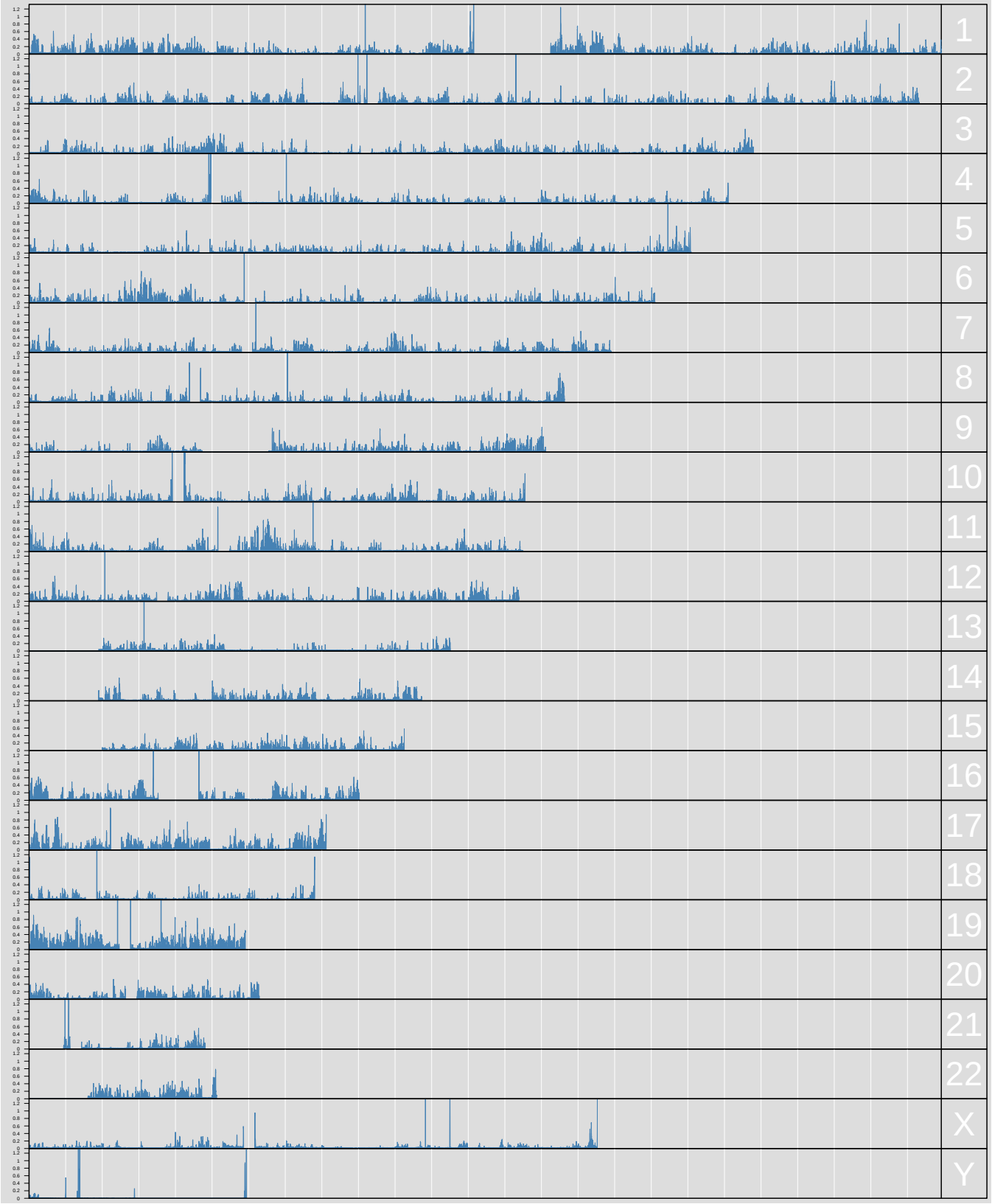
ENCODE CD14+Monocyte Histone ChIP-seq: 671k4

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: 671k4

Genome name: hg19

1.1 BAM file

Size: 1.89 GB

Created: 2013-04-30 04:58:06

Modified: 2013-04-30 04:55:30

Location: > home > zhangz > hts > projects > ks001 > 2013-04_BGI_ChIPseq > bams > novoalign_671k4_-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	35,987,987
Total mapped reads	4,024,595
Total upmapped reads	0
Total mappings	35,987,987
Total mapping locations	3,636,721
Base N%	0.003
(G+C)%	49.4
Mapped to forward strand%	50.01
Duplicated mapping reads%	15.18
Best sequencing quality	41
Average sequencing quality	38.62
Maximum mapping length (bp)	49
Minimum mapping length (bp)	25
Average mapping length (bp)	48.75
Best mapping quality	70
Average mapping quality	41.76
Highest sequencing depth	9,207
Average sequencing depth	0.067
Mapped reads per kilobase	1.39

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	2,757,281,094	96.36
Depth>=1	104,051,512	3.64
Depth>=5	7,965,969	0.28
Depth>=10	2,343,373	0.08
Depth>=20	158,240	0.01
Depth>=30	59,483	0.00
Depth>=50	28,317	0.00
Depth>=100	11,544	0.00
Depth>=1000	1,505	0.00
Depth=9207	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	380,330	336,199	0.08	5,171	91,092,893
chr2	243,199,373	238,207,373	316,633	276,057	0.06	4,560	128,285,576
chr3	198,022,430	194,797,140	224,128	211,253	0.06	280	193,460,439
chr4	191,154,276	187,661,676	209,354	180,853	0.05	9,207	66,939,006
chr5	180,915,260	177,695,260	199,301	185,521	0.05	3,086	171,332,108
chr6	171,115,067	167,395,067	216,505	202,220	0.06	1,140	130,184,090
chr7	159,138,663	155,353,663	201,745	189,134	0.06	246	58,669,049
chr8	146,364,022	142,888,922	179,484	151,607	0.06	6,997	67,487,294
chr9	141,213,431	120,143,431	162,802	152,757	0.07	156	49,501,111
chr10	135,534,747	131,314,747	200,153	173,943	0.07	947	39,020,268
chr11	135,006,516	131,129,516	193,613	177,708	0.07	2,341	73,980,570
chr12	133,851,895	130,481,895	190,216	174,206	0.07	3,917	20,544,439
chr13	115,169,878	95,589,878	95,172	87,640	0.05	1,063	12,398,356
chr14	107,349,540	88,289,540	120,560	111,765	0.07	1,688	71,341,410
chr15	102,531,392	81,694,769	118,062	111,238	0.07	40	8,264,635
chr16	90,354,753	78,884,753	163,088	147,565	0.10	2,606	33,853,739
chr17	81,195,210	77,795,210	187,422	174,456	0.12	425	19,256,013
chr18	78,077,248	74,657,248	80,723	72,850	0.05	642	74,656,465
chr19	59,128,983	55,808,983	205,189	178,995	0.18	5,476	23,974,122
chr20	63,025,520	59,505,520	95,767	89,570	0.08	102	42,919,879
chr21	48,129,895	35,108,702	62,050	49,433	0.08	4,427	316,273
chr22	51,304,566	34,894,566	74,133	69,418	0.10	127	661,655
chrX	155,270,560	151,100,560	125,959	115,454	0.04	3,338	104,487,419
chrY	59,373,566	25,653,566	22,206	16,879	0.04	371	10,332,712

2.3 Depth by genomic feature

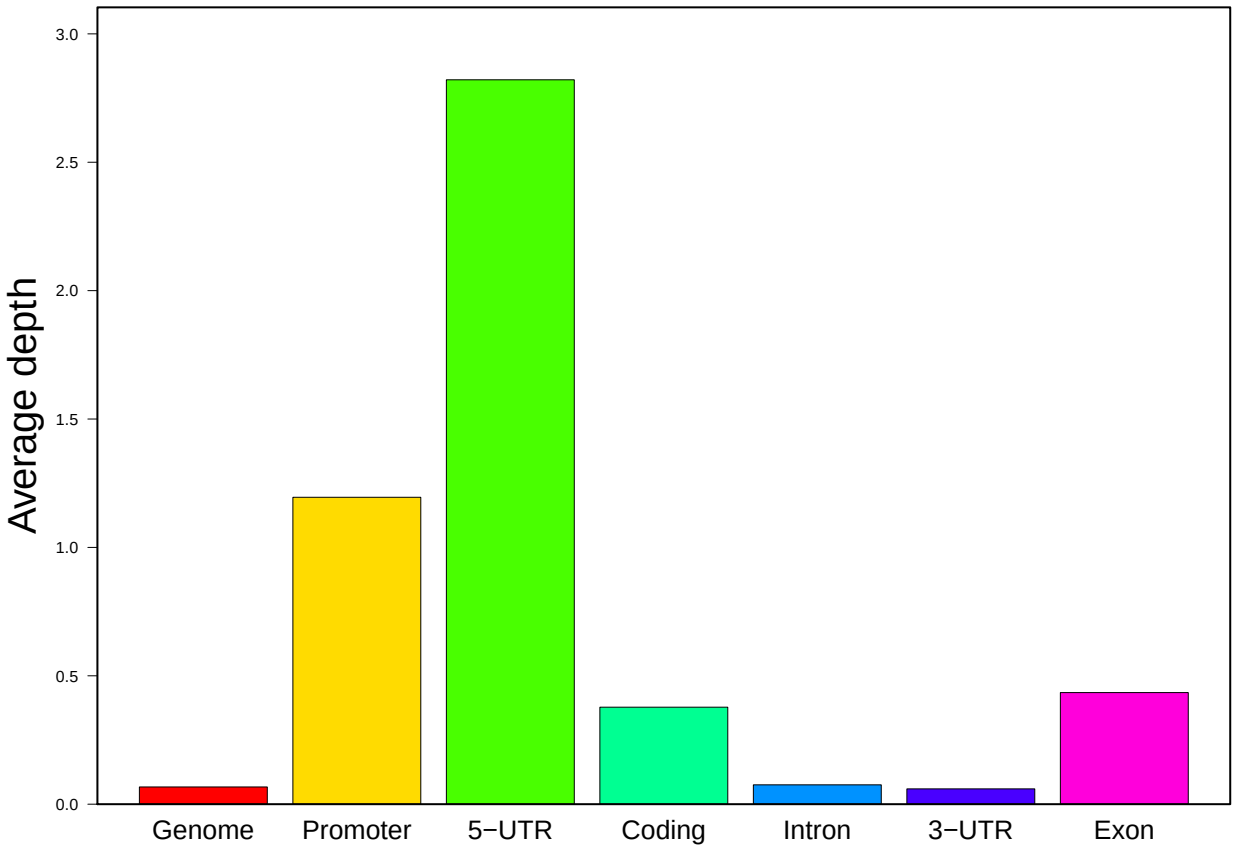


Figure 1: **Average depth of genomic features.** Genomic features are regions annotated based on previous knowledge, such as the RefSeq gene track downloaded from UCSC genome browser. Many applications of high-throughput sequencing technologies, such as exome sequencing and RNA-seq, expect higher depth 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   38.62   41.00   41.00
```

3.1 Quality score categories

Score	Count	Percentage
Score=0	0	0.00
Score>=5	7,075,182	99.83
Score>=10	7,068,376	99.74
Score>=13	7,065,206	99.69
Score>=20	7,047,871	99.45
Score>=30	6,956,591	98.16
Score>=40	3,896,398	54.98
Score=41	3,023,200	42.66

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

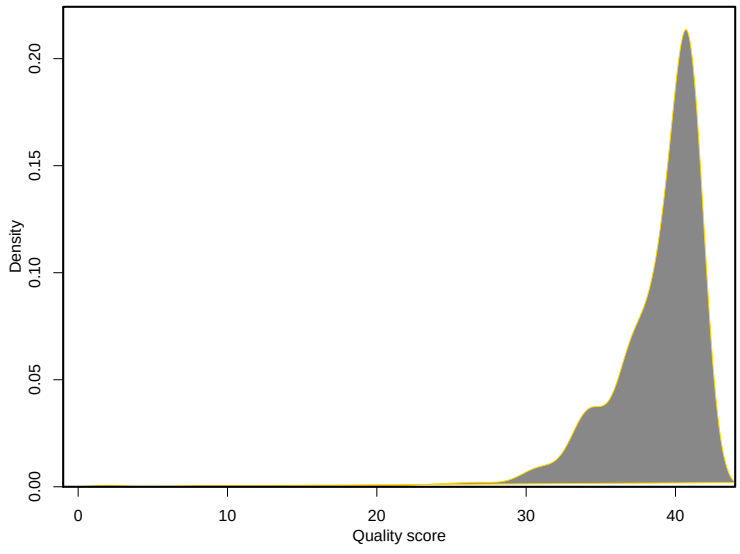


Figure 2: **Score distribution.** This distribution is based on all bases of randomly selected sequencing reads, so position-specific sequencing quality is not considered (see below). The quality scores are calculated by subtracting 33 from the integers corresponding to the ASCII characters in **<QUAL>**.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

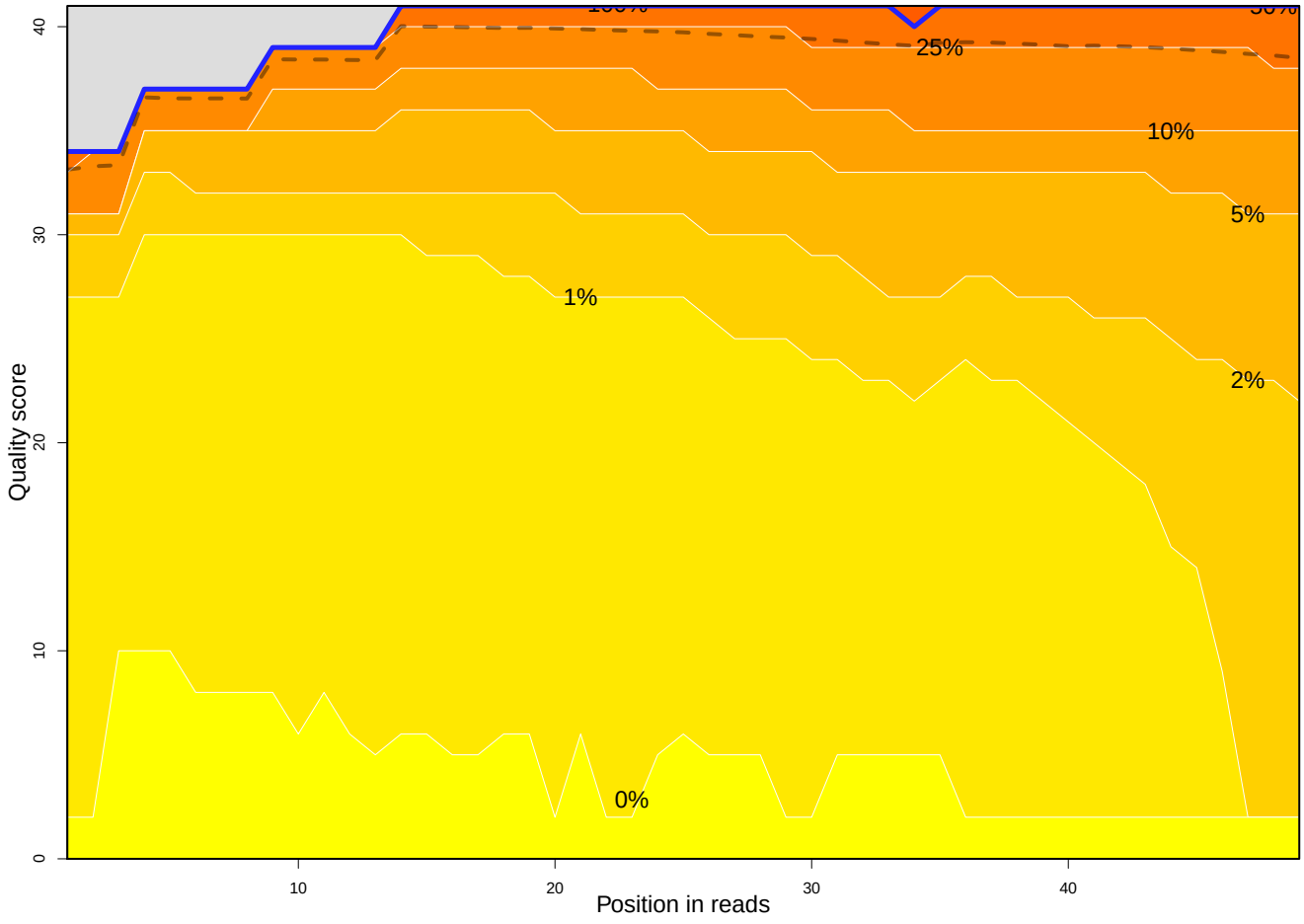


Figure 3: **Position-specific sequencing scores.** This plot shows quality scores at different positions within reads. The dashed lines represents the means of quality scores at different positions; whereas the heat gradient corresponds to percentiles.

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.
  25.00   49.00   49.00   48.75   49.00   49.00
```

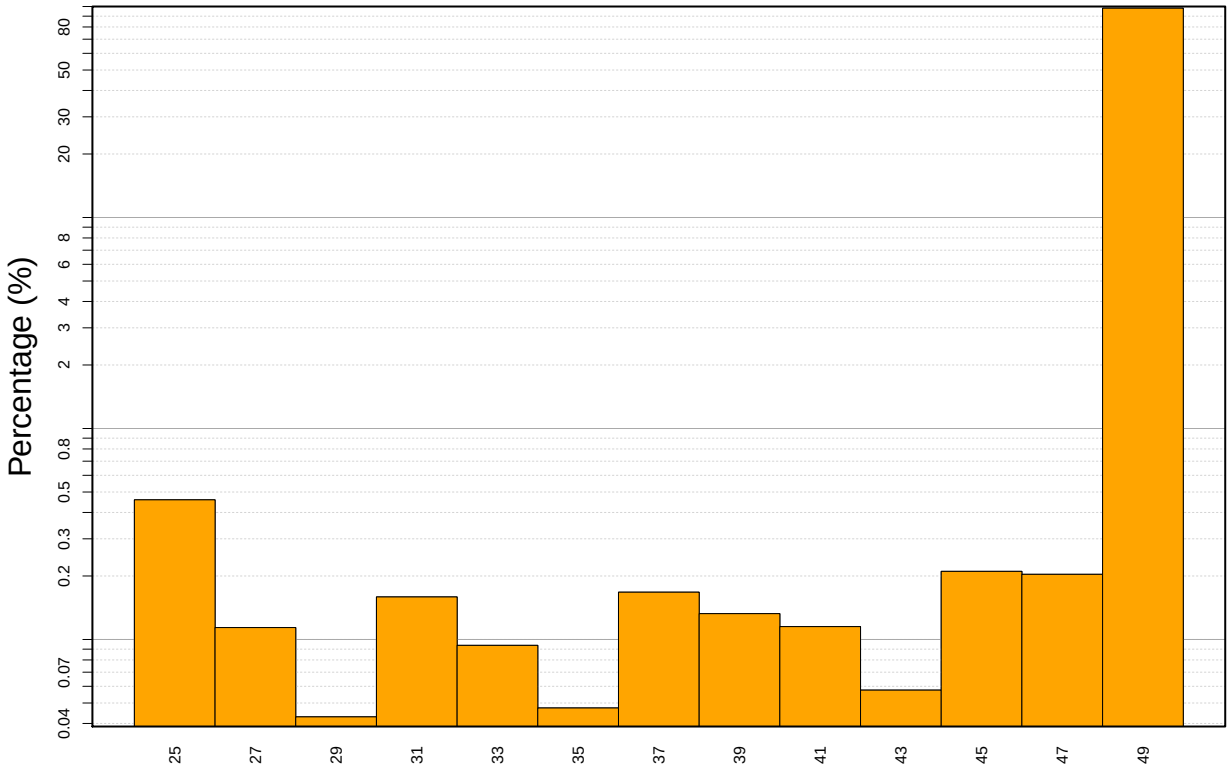


Figure 4: **Frequency of mapping lengths.**

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	17,989,389	49.99
0X20	0	0.00
0X40	0	0.00
0X80	0	0.00
0X100	31,963,392	88.82
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	2,010,196	5.59	-	-	-	-	-	-	-	-	-	-	-
16	2,014,399	5.60	-	-	-	-	X	-	-	-	-	-	-
256	15,988,402	44.43	-	-	-	-	-	-	-	-	X	-	-
272	15,974,990	44.39	-	-	-	-	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   41.76   70.00   70.00
```

4.3.1 Mapping score categories

Score	Count	Percentage
mapq=0	45,533	31.43
mapq>=1	99,325	68.57
mapq>=2	95,969	66.25
mapq>=3	94,514	65.25
mapq>=4	91,808	63.38
mapq>=5	91,759	63.34
mapq>=10	91,489	63.16
mapq>=20	89,654	61.89
mapq>=30	85,866	59.28
mapq>=40	85,053	58.71
mapq=70	80,900	55.85

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

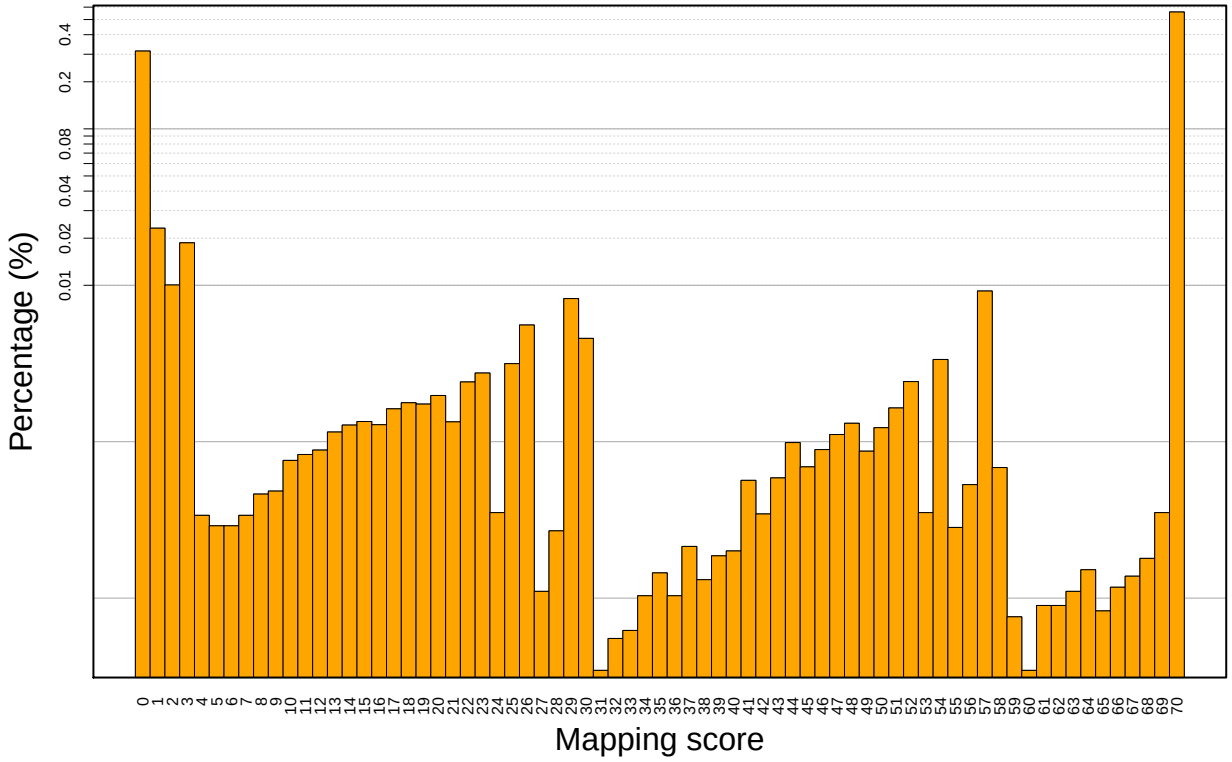


Figure 5: **Mapping score distribution.** By definition, mapping quality equals to -10*Log10(p value), where p value is the likelihood of incorrect mapping; however, its calculation depends on individual programs.

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	4,024,595	100.00
I	26,881	0.67
D	13,704	0.34
S	284,454	7.07
H	72,565	1.80

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.

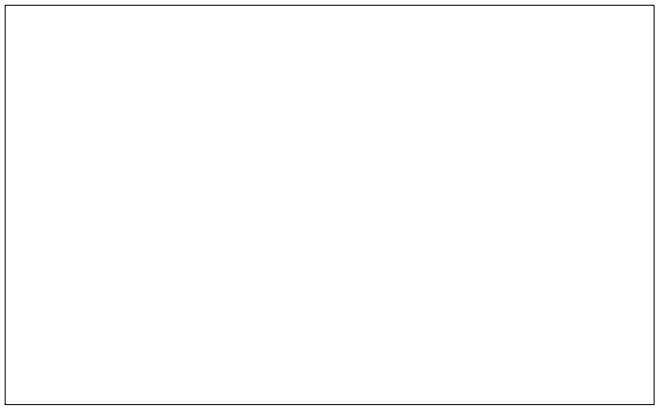


Figure 6: **Distribution of gap size.** If the alignment program tried to align subsequence of the same read to remote locations, **<CIGAR>** will provide the size of gapped regions

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 1.107.

Level	Location_count	Read_count	Percentage
1	3,413,810	3,413,810	84.82
2	194,981	389,962	9.69
3	17,495	52,485	1.30
4	3,620	14,480	0.36
5	1,611	8,055	0.20
6	981	5,886	0.15
7	575	4,025	0.10
8	455	3,640	0.09
9	315	2,835	0.07
10	243	2,430	0.06
>10	2,635	126,987	3.16

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

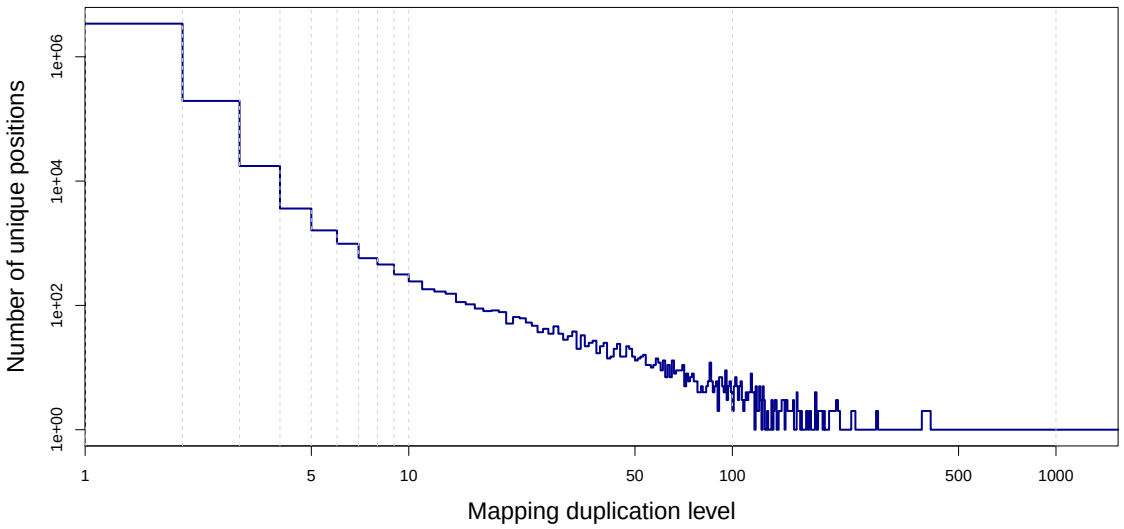


Figure 7: textbfDistribution of duplication levels. The x-axis indicates the number of reads sharing the same mapping location of their 5'-end and the y-axis is the total occurance of each level. Only reads mapped to the forward strand and the first 10 million reads of each chromosome was used to reduce 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	7,087,116	214	0.003
Read	144,858	203	0.140

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(%)	25.13	24.65	24.75	25.47	49.40
Observed/Expected(%)	85.15	120.45	120.87	86.20	120.66

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

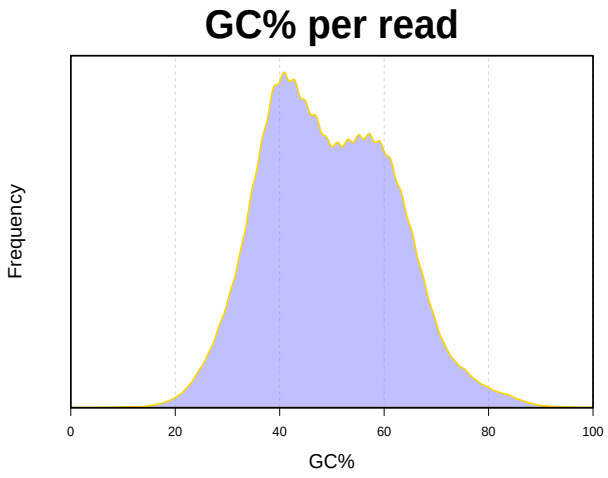


Figure 9: **GC content.** Percentage of C/G bases within each read.

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

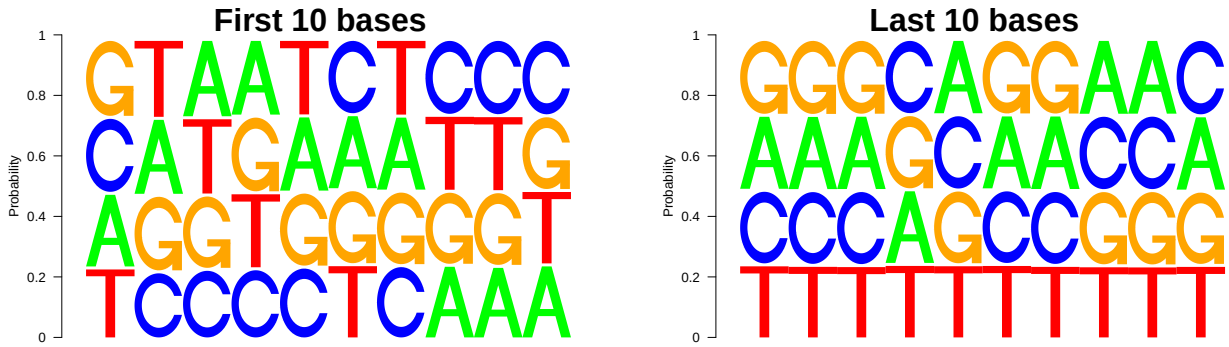


Figure 10: **Single base frequency at both ends.**The base frequency of the first and last 10 bases (the rightmost is the last base) of reads. The frequency was normalized by the overall base frequency with sequencing reads, so this summary indicates the preference of sequencing to start with a given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

5.4.2 First two bases

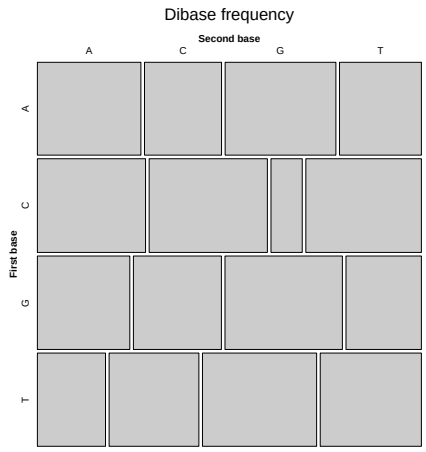


Figure 11: **First two base combination.**This plot summarizes the frequency of the two-base combinations at the 5'-end of reads. The size of the blocks represent their relative frequency after adjusted by their expected frequency based on the position-specific frequency of the first two bases.

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
CGTAC|	141.59	3	0.021
TCGTA|	140.99	4	0.028
|TCGAC	114.23	6	0.053
|CGTAC	146.54	6	0.041
CGATA|	142.12	6	0.042
|TCCGA	113.31	7	0.062
TACGA|	145.29	7	0.048
GTACG|	140.47	8	0.057
GTCGT|	139.66	8	0.057
CGTAT|	140.56	9	0.064

Table 14: Highest frequency

5-mer	Expected_count	Observed_count	Observed/Expected
|TTTTT	165.92	732	4.41
AAAAA|	151.23	685	4.53
TTTTT|	141.26	600	4.25
|AAAAA	162.51	556	3.42
|CCCAG	115.92	517	4.46
GAGGC|	142.95	467	3.27
|CTGGG	150.42	462	3.07
|CCAGG	131.71	455	3.45
|GGTGG	156.25	451	2.89
GCCTC|	138.20	441	3.19

Table 15: Highest relative enrichment

5-mer	Expected_count	Observed_count	Observed/Expected
AAAAA|	151.23	685	4.53
|CCCAG	115.92	517	4.46
|TTTTT	165.92	732	4.41
TTTTT|	141.26	600	4.25
|CCAGG	131.71	455	3.45
|AAAAA	162.51	556	3.42
GAGGC|	142.95	467	3.27
CCTCC|	136.37	440	3.23
TCCTG|	136.62	438	3.21
GCCTC|	138.20	441	3.19
|CTGGG	150.42	462	3.07
|CCTCC	113.72	347	3.05
CCCAG|	135.89	409	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.

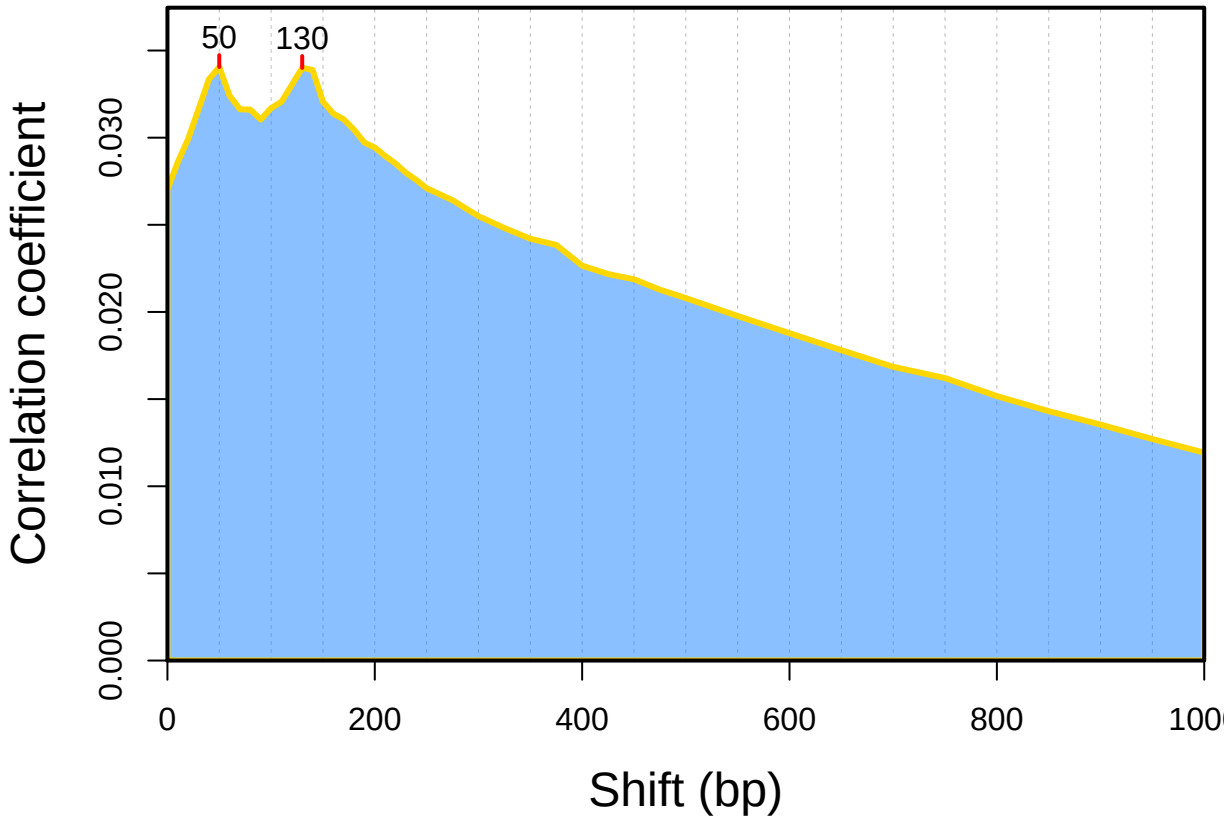


Figure 12: **Strand-strand correlation vs. shifting.**The x-axis is the number of bases to shift the forward strand towards the right. Correlation was calculated between the 5-prime end of mapping locations after removing duplicated mappings.

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	17,098	1,831.46
>=25	10,648	2,155.48
>=50	3,867	2,473.98
>=100	122	4,229.66
>=200	50	5,604.08
>=500	32	5,505.34
>=1,000	24	3,061.75
>=5,000	6	1,248.50
>=10,000	3	649.67
=13,547	1	747.00

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.

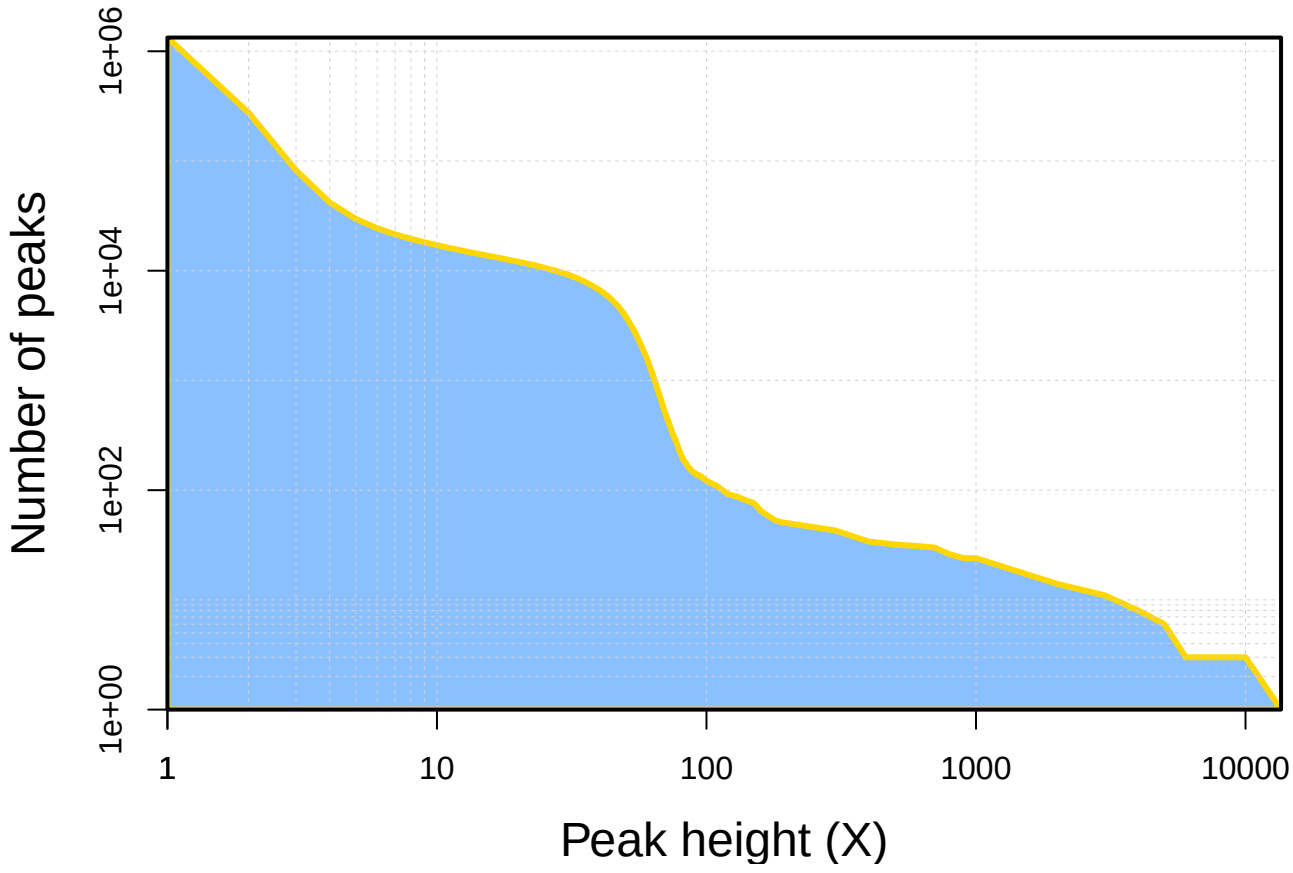


Figure 13: **Peak height distribution.**

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   245.6   200.0 65710.0
```

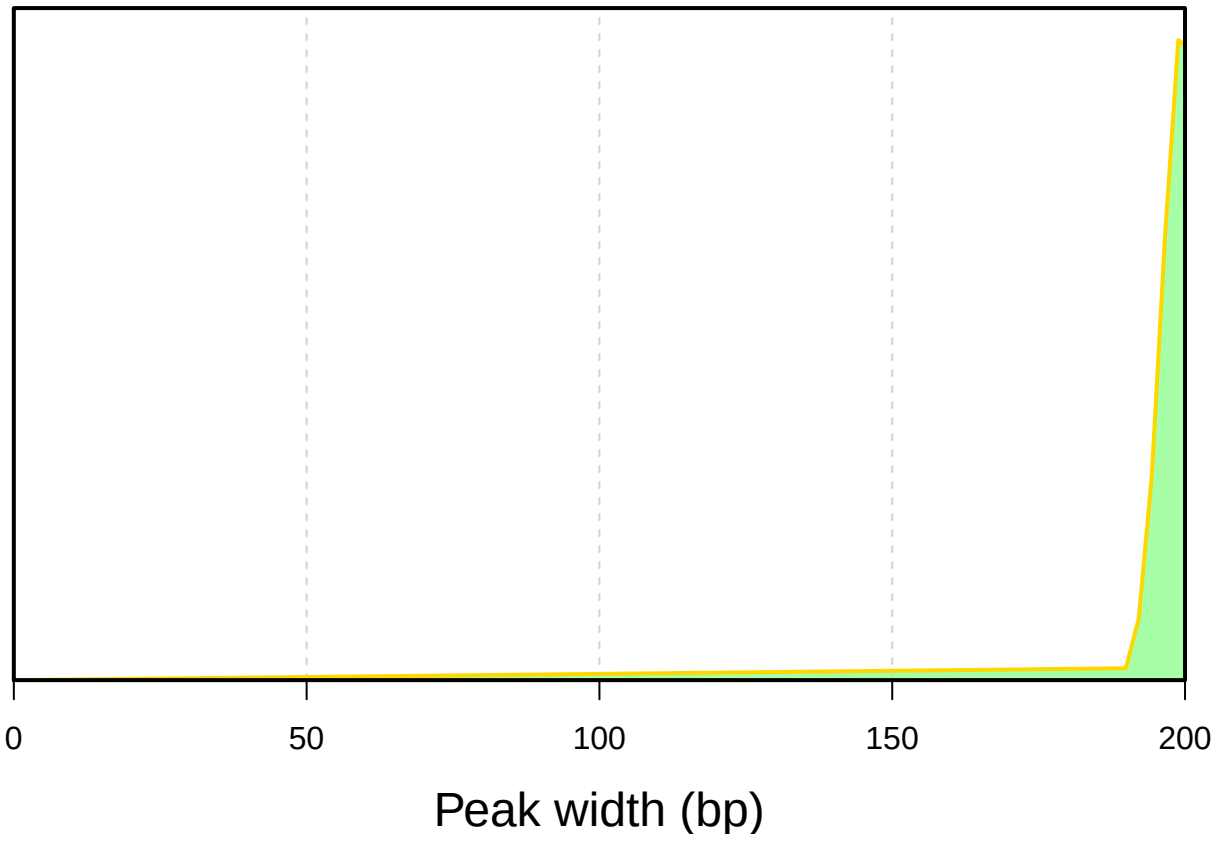


Figure 14: **Peak width distribution.**

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	12,040	11,267	8,747	13,090	906	12,605
Height>=25	9,226	8,840	6,613	9,294	557	9,361
Height>=50	3,484	3,394	2,552	3,512	161	3,521
Height>=100	17	5	5	38	1	21
Height>=200	3	0	0	14	0	3
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>=13547	0	0	0	1	0	0

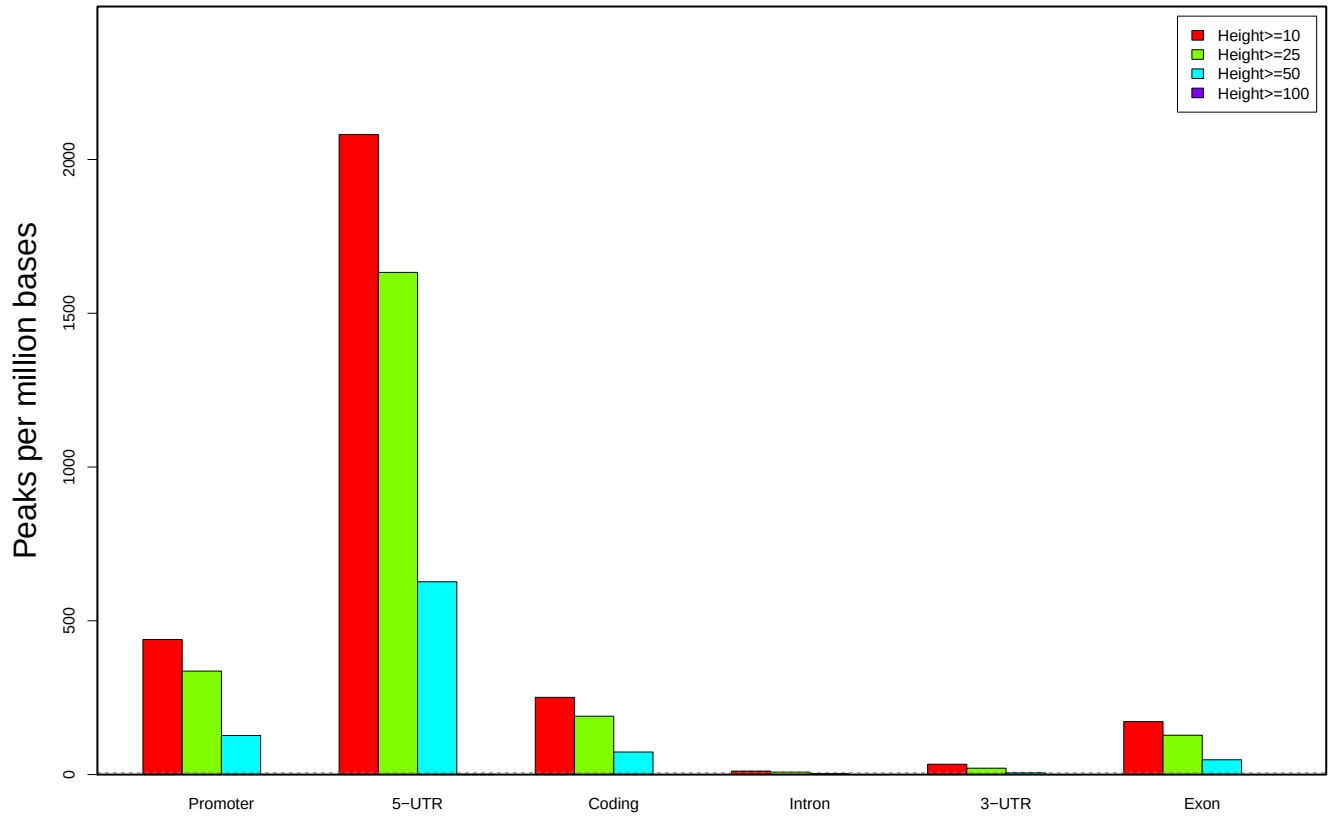


Figure 15: **Peak frequency within genomic features.**The dashed line is the overall frequency of peaks with depth no less than 10 within the whole genome.

6.2.4 Top peaks

Table 18: Top 20 peaks with the highest height.

Chromosome	Start	End	Width	Height
chr8	70,602,127	70,602,873	747	13,547
chr4	70,296,441	70,296,928	488	10,331
chr1	91,852,605	91,853,318	714	10,030
chr2	133,011,729	133,013,357	1,629	5,833
chr21	9,825,282	9,827,865	2,584	5,823
chr19	24,183,922	24,185,250	1,329	5,476
chrX	108,297,199	108,298,005	807	4,360
chr12	20,704,189	20,704,657	469	4,276
chr19	36,066,337	36,066,971	635	3,424
chr5	174,541,592	174,542,319	728	3,147
chr11	77,597,339	77,597,917	579	3,018
chr16	33,960,491	33,966,645	6,155	2,616
chr1	121,478,264	121,485,604	7,341	2,384
chr13	31,418,015	31,418,651	637	2,182
chr14	90,341,220	90,341,668	449	1,690
chr1	145,277,112	145,277,800	689	1,634
chr1	120,543,550	120,544,263	714	1,517
chr2	133,035,419	133,037,126	1,708	1,302
chr1	237,766,173	237,766,755	583	1,276
chr2	9,836	10,356	521	1,264

6.3 TSS

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

6.3.1 Strand-specific depth around TSS

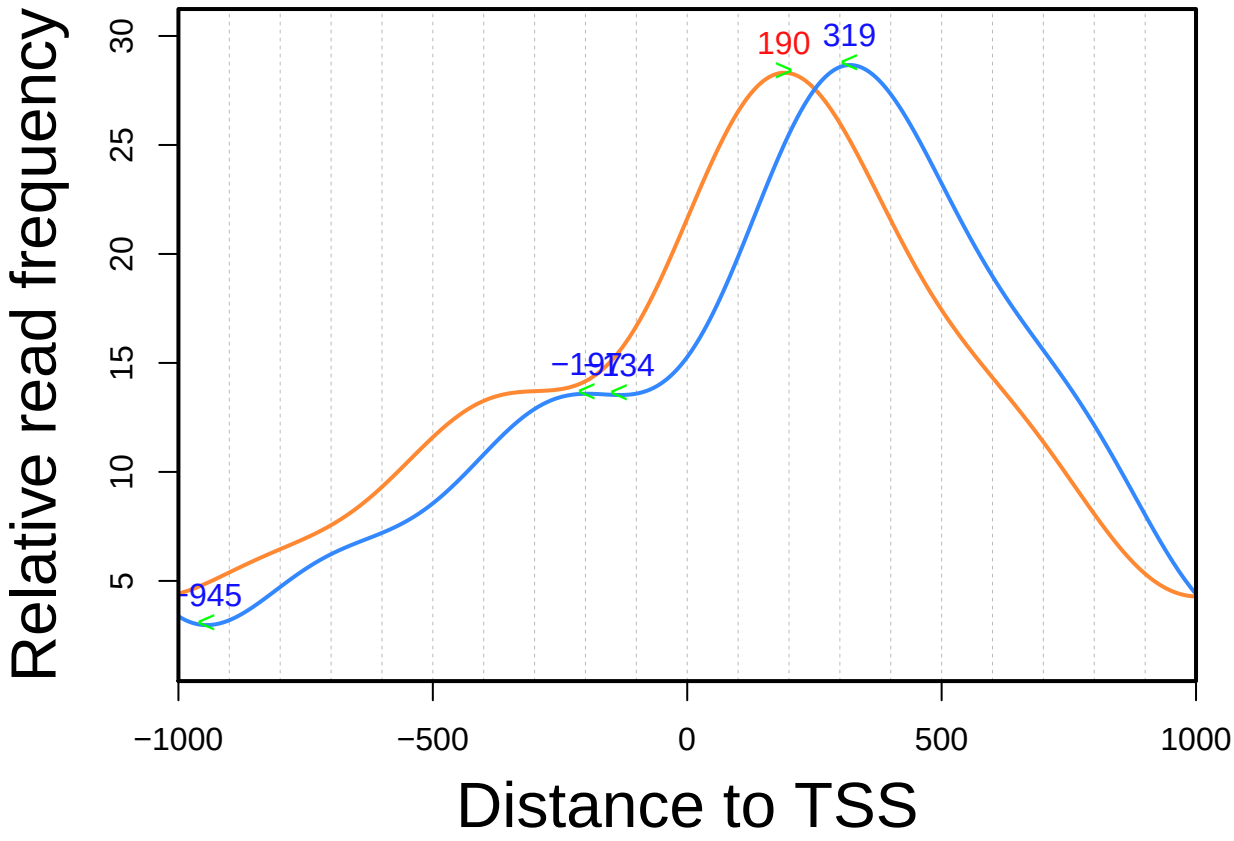


Figure 16: **Read frequency around TSS.**This plot shows the frequency of reads whose 5-prime end was mapped around TSS of RefSeq genes. The read counts were normalized by the global average after duplicated mapping was not removed.

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    3.00   52.00   71.97  128.00  920.00
```

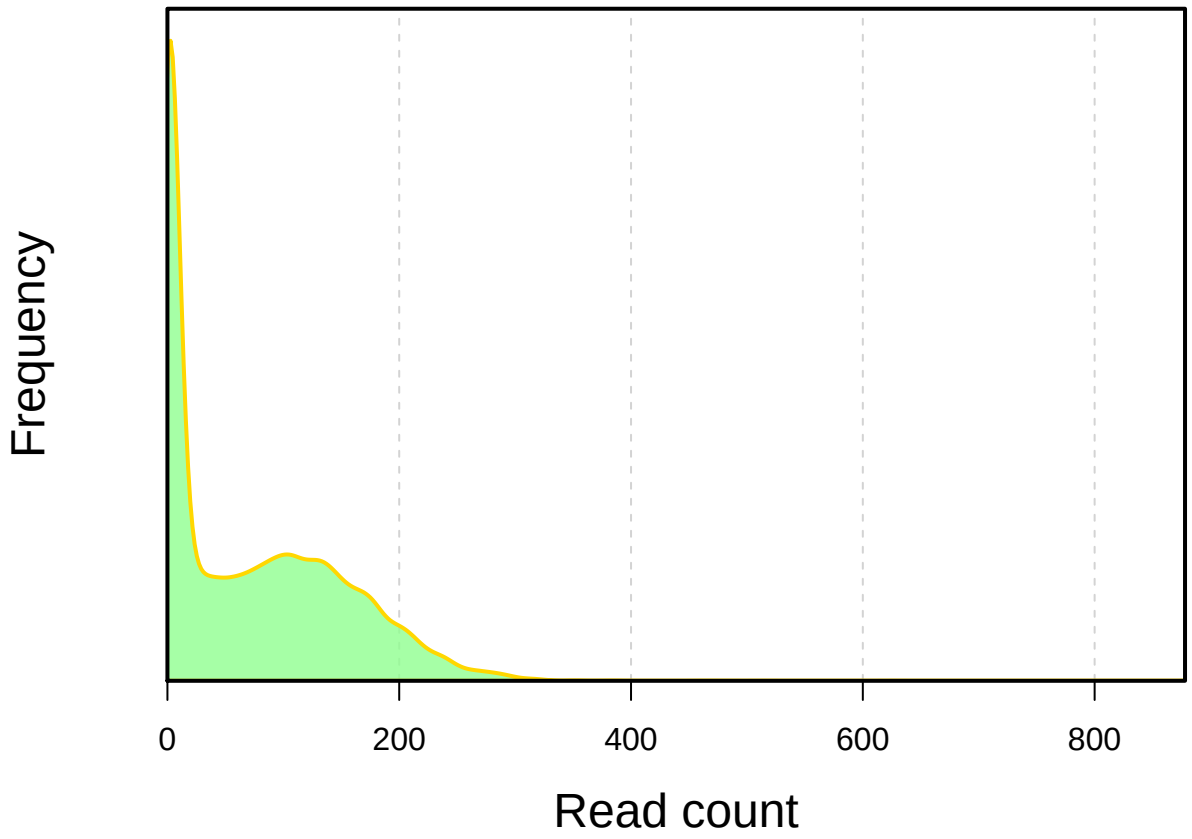


Figure 17: **Read count distribution of genes.**This plot shows the distribution of read counts within the [-1kb, 1kb] region of RefSeq TSSs. Duplicated mapping was excluded.

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

RefSeq_ID	Sense	Antisense	Total
NR_033770	460	460	920
NR_037458	433	453	886
NR_037421	311	315	626
NM_203458	251	264	515
NR_003264	256	225	481
NM_001200001	232	232	464
NM_024408	232	232	464
NR_024077	201	190	391
NM_001002811	203	182	385
NR_003265	192	191	383
NR_029683	177	204	381
NR_040095	194	182	376
NM_017940	167	191	358
NM_101395	187	166	353
NR_003266	165	187	352
NM_014863	166	184	350
NM_015892	166	184	350
NM_001024228	182	168	350
NM_001024227	182	168	350
NM_001159673	164	164	328

7 Alerts

— GC contents of reads are more than 110% (120.7%) of expected percentage.
— There are 13 5-mers overrepresented at either end of reads.