Effecitive_size	Total_reads	Unique_mapping	Average_depth	Maximum_depth	Maximum_location
1	249,250,621	225,280,621	3,384,202	2,963,974	0.75	25,052	120,475,196
2	243,199,373	238,207,373	3,447,076	3,078,225	0.72	8,586	89,390,319
3	198,022,430	194,797,140	2,776,069	2,526,386	0.71	1,050	193,460,450
4	191,154,276	187,661,676	2,667,470	2,360,963	0.71	10,497	48,994,085
5	180,915,260	177,695,260	2,488,091	2,265,395	0.70	297	46,304,704
6	171,115,067	167,395,067	2,429,128	2,185,453	0.72	973	58,669,263
7	159,138,663	155,353,663	2,207,972	1,983,445	0.71	2,999	58,669,050
8	146,364,022	142,888,922	2,042,117	1,845,034	0.71	3,448	42,982,912
9	141,213,431	120,143,431	1,613,007	1,461,558	0.67	950	52,143,120
10	135,534,747	131,314,747	2,180,445	1,741,310	0.82	16,903	39,019,913
11	135,006,516	131,129,516	1,934,055	1,748,569	0.73	2,040	51,162,611
12	133,851,895	130,481,895	1,899,914	1,726,286	0.72	397	34,686,941
13	115,169,878	95,589,878	1,334,620	1,217,240	0.70	28	26,543,847
14	107,349,540	88,289,540	1,269,785	1,154,665	0.72	404	38,424
15	102,531,392	81,694,769	1,160,044	1,053,425	0.71	69	59,718,929
16	90,354,753	78,884,753	1,243,630	1,078,909	0.78	2,100	35,043,689
17	81,195,210	77,795,210	1,208,133	1,087,444	0.77	1,314	22,053,217
18	78,077,248	74,657,248	1,082,174	966,664	0.72	2,783	15,410,292
19	59,128,983	55,808,983	980,453	858,980	0.87	4,888	24,422,247
20	63,025,520	59,505,520	925,813	835,364	0.77	557	26,507,801
21	48,129,895	35,108,702	540,288	480,518	0.76	1,569	1,058,386
22	51,304,566	34,894,566	530,025	474,899	0.75	3,683	661,670
X	155,270,560	151,100,560	2,016,321	1,831,625	0.66	3,163	57,901,412
Y	59,373,566	25,653,566	73,926	29,061	0.13	3,443	9,998,445

2.3 Depth by genomic feature

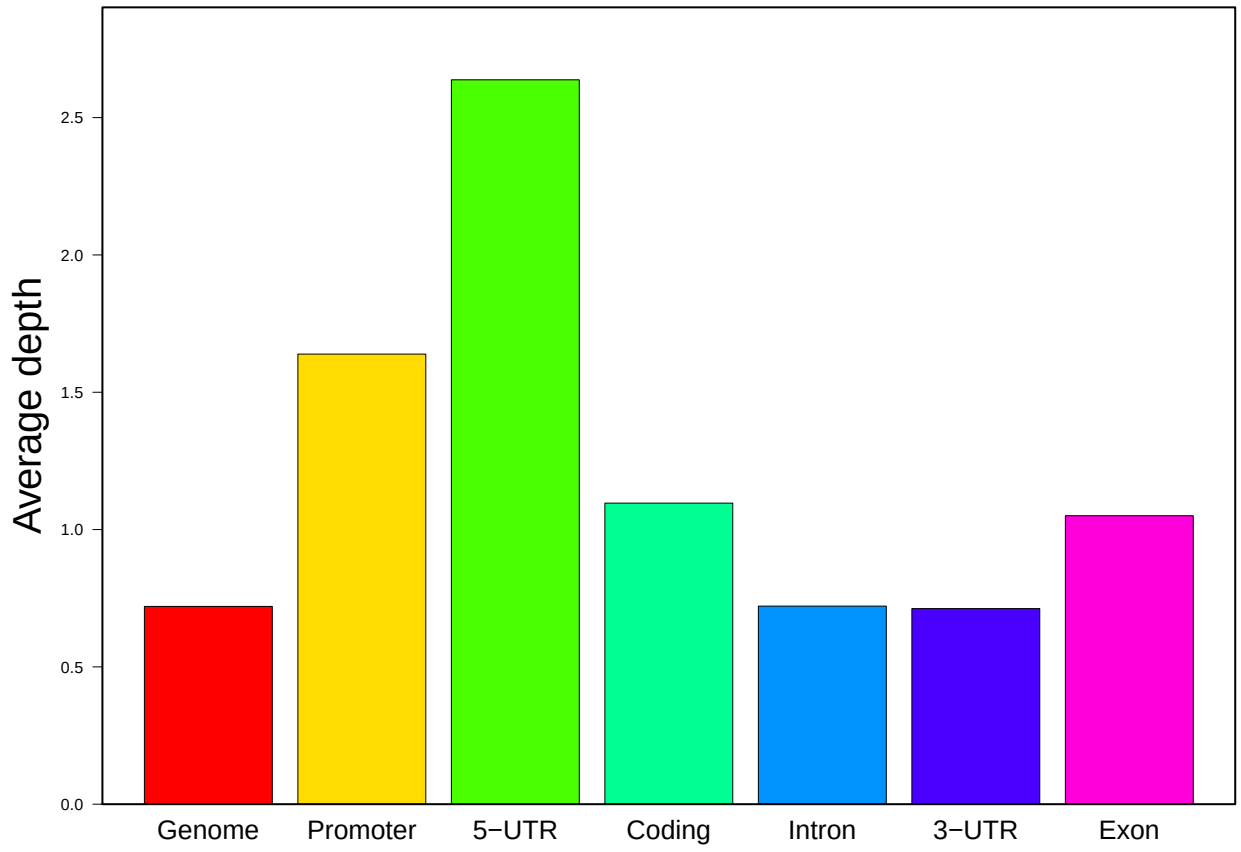


Figure 1: **Average depth of genomic features.** Genomic features are regions annotated based on previous knowledge, such as the RefSeq gene track downloaded from UCSC genome browser. Many applications of high-throughput sequencing technologies, such as exome sequencing and RNA-seq, expect higher depth 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   30.00   30.00   29.89   31.00   41.00
```

3.1 Quality score categories

Score	Count	Percentage
Score=0	0	0.00
Score>=5	7,695,964	99.77
Score>=10	7,677,107	99.53
Score>=13	7,658,173	99.28
Score>=20	7,611,089	98.67
Score>=30	6,440,012	83.49
Score>=40	229	0.00
Score=41	88	0.00

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

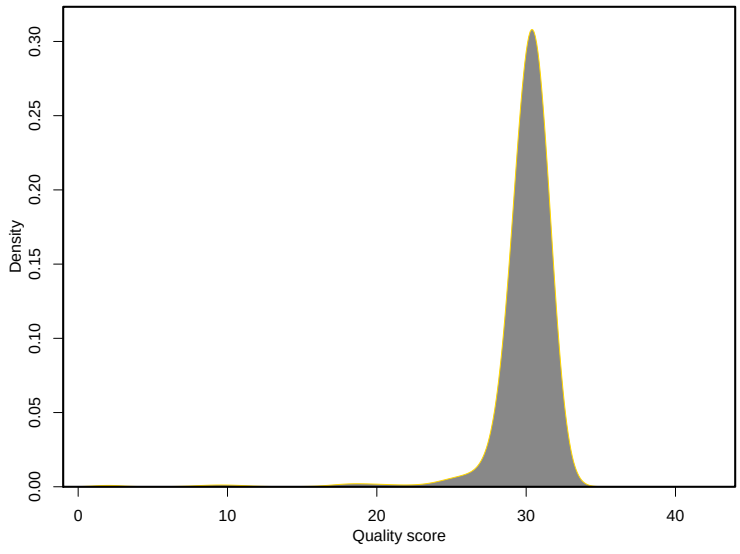


Figure 2: **Score distribution.** This distribution is based on all bases of randomly selected sequencing reads, so position-specific sequencing quality is not considered (see below). The quality scores are calculated by subtracting 33 from the integers corresponding to the ASCII characters in **<QUAL>**. 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

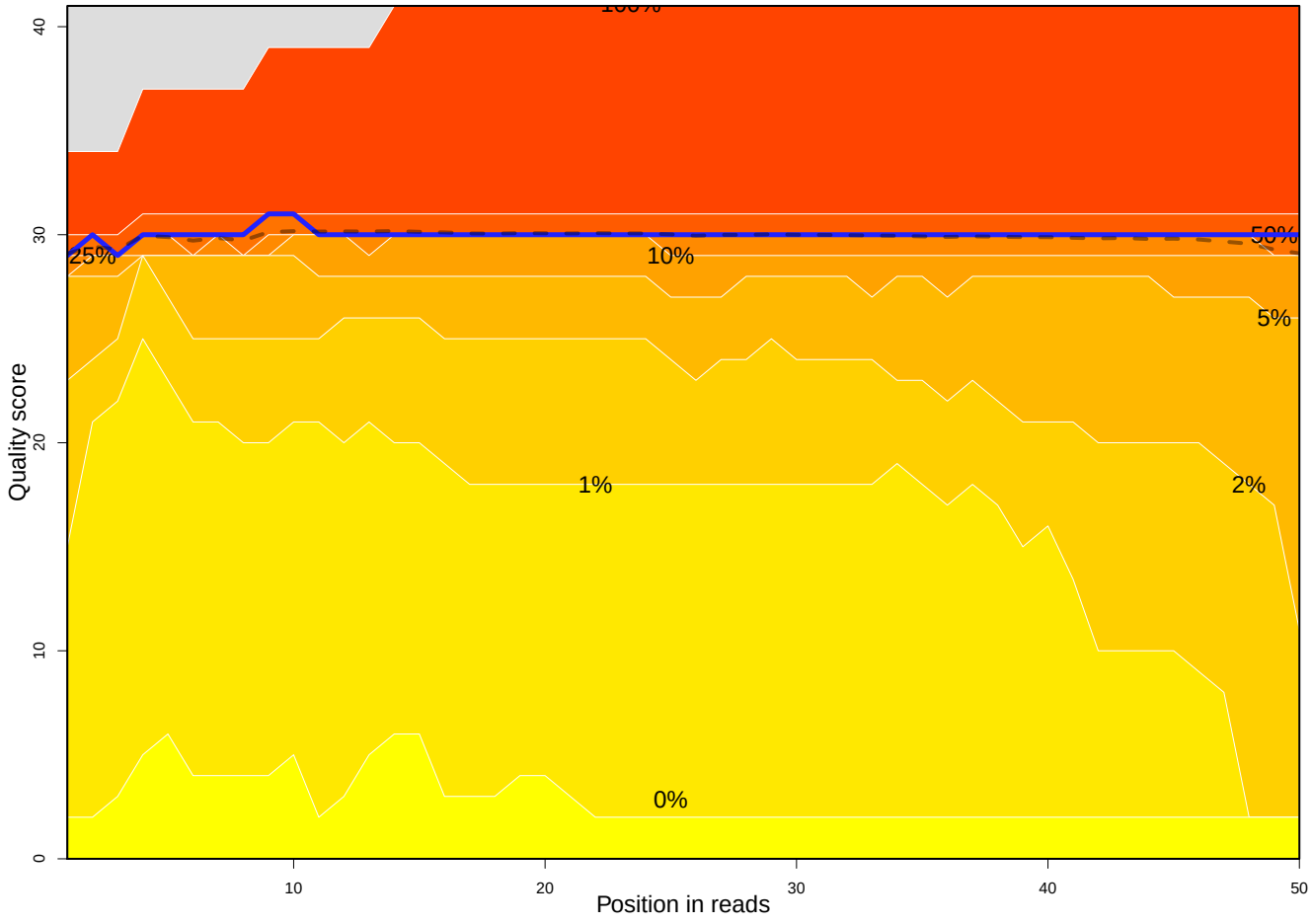


Figure 3: **Position-specific sequencing scores.** This plot shows quality scores at different positions within reads. The dashed lines represents the means of quality scores at different positions; whereas the heat gradient corresponds to percentiles.

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.
  50      50      50      50      50      50
```

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	20,727,281	50.02
0X20	0	0.00
0X40	0	0.00
0X80	0	0.00
0X100	0	0.00
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 flag categories.

Value	Count	Percentage	0X1	0X2	0X4	0X8	0X10	0X20	0X40	0X80	0X100	0X200	0X400
0	20,707,477	49.98	-	-	-	-	-	-	-	-	-	-	-
16	20,727,281	50.02	-	-	-	-	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   70.00   70.00   66.13   70.00   70.00
```

4.3.1 Mapping score categories

Score	Count	Percentage
mapq=0	33	0.02
mapq>=1	154,240	99.98
mapq>=2	154,218	99.96
mapq>=3	154,185	99.94
mapq>=4	154,152	99.92
mapq>=5	154,098	99.89
mapq>=10	153,710	99.64
mapq>=20	151,080	97.93
mapq>=30	145,367	94.23
mapq>=40	144,070	93.39
mapq=70	137,471	89.11

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

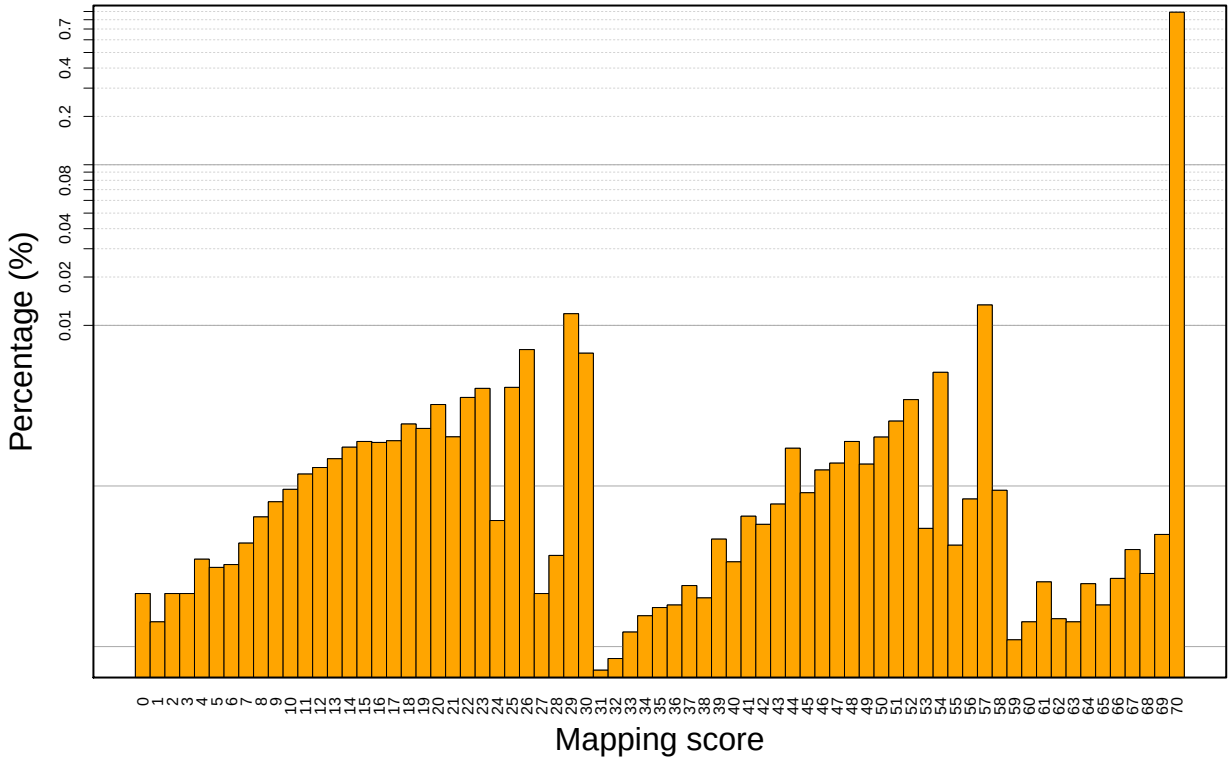


Figure 5: **Mapping score distribution.** By definition, mapping quality equals to -10*Log10(p value), where p value is the likelihood of incorrect mapping; however, its calculation depends on individual programs.

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	41,434,758	100.00
I	140,604	0.34
D	115,779	0.28
S	1,966,527	4.75

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 1.121.

Level	Location_count	Read_count	Percentage
1	33,486,211	33,486,211	80.817
2	3,160,618	6,321,236	15.256
3	251,636	754,908	1.822
4	24,172	96,688	0.233
5	6,063	30,315	0.073
6	3,612	21,672	0.052
7	2,694	18,858	0.046
8	2,200	17,600	0.042
9	1,806	16,254	0.039
10	1,514	15,140	0.037
>10	14,866	655,876	1.583

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

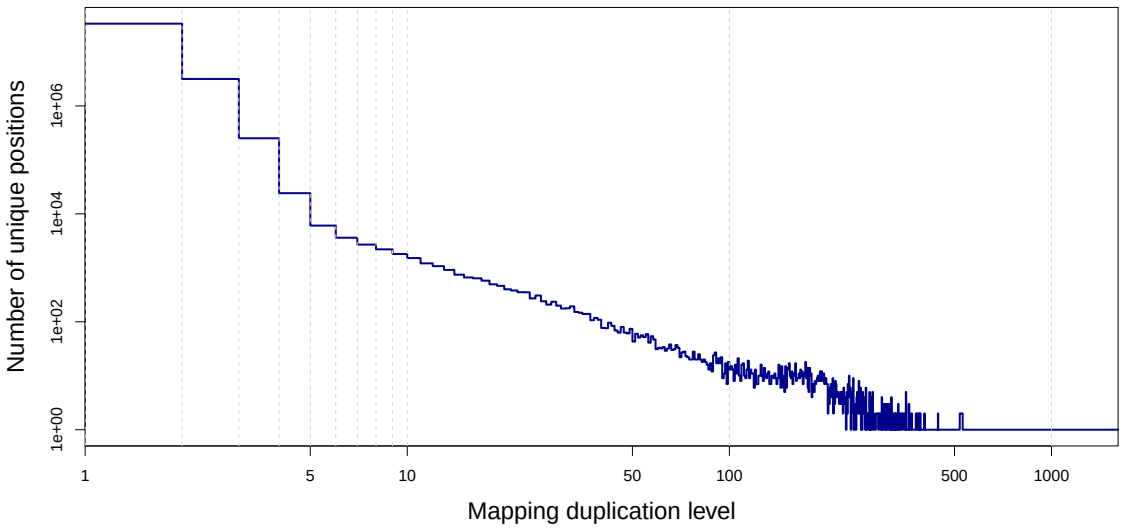


Figure 7: textbfDistribution of duplication levels. The x-axis indicates the number of reads sharing the same mapping location of their 5'-end and the y-axis is the total occurance of each level. Only reads mapped to the forward strand and the first 10 million reads of each chromosome was used to reduce 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	7,713,650	204	0.0026
Read	154,273	179	0.1160

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(%)	28.22	21.96	21.57	28.24	43.53
Observed/Expected(%)	95.64	107.31	105.33	95.59	106.32

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

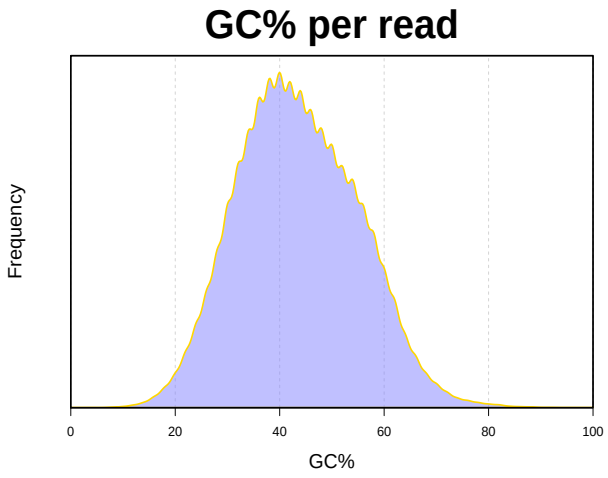


Figure 9: **GC content.** Percentage of C/G bases within each read.

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

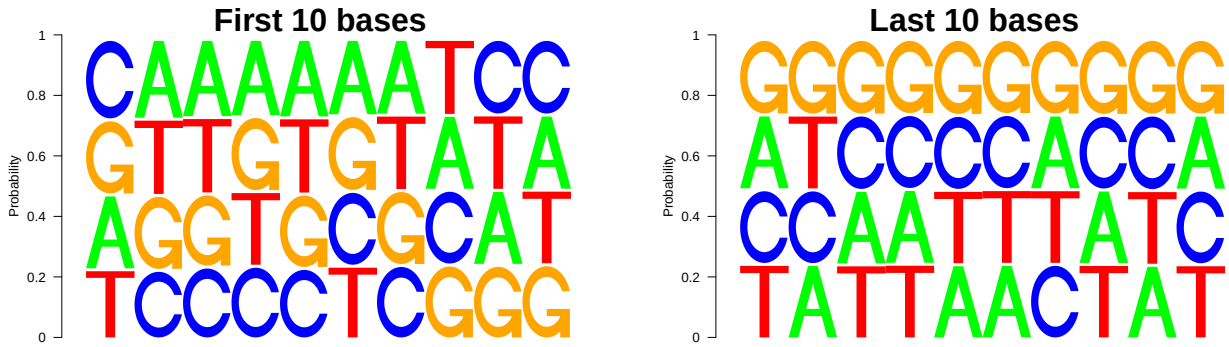


Figure 10: **Single base frequency at both ends.**The base frequency of the first and last 10 bases (the rightmost is the last base) of reads. The frequency was normalized by the overall base frequency with sequencing reads, so this summary indicates the preference of sequencing to start with a given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

5.4.2 First two bases

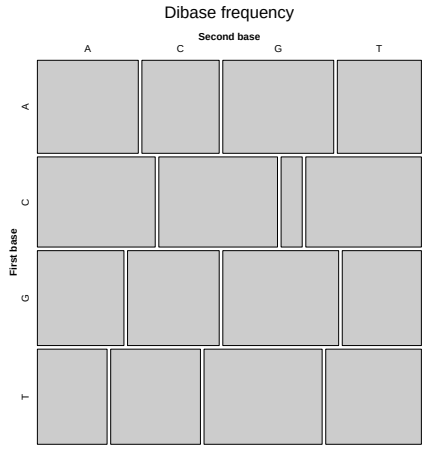


Figure 11: **First two base combination.**This plot summarizes the frequency of the two-base combinations at the 5'-end of reads. The size of the blocks represent their relative frequency after adjusted by their expected frequency based on the position-specific frequency of the first two bases.

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
|TCGCG	76.33	0	0.000
|TCGAC	100.39	2	0.020
|CGCCG	74.73	3	0.040
|CGTCG	102.40	3	0.029
ACGCG|	100.75	4	0.040
|ACCGA	116.23	5	0.043
|TACGC	112.17	5	0.045
|ACGCG	82.79	5	0.060
CGACG|	100.81	5	0.050
GCGAA|	125.81	6	0.048

Table 14: Highest frequency

5-mer	Expected_count	Observed_count	Observed/Expected
|AAAAA	329.74	889	2.70
AAAAA|	269.49	717	2.66
TTTTT|	271.55	709	2.61
|TTTTT	266.48	608	2.28
|AAAAT	320.21	519	1.62
|ATTTT	289.04	496	1.72
ATTTT|	270.76	484	1.79
AAAAT|	266.63	478	1.79
|TAAAA	304.01	473	1.56
|AAATA	318.98	453	1.42

Table 15: Highest relative enrichment

5-mer	Expected_count	Observed_count	Observed/Expected
CCCAG|	98.73	353	3.58
GGAGG|	97.36	322	3.31
|CCAGG	98.85	324	3.28
|CTGGG	101.26	328	3.24
|CCCAG	94.80	302	3.19
|CCTGG	96.58	304	3.15
CCTCC|	101.63	315	3.10
CTGGG|	99.89	308	3.08
|GGCTG	90.34	277	3.07

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.

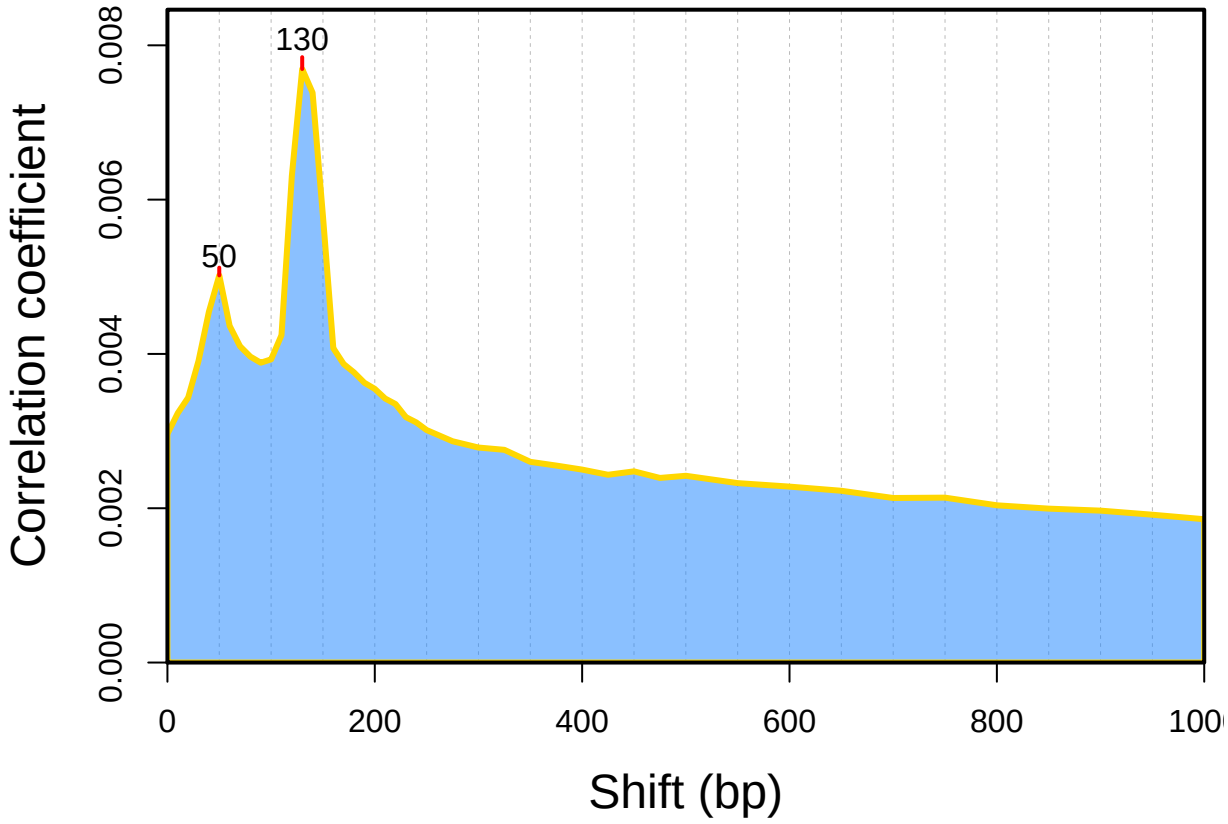


Figure 12: **Strand-strand correlation vs. shifting.**The x-axis is the number of bases to shift the forward strand towards the right. Correlation was calculated between the 5-prime end of mapping locations after removing duplicated mappings.

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 5X depth.

Height	Count	Average_width
>=10	37,287	135.78
>=25	2,734	169.84
>=50	1,452	158.31
>=100	915	177.17
>=200	546	201.73
>=500	245	209.47
>=1,000	104	247.15
>=5,000	16	364.12
>=10,000	7	368.86
=25,417	1	1,175.00

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.

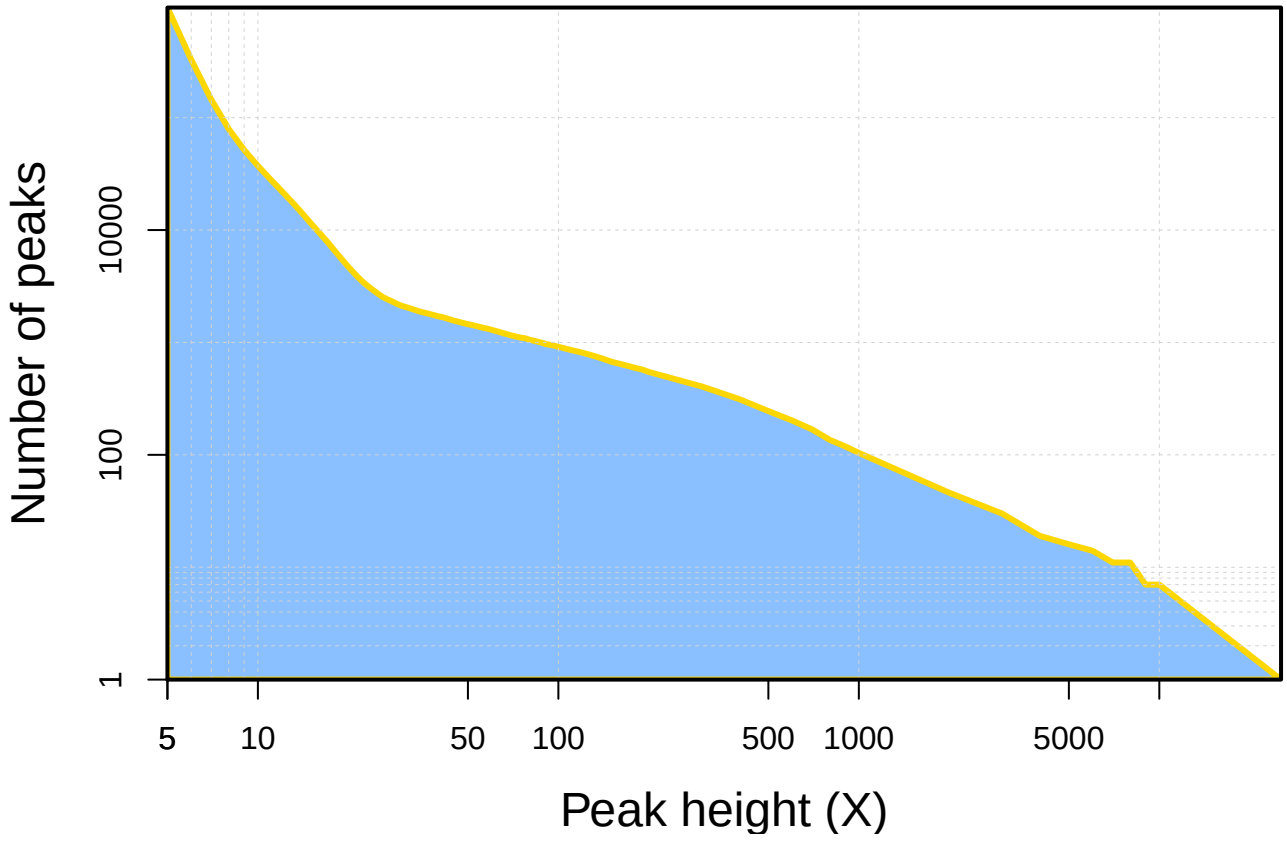


Figure 13: **Peak height distribution.**

6.2.2 Peak width

Peak width is the size of a continuous region with a minimum of 5X depth.

Peak width summary:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    6.00   14.00   22.97   29.00 2850.00
```

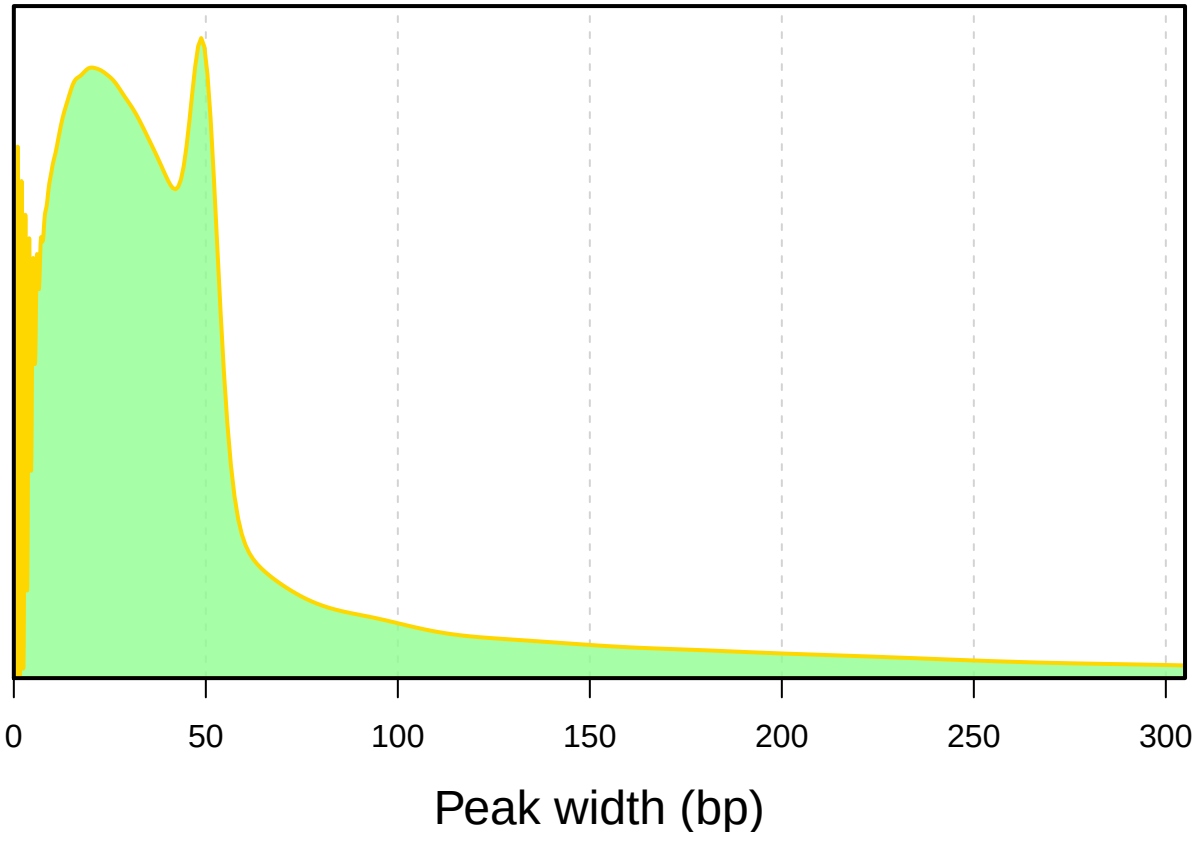


Figure 14: **Peak width distribution.**

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	9,860	7,684	6,911	19,941	282	12,374
Height>=25	133	203	216	347	5	292
Height>=50	5	1	1	26	0	6
Height>=100	3	0	0	9	0	2
Height>=200	0	0	0	6	0	0
Height>=500	0	0	0	3	0	0
Height>=1000	0	0	0	1	0	0

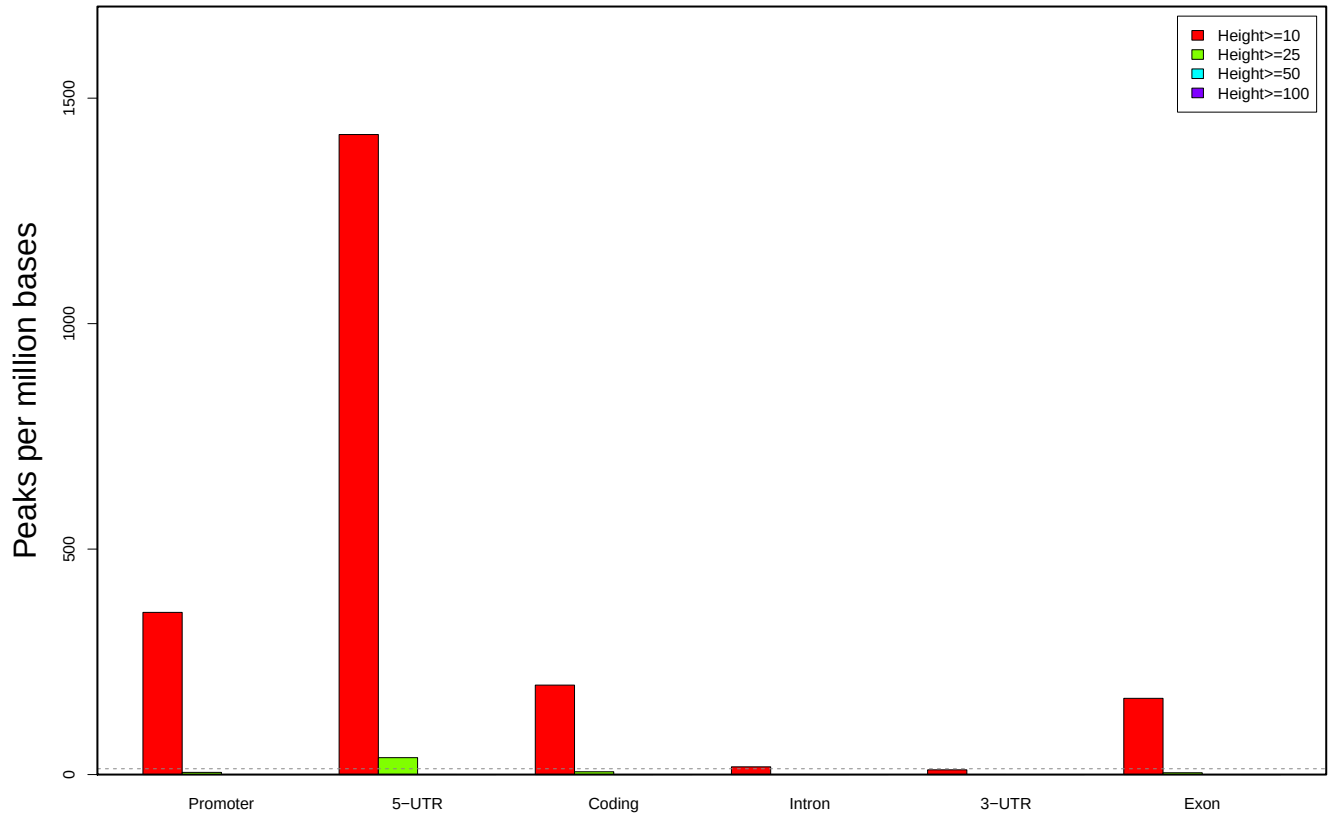


Figure 15: **Peak frequency within genomic features.**The dashed line is the overall frequency of peaks with depth no less than 10 within the whole genome.

6.2.4 Top peaks

Table 18: Top 20 peaks with the highest height.

Chromosome	Start	End	Width	Height
chr1	121,484,284	121,485,458	1,175	25,417
chr10	42,379,661	42,379,983	323	16,904
chr10	42,385,004	42,385,286	283	15,418
chr10	42,392,684	42,392,800	117	14,823
chr10	42,396,788	42,396,985	198	14,551
chr10	42,384,679	42,384,817	139	13,425
chr4	49,151,599	49,151,945	347	10,546
chr10	42,385,356	42,385,602	247	8,854
chr2	89,875,165	89,875,473	309	8,631
chr10	42,395,970	42,396,537	568	8,515
chr10	42,387,006	42,387,508	503	8,460
chr10	42,380,221	42,380,540	320	6,996
chr4	49,150,811	49,151,440	630	6,869
chr10	42,393,361	42,393,497	137	6,566
chr10	42,599,496	42,599,926	431	5,943
chr2	92,269,460	92,269,558	99	5,166
chr10	42,599,928	42,600,255	328	4,963
chr19	27,731,883	27,732,395	513	4,910
chr10	42,596,792	42,597,052	261	4,370
chr22	16,861,621	16,861,721	101	3,684

6.3 TSS

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

6.3.1 Strand-specific depth around TSS

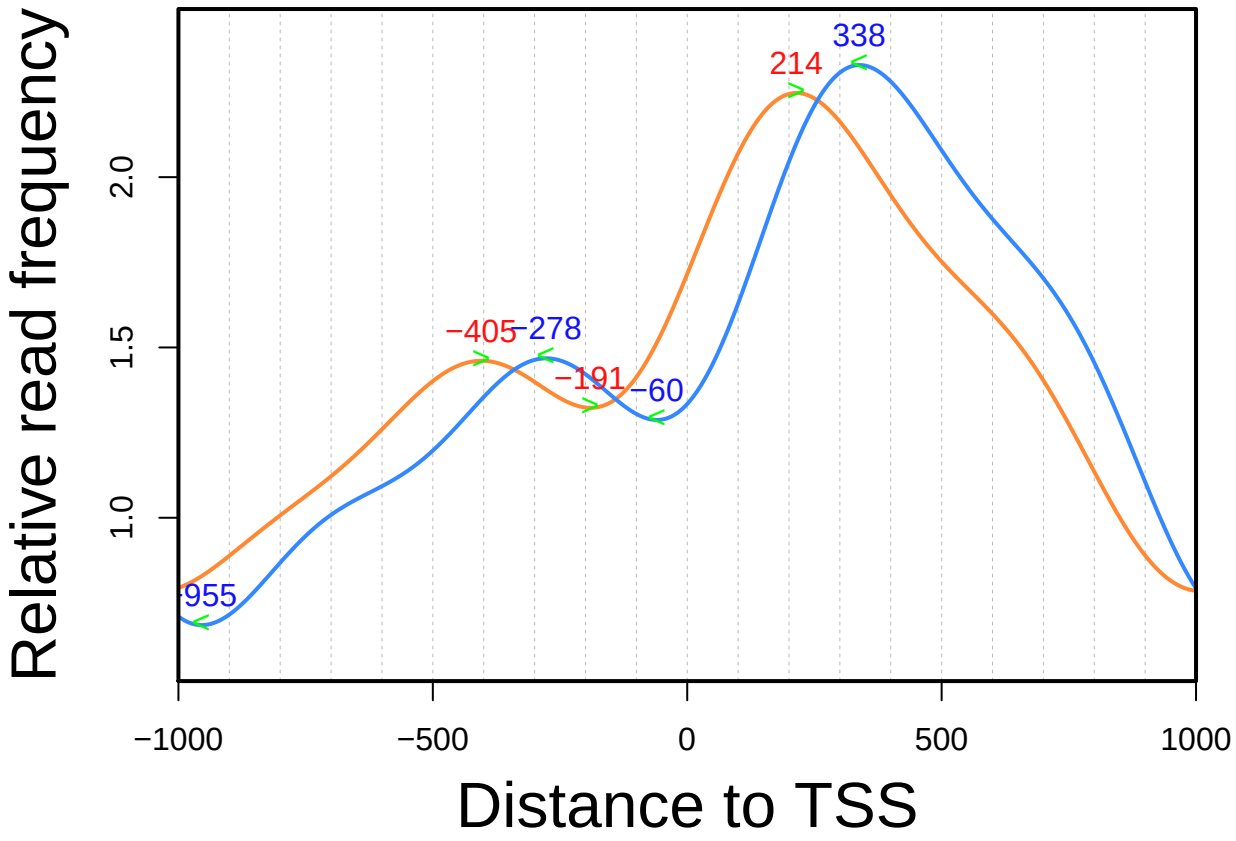


Figure 16: **Read frequency around TSS.**This plot shows the frequency of reads whose 5-prime end was mapped around TSS of RefSeq genes. The read counts were normalized by the global average after duplicated mapping was not removed.

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.
0.00   28.00   55.00   75.85  119.00  464.00
```

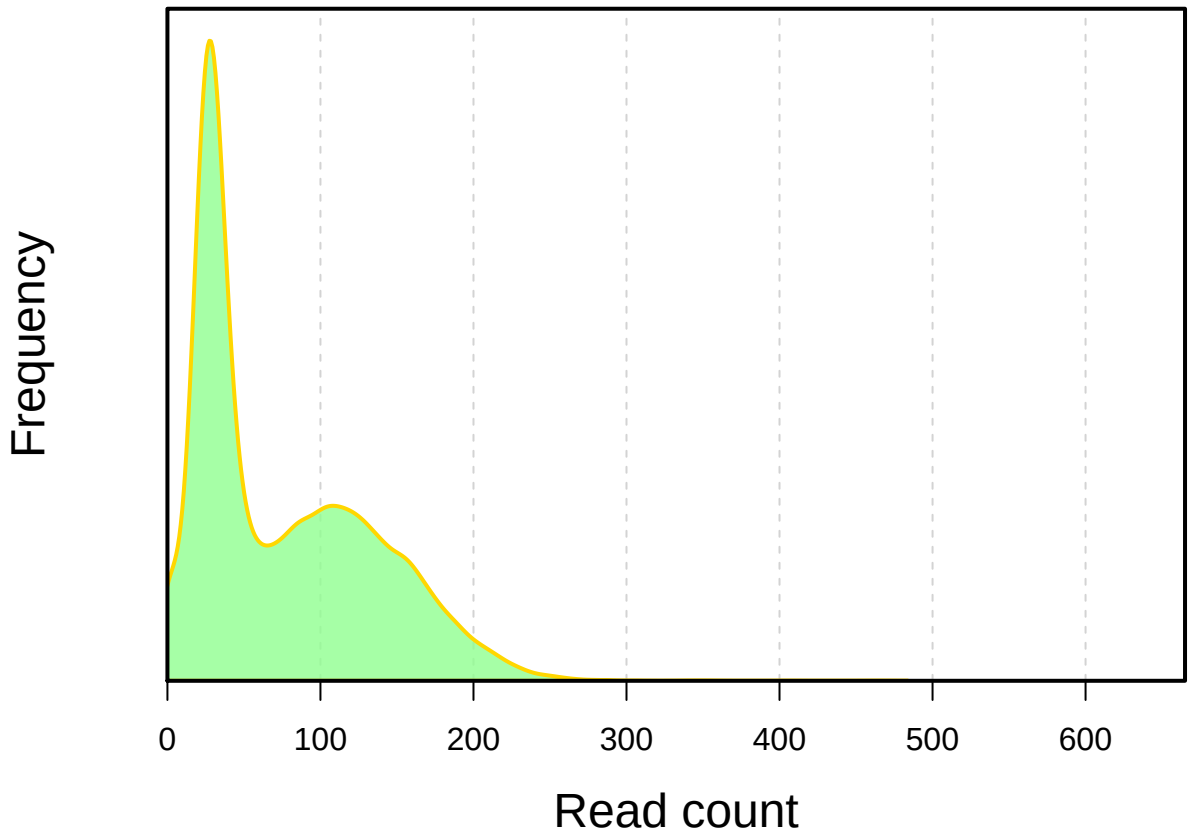


Figure 17: **Read count distribution of genes.**This plot shows the distribution of read counts within the [-1kb, 1kb] region of RefSeq TSSs. Duplicated mapping was excluded.

Table 19: Top 20 genes with the highest read counts around TSS

RefSeq_ID	Sense	Antisense	Total
NR_027412	246	218	464
NR_029683	201	205	406
NM_017940	174	182	356
NM_001002811	172	179	351
NM_182931	198	146	344
NR_024586	144	194	338
NR_037458	160	144	304
NM_001198832	154	150	304
NM_031369	133	156	289
NM_031435	147	140	287
NM_144982	142	144	286
NM_031263	155	129	284
NM_053043	148	134	282
NM_030912	153	128	281
NM_002199	146	134	280
NM_001122964	136	142	278
NR_030366	136	141	277
NM_001242534	141	136	277
NM_152837	130	144	274
NM_014462	138	132	270

7 Alerts

— Less than 1% (0.37%) of the total reads were randomly selected to summarize sequencing quality, mapping quality, mismatch frequency and base frequency.
— There are 9 5-mers overrepresented at either end of reads.