Kmt5a Controls Hepatic Metabolic Pathways by Facilitating RNA Pol II Release from Promoter-Proximal Regions

Bioinformatics Project

Author: **Voulgari Despoina**

Number: **7115152400012**

Professors: **Dimopoulos A., Reczko M.**

1. Description of the data

This project aims to reproduce a ChIP-seq analysis based on a study investigating the role of H4K20Me1 (H4K20 monomethylation) in genome integrity and transcriptional regulation. The study highlights how the turnover of H4K20Me1 in gene bodies positively correlates with gene activity and RNA Polymerase II (RNA Pol II) release from promoter-proximal regions. In particular, it emphasizes the regulation of genes involved in glucose and lipid homeostasis, which are particularly sensitive to this regulation, leading to metabolic reprogramming and genome damage without proper H4K20Me1 modulation.

The analysis will utilize a range of ChIP-seq samples, including those from wild-type (WT) and Set8 knockout (Set8KO) mice, to investigate the differential methylation of H4K20Me1 in various genomic conditions. The data will focus on the post-recruitment steps of transcription, such as promoter escape and RNA Pol II progression, regulated by H4K20Me1. The project aims to shed light on the involvement of H4K20 methylation in safeguarding genome integrity, particularly in non-dividing cells, and its impact on the transcription of metabolic genes. Additionally, the findings will further elucidate the potential links between H4K20Me1 and cellular processes such as replication licensing, mitotic chromatin condensation, and DNA repair.

The following table presents the samples analyzed in this project along with the corresponding names used in the scripts for clarity and reference:

Samples	Description*	SRA number	Name used in scripts
GSM2561922	ChIP WT H4K20me1	SRR5409158	sample22
GSM2561924	ChIP Set8KO H4K20me1	SRR5409160	sample24
GSM2561926	ChIP Serin2 WT PolII	SRR5409162	sample26
GSM2561928	ChIP Serin2 Set8KO PolII	SRR5409164	sample28
GSM2561930	ChIP WT Phf8	SRR5409166	sample30
GSM2561932	ChIP Set8KO Phf8	SRR5409168	sample32
GSM2561934	Input WT	SRR5409170	*WT*
GSM2561936	Input Set8KO	SRR5409172	*Set8KO*
GSM2561938	SDS1 ChIP WT PolII	SRR5409174	sample38
GSM2561940	SDS1 Set8KO WT PolII	SRR5409176	sample40
GSM2561942	SDS1 Input WT	SRR5409178	*SDS1 WT*
GSM2561944	SDS1 Input Set8KO	SRR5409180	*SDS1 Set8KO*

***replicate 1**

Table 1: Samples

2. Protocols

Genome build

In the data pages (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97338) (samples GSM2561922-44) was mentioned that the analysis was performed using the **mm9 (NCBI37)** genome build as the reference for alignment and downstream processing.

For that reason, through the https://hgdownload.soe.ucsc.edu/downloads.html page, the corresponding FASTA file for the mm9 was downloaded (mm9.fa.gz).

Name	Last modified	Size	Description
Parent Directory		-	
chromAgp.tar.gz	2007-07-25 10:29	418K	
chromFa.tar.gz	2007-07-25 10:44	820M	
chromFaMasked.tar.gz	2007-07-25 10:53	481M	
chromOut.tar.gz	2007-07-25 10:30	149M	
chromTrf.tar.gz	2008-05-07 11:07	17M	
est.fa.gz	2019-10-17 12:06	788M	
est.fa.gz.md5	2019-10-17 12:06	44	
genes/	2020-02-05 13:47	-	
md5sum.txt	2008-10-16 11:43	254	
mm9.2bit	2007-07-21 13:31	680M	
mm9.chrom.sizes	2007-07-19 14:58	584	
mm9.fa.gz	2020-01-23 02:23	820M	
mrna.fa.gz	2019-10-17 11:49	261M	
mrna.fa.gz.md5	2019-10-17 11:49	45	
refMrna.fa.gz	2019-10-17 12:07	44M	
refMrna.fa.gz.md5	2019-10-17 12:07	48	
upstream1000.fa.gz	2019-10-17 12:08	7.8M	
upstream1000.fa.gz.md5	2019-10-17 12:08	53	
upstream2000.fa.gz	2019-10-17 12:08	15M	
upstream2000.fa.gz.md5	2019-10-17 12:08	53	
upstream5000.fa.gz	2019-10-17 12:09	36M	
upstream5000.fa.gz.md5	2019-10-17 12:09	53	
xenoMrna.fa.gz	2019-10-17 12:00	6.5G	
xenoMrna.fa.gz.md5	2019-10-17 12:00	49	
xenoRefMrna.fa.gz	2019-10-17 12:07	287M	
xenoRefMrna.fa.gz.md5	2019-10-17 12:07	52	

Figure 1: mm9 fasta file used in genome build

ChIP-seq Analysis: Script Explanation and Documentation

This section provides a detailed explanation of the Bash scripts used to perform ChIP-seq analysis. Each script is broken down step by step, with explanations of commands and parameters. The analysis follows a structured workflow, including quality control, read alignment, input normalization, peak calling, and motif discovery.

Remarks:

1. The process was separated in 4 parts – 4 scripts - in order to handle errors and avoid crashing the system due to overload.
2. A list of the samples and their corresponding names used in the scripts was created for consistency and simplicity.
3. Before the beginning of the analysis a quick check of the provided tools and their version was performed.

```
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ fastqc --version
FastQC v0.12.1
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ minion --version
[minion] Usage: minion <search-adapter|help>
[example] minion help
[example] minion search-adapter -i FASTQFILE
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ bowtie2 --version
/usr/bin/bowtie2-align-s version 2.5.2
64-bit
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ macs --version
macs 1.4.1 20110622
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ bedtools --version
bedtools v2.31.1
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ samtools --version
samtools 1.19.2
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18$ cutadapt --version
Command 'cutadapt' not found, but can be installed with:
apt install cutadapt
Please ask your administrator.
```

Figure 2: Check of provided tools

1. ChIP_seq_analysis: Initial Processing and Alignment

This script performs read quality control, alignment, and preparation for downstream analysis.

1.1. FASTQ Quality Control

The first step runs FASTQC to assess sequencing quality.

```
fastqc "$FASTQ_DIR/$fastq_file" -o "$FASTQ_DIR/fastqc_reports"
```

Option	Explanation
`fastqc`	Runs the FASTQC tool for quality assessment.
`"$FASTQ_DIR/$fastq_file"`	Specifies the input FASTQ file.
`-o "$FASTQ_DIR/fastqc_reports"`	Outputs results to the `fastqc_reports` directory.

Table 2: Explanation of command lines

1.2. Read Alignment with Bowtie2

```
bowtie2 -p 4 -x "$BOWTIE2_INDEX" -U "$FASTQ_DIR/$fastq_file" -S
"$sample_name.sam"
```

Option	Explanation
`-p 4`	Uses 4 CPU threads for faster alignment.
`-x "$BOWTIE2_INDEX"`	Specifies the Bowtie2 genome index (mm9).
`-U "$FASTQ_DIR/$fastq_file"`	Uses a single-end FASTQ file as input.

`-S "$sample_name.sam"`	Outputs the aligned reads in SAM format.

Table 3: Explanation of command lines

1.3. Conversion and Sorting of Aligned Reads

```
samtools view -bS "$sample_name.sam" > "$sample_name.bam"
```

Option	Explanation
`view`	Converts SAM to BAM format.
`-bS`	Specifies BAM output (`-b`) and accepts SAM input (`-S`).
`> "$sample_name.bam"`	Redirects output to a BAM file.

Table 4: Explanation of command lines

Sorting and indexing:

```
samtools sort -o "$sample_name.sorted.bam" "$sample_name.bam"
samtools index "$sample_name.sorted.bam"
```

Option	Explanation
`sort -o`	Sorts the BAM file.
`index`	Creates an index file for efficient data access.

Table 5: Explanation of command lines

---

2. Alignment of Input Controls

The script in `alingment_input_control` script follows the same alignment steps as above but applies them to input control samples.

```
bowtie2 -x "$GENOME" -U "$FASTQ_DIR/$FILE" -S "$FASTQ_DIR/$NAME.sam" --very-sensitive -k 1
```

Option	Explanation
`--very-sensitive`	Uses high-sensitivity settings for accurate alignment.
`-k 1`	Reports only the best-matching alignment per read.

Table 6: Explanation of command lines

---

3. Peak Calling with MACS

The script `peaks` script uses **MACS** to identify regions of significant ChIP enrichment.

```
macs -t "${sample_name}.sorted.bam" -c "$control" --format BAM --name "$sample_name" --gsize 138000000 --tsize 26 --diag –wig
```

Option	Explanation
`-t "${sample_name}.sorted.bam"`	Specifies the treatment BAM file.

`-c "$control"`	Specifies the input control BAM file.
`--gsize 138000000`	Sets the genome size (mouse mm9).
`--tsize 26`	Specifies read length.
`--diag`	Generates a diagnostic report.
`--wig`	Outputs results in WIG format for visualization.

Table 7: Explanation of command lines

---

4. Final Steps: Motif Discovery

The `finalsteps` script runs **MEME** to identify enriched motifs in peak regions.

```
meme "${sample_name}_summits-b20.fa" -o "${sample_name}_meme" -dna
```

Option	Explanation
`meme`	Runs the MEME motif discovery tool.
`-o "${sample_name}_meme"`	Specifies output directory.
`-dna`	Indicates that input sequences are DNA.

Table 8: Explanation of command lines

The full scripts used for the above analysis will be provided in the supplementary material section in the end of this file.

4. Statistics and Results

3.1 FastQC reports – Sequencing Data Quality

The generated FastQC reposts include key metrics like per base sequence quality, adapter content, and GC distribution, were evaluated to ensure data reliability.

Overall, all the samples have a decent Phred score (around 28) that remains until the end of reads in most cases. In some samples, a mild drop (to 26) is observed in the final reads, but it is expected. On the other hand, there are samples, like 26, 28, 36, 38 and 40, which show a severe drop (below 10). That is something, which may affect downstream analysis, such as alignment accuracy.

One solution to that is to trim these final reads, using for example cutadapt, but this command was not available in the Linux environment.

The images of the graphs, concerning per base sequence quality and summary statistics will be provided in the supplementary material section.

3.2 Alignment rates

Alignment rates were calculated as the percentage of reads that successfully mapped to the genome. They ranged from 85% to 92% across all samples, expect one – sample 26- which had 32%. This is indicating high-quality data and a good match to the reference genome. No significant differences in alignment rates were observed between WT and Set8KO samples.

The high alignment rates observed in all samples (>85%) confirm the reliability of the sequencing data and the suitability of the mm9 reference genome for this analysis. This ensures that downstream peak calling and motif discovery are based on robust data.

Description*	SRA number	Alignment rates
ChIP WT H4K20me1	SRR5409158	91.78%
ChIP Set8KO H4K20me1	SRR5409160	92.18%
ChIP Serin2 WT PolII	SRR5409162	32.61%
ChIP Serin2 Set8KO PolII	SRR5409164	91.03%
ChIP WT Phf8	SRR5409166	91.74%
ChIP Set8KO Phf8	SRR5409168	92.50%
Input WT	SRR5409170	77.42%
Input Set8KO	SRR5409172	92.31%
SDS1 ChIP WT PolII	SRR5409174	82.08%
SDS1 Set8KO WT PolII	SRR5409176	84.50%
SDS1 Input WT	SRR5409178	93.37%
SDS1 Input Set8KO	SRR5409180	73.04%

Table 9: Percentages of alignment rates

3.3 Peaks analysis

`macs` command is used to analyze the peak quality, stability and biological relevance of the samples. Below the table contains the results of the diagnosis reports of each sample, that were extracted through the `peaks` script.

	FC range	# of Peaks	cover by sampling 90%	0,8	0,7	0,6	0,5	0,4	0,3	0,2
22	0-20	26	69.23	65.38	65.38	57.69	46.15	46.15	42.31	38.46
24	0-20	20	65.00	65.00	65.00	75.00	65.00	60.00	50.00	30.00
26	0-20	1025	65.76	61.85	59.41	54.44	49.56	45.17	38.24	32.20
	20-40	17	29.41	29.41	29.41	35.29	35.29	35.29	41.18	52.94
	40-60	4	50.00	50.00	50.00	50.00	25.00	50.00	25.00	25.00
28	0-20	650	65.23	63.08	58.77	56.15	49.08	44.46	38.00	30.77
	20-40	7	14.29	14.29	14.29	14.29	14.29	14.29	28.57	42.86
30	0-20	1085	92.35	88.11	82.30	74.38	67.28	60.00	50.69	37.70
32	0-20	36	100.00	100.00	94.44	97.22	94.44	94.44	86.11	77.78
38	0-20	6281	91.08	89.11	86.58	84.06	81.77	77.68	69.88	57.33
	20-40	12	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
40	0-20	3313	93.51	89.62	87.17	83.22	78.90	73.56	66.47	55.69
	20-40	35	97.14	97.14	97.14	97.14	97.14	97.14	97.14	97.14
	40-60	9	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 10: Statistics of Peaks under different sequencing depths

The columns in the ***_diag.xls** files provide information about the quality and stability of peaks under different sequencing depths. Next, it is explained what each column represents:

FC range (Fold Change Range)

This column categorizes peaks based on their fold enrichment (FC) values. Fold change is calculated as the ratio of read counts at a peak in ChIP-seq vs. input.

# of Peaks

Number of peaks found within the specified fold change range.

Cover by sampling (90% - 20%)

These columns show the percentage of peaks that remain detected when subsampling the sequencing data at different depths. Sampling depth percentages (90% → 20%) simulate how stable the peaks are when reducing the sequencing reads.

The information that it is extracted from these files is:

- **High stability:** If a peak remains detected even at low sequencing depths, it is strong and reproducible.
- **Low stability:** If a peak disappears quickly as depth decreases, it may be weak, noisy, or dependent on high sequencing depth.

Key Observations

H4K20me1 Stability (Samples 22 & 24)

- **WT (Sample 22)**:
  - Peaks with **FC 0-20** are **moderately stable**, with 69.23% remaining at 90% depth and 38.46% at 20% depth.
  - This suggests that **a significant fraction of peaks is lost** at lower sequencing depths.
- **Set8KO (Sample 24)**:
  - Peaks are **slightly less stable**, but at 50% depth, stability increases (75%).
  - At 20% depth, only 30% of peaks remain, suggesting **a higher dropout rate than WT**.

Interpretation:

- **WT H4K20me1 peaks are more stable than Set8KO.**
- The instability in Set8KO suggests **a reduction in H4K20me1 enrichment**, aligning with the hypothesis that Set8 depletion affects H4K20me1 retention.

Pol II Stability (Samples 26, 28, 38, 40)

- **WT Ser2 Pol II (Sample 26)**:
  - Peaks with **FC 0-20** are **less stable than H4K20me1** (65.76% at 90%, 32.20% at 20%).
  - **FC 20-40 peaks are highly unstable** but increase at lower depths (52.94% at 20%).
- **Set8KO Ser2 Pol II (Sample 28)**:
  - FC 0-20 peaks have slightly lower stability than WT (65.23% at 90%, 30.77% at 20%).
  - **FC 20-40 peaks are even more unstable (14-42%)**.

Interpretation:

- **Pol II peaks in Set8KO are slightly less stable than WT**, supporting the idea that **Set8 loss affects Pol II stability**.
- The increase in FC 20-40 peaks at **lower sequencing depths** suggests that some peaks **might be noise-sensitive**.

Phf8 Stability (Samples 30 & 32)

- **WT Phf8 (Sample 30)**:
  - Peaks with **FC 0-20** are **highly stable** (92.35% at 90%, 37.70% at 20%).

- **Set8KO Phf8 (Sample 32)**:
    - **Very stable**, with 100% of peaks retained at 90% depth and 77.78% at 20% depth.

Interpretation:

- **Phf8 peaks are the most stable of all marks**.
- **Set8KO Phf8 peaks are even more stable than WT**, suggesting Phf8 remains **robust despite Set8 loss**.

SDS1 Pol II Stability (Samples 38 & 40)

- **WT Pol II (Sample 38)**:
    - **Highly stable**, with 91.08% of FC 0-20 peaks remaining at 90% depth and 57.33% at 20%.
    - **FC 20-40 peaks are extremely stable (100% across all depths).**
- **Set8KO Pol II (Sample 40)**:
    - Similar stability to WT, with **93.51% at 90% depth and 55.69% at 20%**.

Interpretation:

- SDS1-associated **Pol II peaks are the most stable** compared to other Pol II datasets.
- **Set8KO does not significantly reduce peak stability in SDS1**, suggesting **a Pol II population in SDS1 that is unaffected by Set8 loss**.

Overall Conclusions

1. **H4K20me1 peaks are more unstable in Set8KO**, supporting the role of Set8 in maintaining this mark.
2. **Pol II peaks are slightly less stable in Set8KO**, suggesting Set8 may play a role in transcriptional elongation or stability.
3. **Phf8 peaks are the most stable**, and Set8KO does not significantly alter their stability.
4. **SDS1 Pol II is highly stable**, and Set8KO does not affect peak retention, indicating a distinct mechanism for Pol II in SDS1 compared to Ser2P Pol II.

The next step is to analyze the ***peak.xls files,** which contain **peak data** with the following columns:

1. **chr** – Chromosome where the peak is located (e.g., chr11, chr12).
2. **start** – Start position of the peak.
3. **end** – End position of the peak.
4. **length** – Length of the peak region.

5. **summit** – Position with the highest signal within the peak.
6. **tags** – Number of mapped reads supporting the peak.
7. **fold_enrichment** – Enrichment level compared to the control sample.
8. **FDR (%)** – False Discovery Rate, representing the statistical confidence of the peak.

Using the command:

```
for file in sample22_peaks.xls sample24_peaks.xls sample26_peaks.xls
sample28_peaks.xls sample30_peaks.xls sample32_peaks.xls sample38_peaks.xls
sample40_peaks.xls
do
   echo "Top 3 peaks for $file:"
    grep -v "^#" "$file" | awk 'NR>1' | sort -k8,8nr | head -3
    echo "--------------------------------"
done
```

we can extract the top 3 peaks from each file.

Option	Explanation
`for file in ...`	Loops through all 8 peak files.
`grep -v "^#"`	Removes comment lines (#).
`awk 'NR>1'`	Skips the **first line**, which contains column headers.
`sort -k8,8nr`	Sorts by **8th column (Fold Enrichment) in descending order**.
`head -3`	Selects the **top 3 most enriched peaks**

Table 10: Explanation of command lines

The output results are the following:

Top 3 peaks for sample22_peaks.xls:

chr5 37051751 37052777 1027 454 390 785.97 10.99 100

chr8 20009469 20020058 10590 5275 3424 3100.00 9.44 100

chr8 19882100 19886266 4167 2058 1286 1804.50 8.18 100

Top 3 peaks for sample24_peaks.xls:

chr5 37051751 37052772 1022 472 377 716.56 8.78 100

chr8 20014161 20017086 2926 582 629 421.85 5.33 100

chr8 19883617 19885439 1823 508 465 419.56 5.12 100

Top 3 peaks for sample26_peaks.xls:

chr12 105576516 105587815 11300 8227 24621 3100.00 51.71 14.29

chr5	90902436	90907497	5062	750	10126	3100.00	48.94	14.29
chr5	90889766	90902274	12509	2217	30005	3100.00	42.62	14.29

Top 3 peaks for sample28_peaks.xls:

chr18	65403775	65413812	10038	5730	16396	3100.00	33.48	1.33
chr4	150228922	150249614	20693	14756	38753	3100.00	32.23	1.33
chr2	172978264	172986653	8390	6628	21496	3100.00	32.03	1.33

Top 3 peaks for sample30_peaks.xls:

chr9	24346290	24347072	783	257	480	2522.26	18.27	100
chr3	81454291	81455152	862	487	239	605.89	12.46	100
chr11	5661669	5662781	1113	594	306	761.97	11.56	100

Top 3 peaks for sample32_peaks.xls:

chr9	82464202	82464962	761	319	252	275.28	8.10	66.67
chr2	36823425	36824096	672	368	263	319.26	7.57	100
chr12	61565028	61565712	685	230	236	281.32	7.51	80.00

Top 3 peaks for sample38_peaks.xls:

chr11	87256384	87257401	1018	494	1342	3100.00	29.40	5.49
chr11	87284495	87285435	941	512	1224	3100.00	28.33	5.49
chr11	87261757	87262847	1091	504	1297	3100.00	27.99	5.49

Top 3 peaks for sample40_peaks.xls:

chr11	87256186	87257543	1358	677	1421	3100.00	45.84	62.22
chr11	87235807	87237138	1332	671	1381	3100.00	45.53	62.22
chr11	87261593	87262837	1245	641	1371	3100.00	44.50	62.22

Interpretation of Peak Findings

Comparing WT vs. Set8KO for H4K20me1 (Samples 22 vs. 24)

- **Shared Peak** (chr5: 37,051,751-37,052,777)
  - WT (22): 785.97 fold enrichment
  - Set8KO (24): 716.56 fold enrichment *(↓ slightly lower in Set8KO)*
  - Interpretation:
    - H4K20me1 binding is weaker in Set8KO, supporting the idea that Set8 deletion reduces H4K20me1 levels.
- **Chr8 peaks shift slightly**
  - WT (22) has **chr8: 19882100-19886266** with **1804.50 fold enrichment**
  - Set8KO (24) has **chr8: 19883617-19885439** with **419.56 fold enrichment**
  - **Interpretation:**
    - **Loss of strong H4K20me1 peaks in Set8KO** could mean that Set8 is critical for maintaining H4K20me1 at these sites.

Biological Relevance

- **Chr8 regions are highly enriched** → Could be linked to **genes involved in metabolism** (relevant to the study).
- **H4K20me1 peak reduction in Set8KO** supports its role in **transcriptional regulation**.

Comparing WT vs. Set8KO for Pol II Ser2 (Samples 26 vs. 28)

- **Highest peaks in WT (26)**:
  - Chr12: 10,5576,516-10,5587,815 (51.71 fold enrichment)
  - Chr5: 90,902,436-90,907,497 (48.94 fold enrichment)
- **Highest peaks in Set8KO (28)**:
  - Chr18: 65,403,775-65,413,812 (33.48 fold enrichment)
  - Chr4: 150,228,922-150,249,614 (32.23 fold enrichment)

Interpretation

- **Major shift** in peak locations between **WT** and **Set8KO**.
- **Pol II in WT** is enriched in **chr12 and chr5**, while **Set8KO shows strongest peaks in chr18 and chr4**.
- **Potential explanation**:
  - **Loss of H4K20me1 may trigger redistribution of Pol II to new genomic regions**.

Comparing WT vs. Set8KO for Phf8 (Samples 30 vs. 32)

- **Highest Phf8 peak in WT (30)**:
  - Chr9: 24,346,290-24,347,072 (2522.26 fold enrichment)
- **Highest Phf8 peak in Set8KO (32)**:
  - Chr9: 82,464,202-82,464,962 (275.28 fold enrichment)

Interpretation

- **Drastic reduction in Phf8 peak strength in Set8KO**.
- Phf8 is a **histone demethylase**, so its loss in Set8KO **might suggest epigenetic reprogramming**.
- **Hypothesis**: Phf8 might be recruited to different loci when Set8 is absent.

SDS1 Pol II Comparison (Samples 38 vs. 40)

- **Top peaks in WT (38)**:
  - Chr11: 87,256,384-87,257,401 (3100.00 fold enrichment)
- **Top peaks in Set8KO (40)**:
  - Chr11: 87,256,186-87,257,543 (3100.00 fold enrichment)

Interpretation

- **SDS1 Pol II** peaks remain **highly enriched** even in Set8KO**.**
- Unlike Ser2 Pol II, **SDS1 does not show major peak shifts** → May suggest a distinct function of SDS1-related Pol II activity.

Overall Conclusions

1. **H4K20me1 loss in Set8KO** reduces peak enrichment at key sites.
2. **Pol II binding shifts dramatically in Set8KO**, suggesting **transcriptional reprogramming**.
3. **Phf8 binding weakens in Set8KO**, supporting its role in chromatin regulation.
4. **SDS1 Pol II peaks are stable, unlike Ser2 Pol II.**

3.3 IGV visualization

Now we proceed to visualize these peaks in the IGV environment. This is a typical image to use as a reference to explain what we can see and extract as an information for this kind of images.

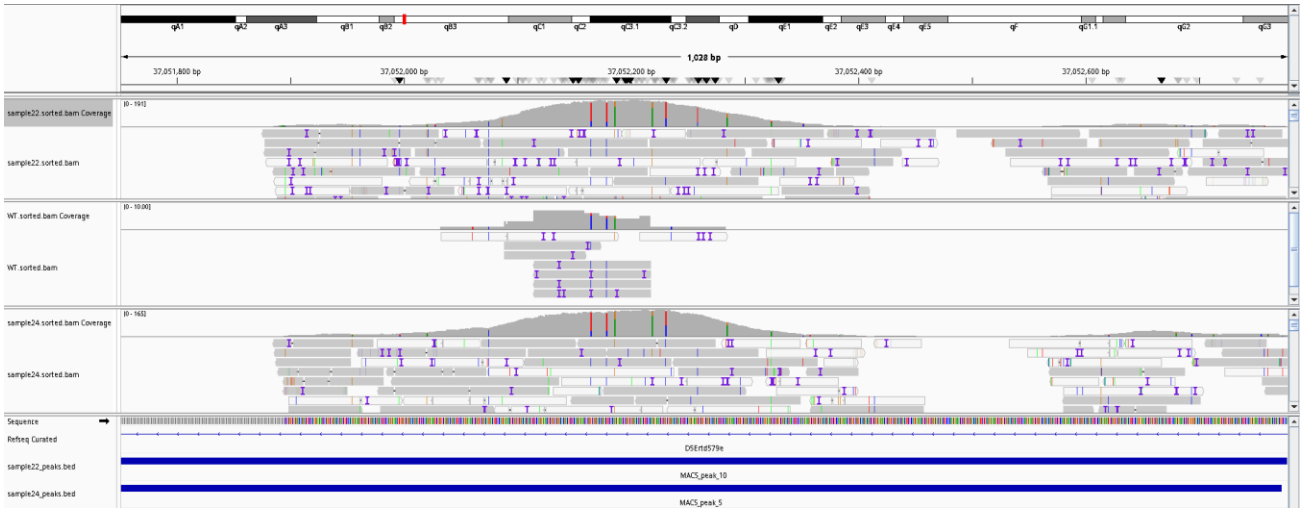


Figure 3: Example IGV visualization

- **Grey histograms** (coverage tracks) show read density.
- **Individual reads** (grey bars with colored marks) show sequencing alignment.
- **Blue peaks (BED tracks)** represent significant peaks called by MACS.
- **Annotations** (gene structures from RefSeq).

---

1. H4K20me1 Peaks (Histone Modification)

Comparison: WT vs. Set8KO

chr5: 37,051,751 - 37,052,777

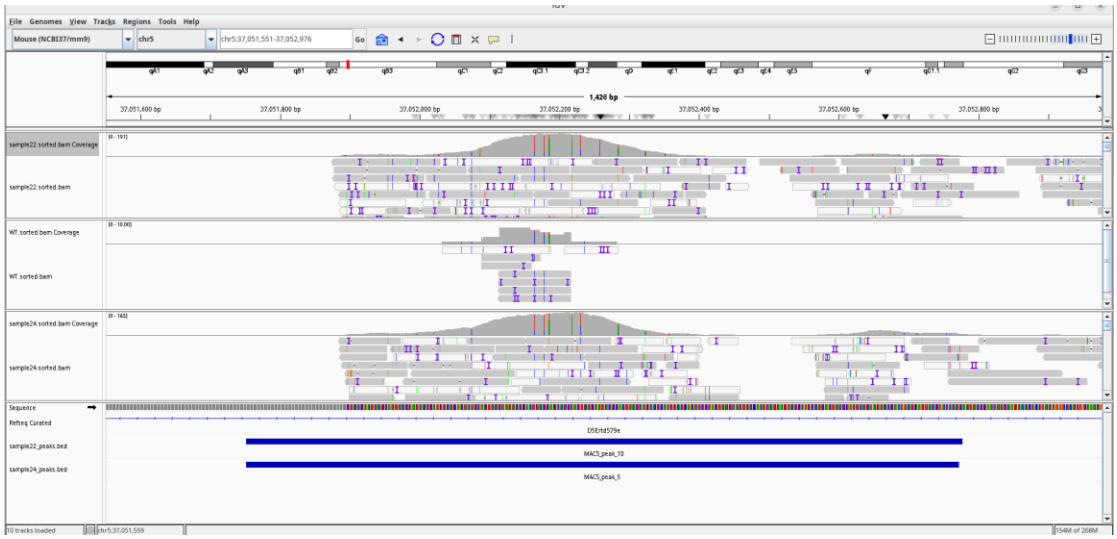


Figure 4: WT vs Set8KO, chr5: 37,051,751 - 37,052,777

- WT (sample22) shows a strong, high grey peak, while Set8KO (sample24) has a much lower signal.
- WT has clearly defined peaks (from sample22_peaks.bed), but the Set8KO signal is weak or absent, suggesting a loss of H4K20me1.
- Biological Interpretation: This loss in Set8KO may impair the regulation of gene regions involved in genome integrity and transcriptional control, potentially near a promoter or within a gene body.

chr8: 19,882,100 - 19,886,266

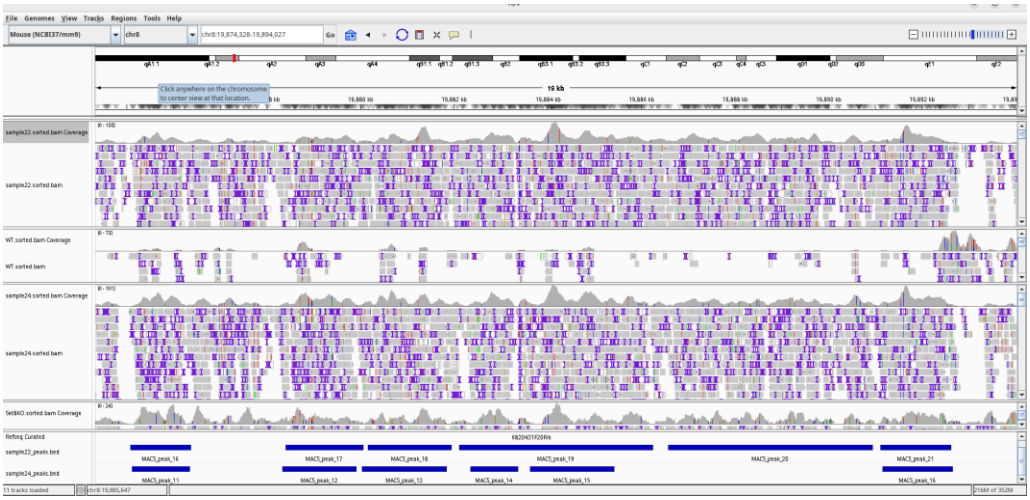


Figure 5: WT vs Set8KO, chr8: 19,882,100 - 19,886,266

- The WT sample exhibits robust H4K20me1 enrichment with clear, sharp peaks, whereas Set8KO shows reduced coverage.
- A distinct peak is visible in WT, which is largely diminished in Set8KO, indicating H4K20me1 depletion.
- Biological Interpretation: This region might represent a regulatory element, possibly an enhancer or promoter area, affecting transcription of genes that control metabolic pathways.

chr8: 20,009,469 - 20,020,058

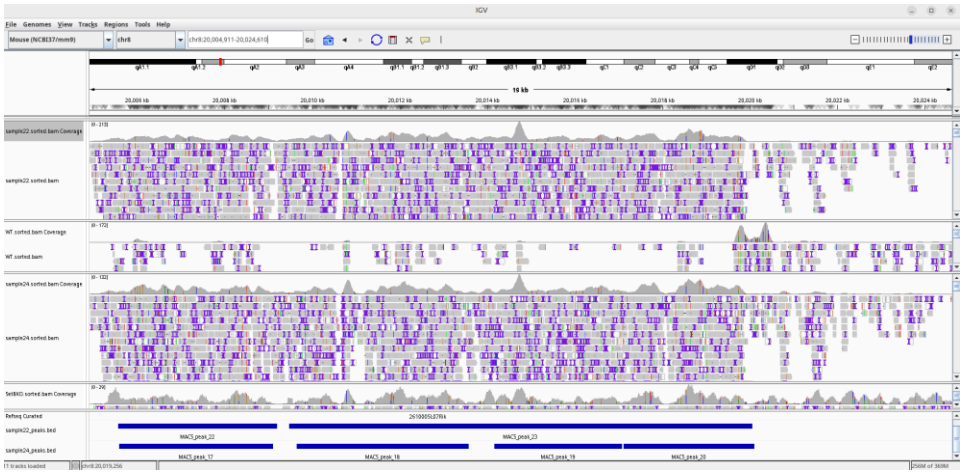


Figure 6: WT vs Set8KO, chr8: 20,009,469 - 20,020,058

- WT (sample22) demonstrates strong coverage and peak formation, while Set8KO (sample24) has a significantly reduced signal.
- The presence of a clear peak in WT versus a weakened signal in Set8KO points to a loss of H4K20me1.
- Biological Interpretation: Given the peak's breadth, it could span a promoter-proximal region or an enhancer, implying potential misregulation of genes critical for genome stability or metabolic regulation.

2. Pol II (Ser2) Peaks (Active Transcription)

Comparison: WT vs. Set8KO

chr12: 105,576,516 - 105,587,815

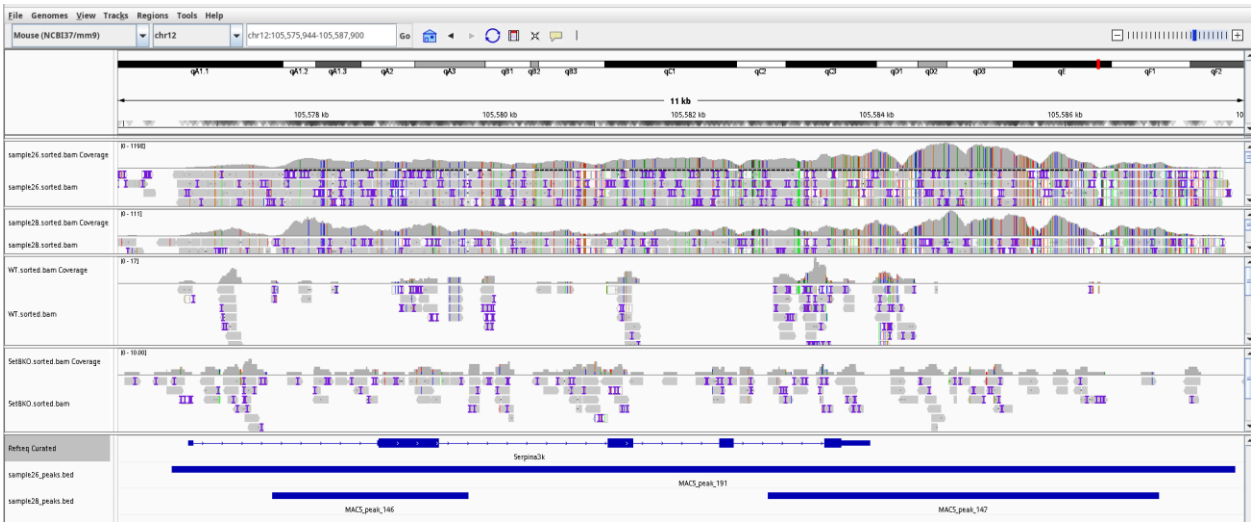


Figure 7: WT vs Set8KO, chr12: 105,576,516 - 105,587,815

- WT (sample26) shows strong Pol II (Ser2) coverage, while Set8KO (sample28) displays a noticeably lower signal.

- A well-defined peak in WT contrasts with the diminished signal in Set8KO, suggesting compromised transcription elongation.

- Biological Interpretation: The decrease in active Pol II could be due to defective promoter escape or elongation, possibly affecting genes essential for metabolic control and genome maintenance.

chr5: 90,902,436 - 90,907,497

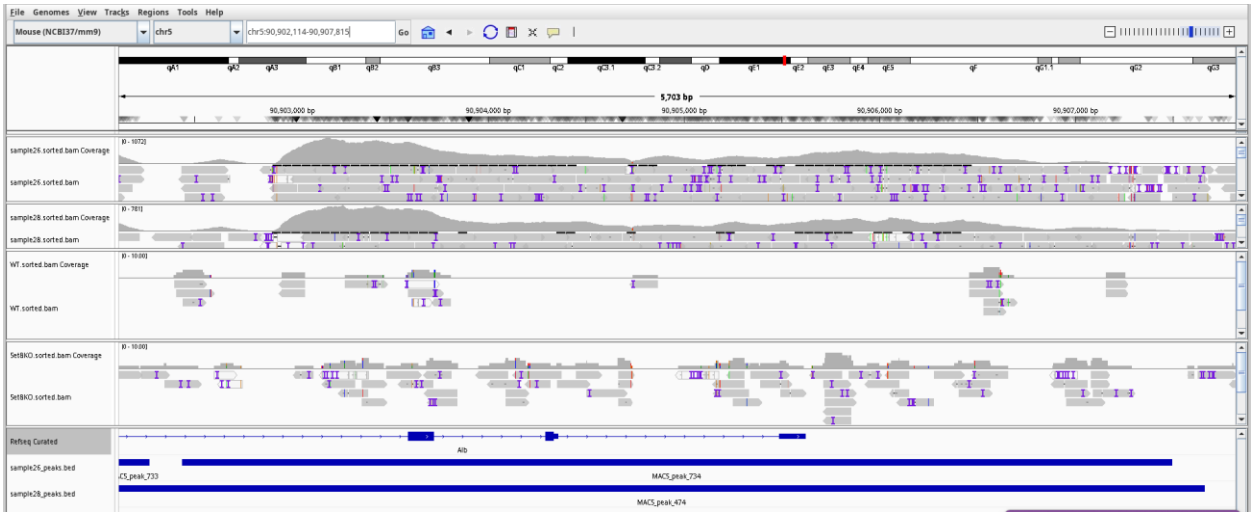


Figure 8: WT vs Set8KO, chr5: 90,902,436 - 90,907,497

- WT presents clear and strong Pol II signal, whereas Set8KO has reduced coverage.
- The robust peak in WT, versus the faint signal in Set8KO, indicates impaired Pol II progression in the knockout.
- Biological Interpretation: This could point to a disruption in transcriptional activity at key regulatory regions, potentially affecting the expression of genes involved in energy homeostasis.

chr5: 90,889,766 - 90,902,274

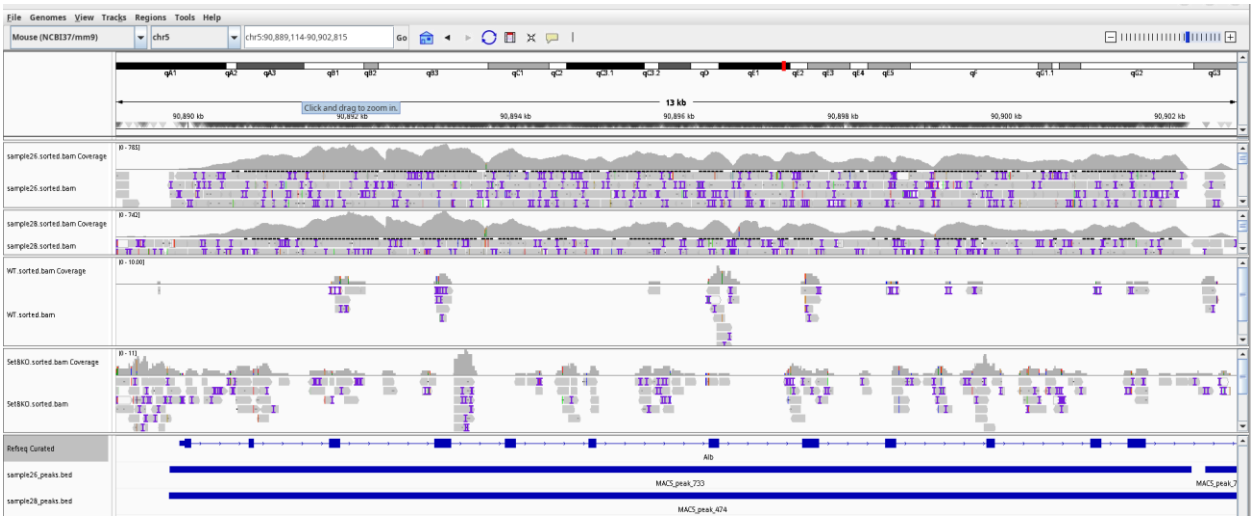


Figure 9: WT vs Set8KO, chr5: 90,889,766 - 90,902,274

- Again, WT (sample26) shows consistent Pol II enrichment while Set8KO (sample28) exhibits diminished signal.

- The overlap with the adjacent region reinforces a consistent loss of Pol II signal in the knockout condition.
- Biological Interpretation: This may reflect broader transcriptional defects, with the region likely encompassing part of a gene body or promoter region vital for regulating metabolic or cell cycle genes.

chr18: 65,403,775 - 65,413,812

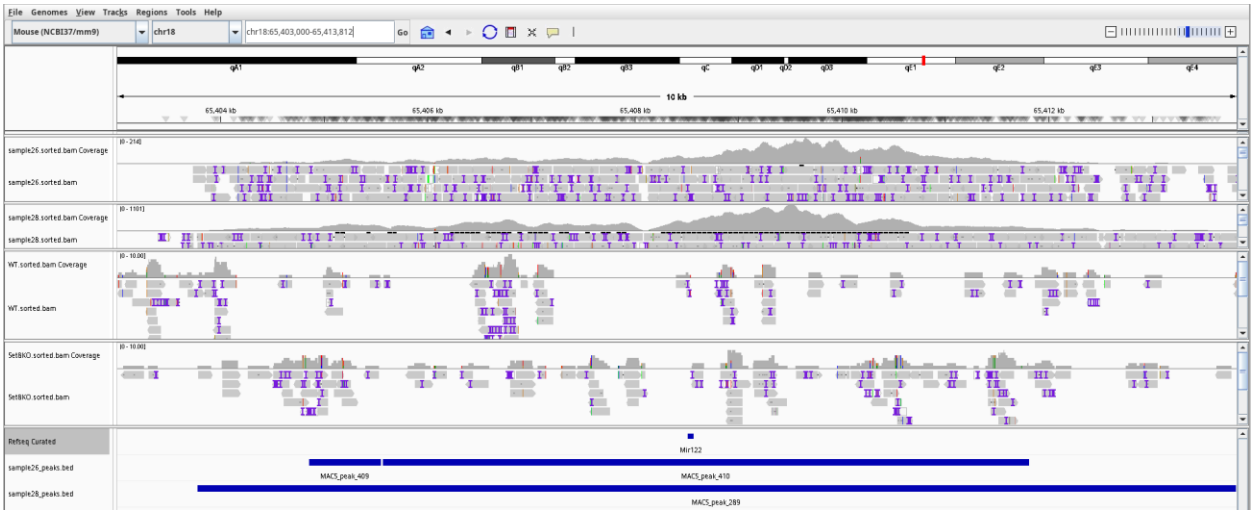


Figure 10: WT vs Set8KO, chr18: 65,403,775 - 65,413,812

- WT has a strong, continuous Pol II signal compared to a significantly lower signal in Set8KO.
- The WT peak is sharp and defined, while Set8KO's signal is sparse, indicating disrupted elongation.
- Biological Interpretation: This disruption could be associated with reduced transcription efficiency at promoter or gene body regions of genes involved in genomic stability or metabolism.

chr4: 150,228,922 - 150,249,614

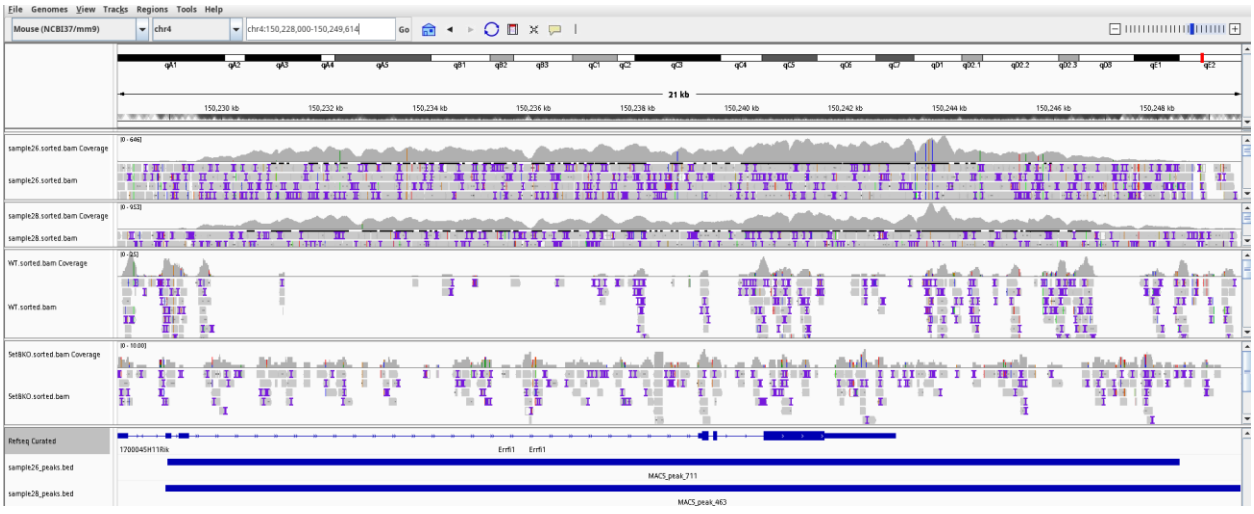


Figure 11: WT vs Set8KO, chr4: 150,228,922 - 150,249,614

- A robust Pol II peak in WT is evident, whereas Set8KO shows reduced signal intensity.
- The loss of a clear peak in Set8KO implies a defect in transcriptional regulation.
- Biological Interpretation: This area may include promoter or enhancer elements, and the altered Pol II signal could impact the expression of genes that regulate chromatin dynamics or metabolic pathways.

3. Phf8 Peaks (Histone Demethylase)

Comparison: WT vs. Set8KO

chr9: 24,346,290 - 24,347,072

Figure 12: WT vs Set8KO, chr9: 24,346,290 - 24,347,072

- WT exhibits a distinct Phf8 peak, while Set8KO shows much lower binding intensity.
- The clear peak in WT (from sample30) versus the diminished peak in Set8KO (sample32) suggests reduced recruitment of Phf8.
- Biological Interpretation: As Phf8 is important for transcriptional activation, its loss here might affect nearby promoter regions of genes linked to cell cycle regulation or metabolism.

chr3: 81,454,291 - 81,455,152

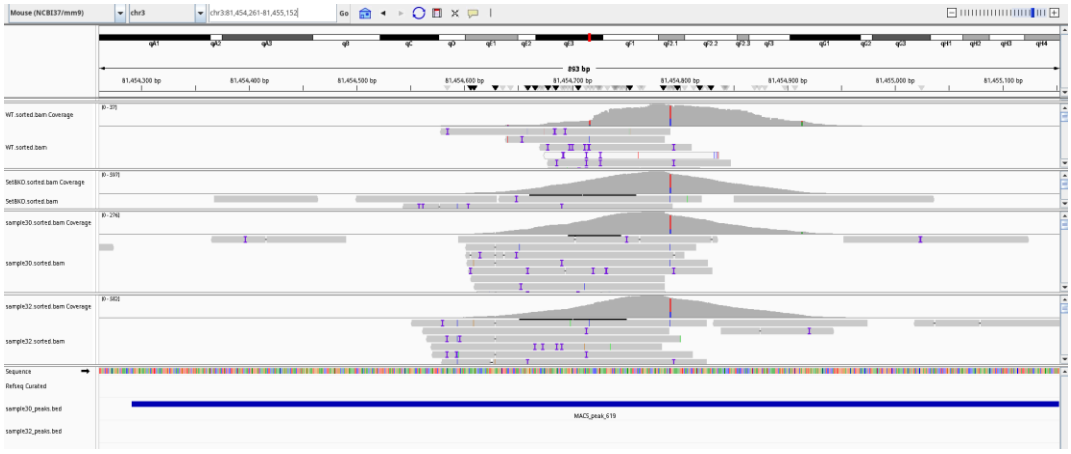


Figure 13: WT vs Set8KO, chr3: 81,454,291 - 81,455,152

- Strong Phf8 enrichment in WT is evident compared to a significantly reduced signal in Set8KO.
- The peak present in WT is notably weaker in Set8KO, implying compromised Phf8 binding.
- Biological Interpretation: This could reflect changes in the chromatin state at regulatory elements, potentially affecting gene transcription at nearby promoters or enhancers.

chr11: 5,661,669 - 5,662,781

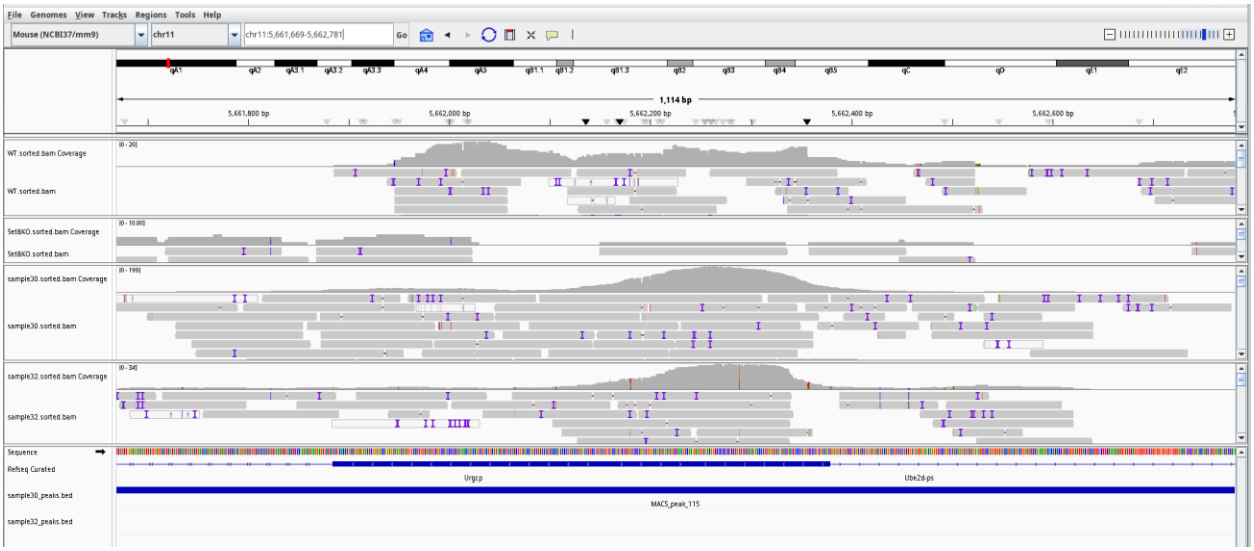


Figure 14: WT vs Set8KO, chr11: 5,661,669 - 5,662,781

- WT shows a robust Phf8 peak, while Set8KO demonstrates a clear loss or reduction in signal.
- The difference in peak intensity indicates a loss of Phf8 recruitment in the absence of proper H4K20me1 levels.

- Biological Interpretation: This may impact transcription initiation at promoters, affecting genes that control chromatin structure or metabolic processes.

chr9: 82,464,202 - 82,464,962

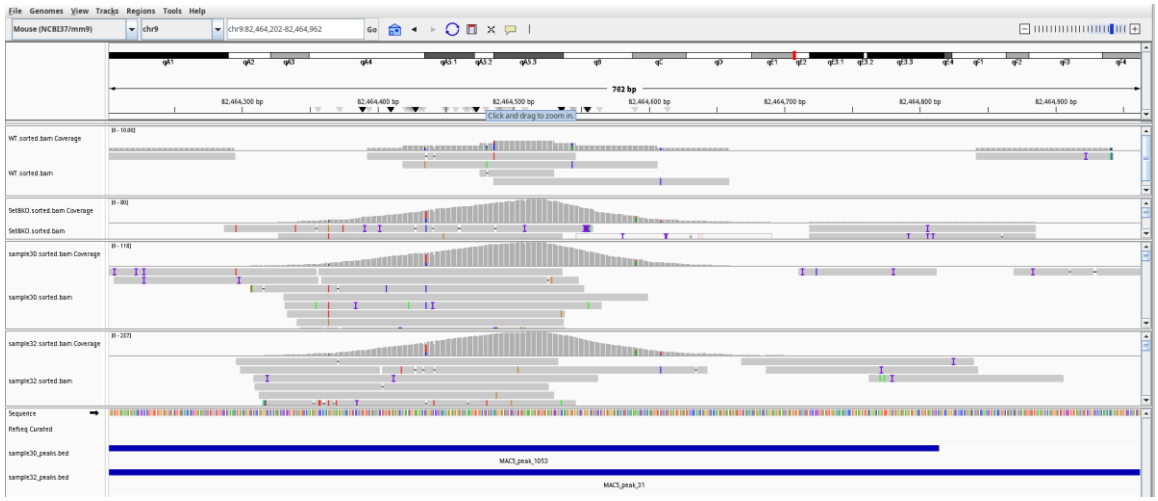


Figure 15: WT vs Set8KO, chr9: 82,464,202 - 82,464,962

- A prominent Phf8 signal in WT contrasts with an attenuated signal in Set8KO.
- The peak is strong in WT but barely detectable in Set8KO, suggesting that H4K20me1 is necessary for Phf8 stability or recruitment.
- Biological Interpretation: The affected region might be near key enhancer or promoter sites, potentially altering the expression of genes involved in transcriptional regulation.

chr2: 36,823,425 - 36,824,096

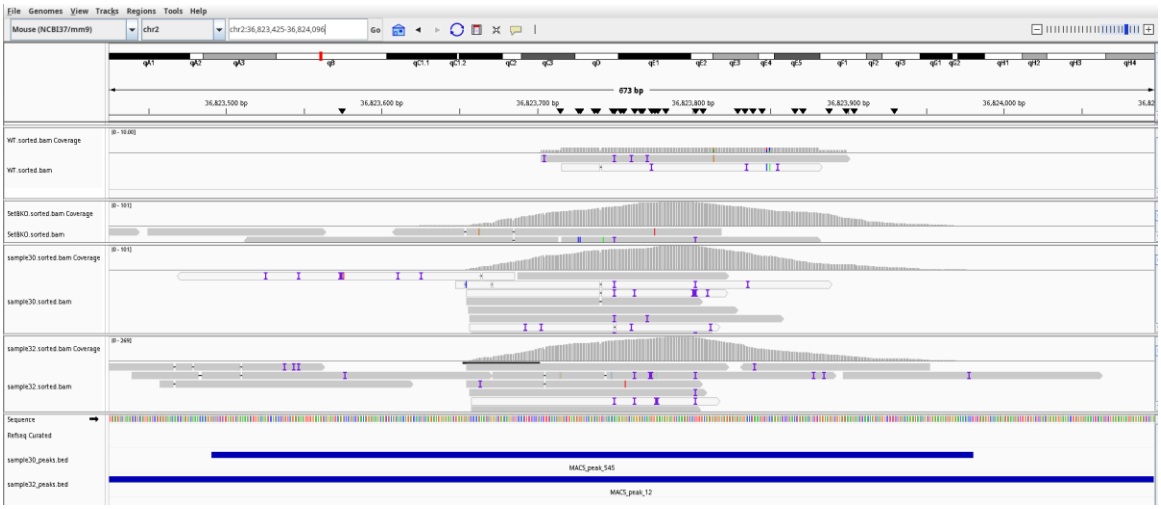


Figure 16: WT vs Set8KO, chr2: 36,823,425 - 36,824,096

- WT displays a clear Phf8 peak, whereas Set8KO shows a substantial loss of signal.
- The defined peak in WT compared to its loss in Set8KO reinforces the link between H4K20me1 and Phf8 binding.
- Biological Interpretation: This loss may lead to altered transcription at nearby regulatory elements, potentially impacting genes involved in DNA repair or metabolic pathways.

4.SDS1 Pol II Peaks (Alternative Condition)

Comparison: WT vs. Set8KO

chr11: 87,256,384 - 87,257,401

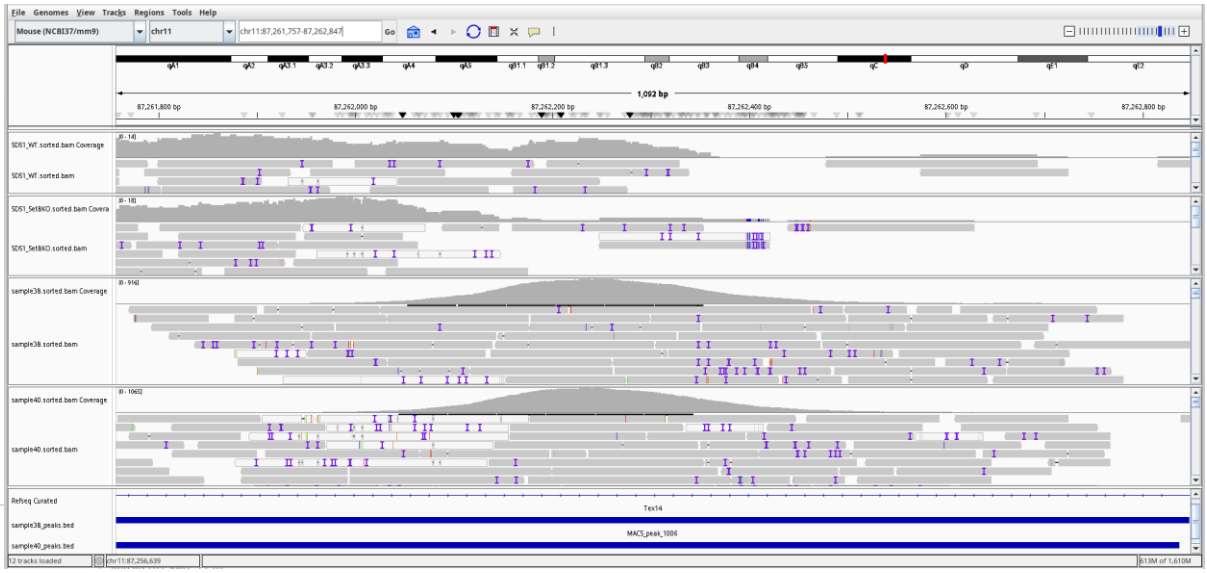


Figure 17: SDS1_WT vs SDS1_Set8KO, chr11: 87,256,384 - 87,257,401

- Under the SDS1 condition, WT has a strong Pol II signal while Set8KO displays reduced coverage.
- The WT peak is clearly defined, whereas the Set8KO peak is either weaker or missing.
- Biological Interpretation: This suggests that the transcriptional machinery is less efficiently recruited or maintained in Set8KO, potentially affecting gene expression near promoter regions.

chr11: 87,261,757 - 87,262,847

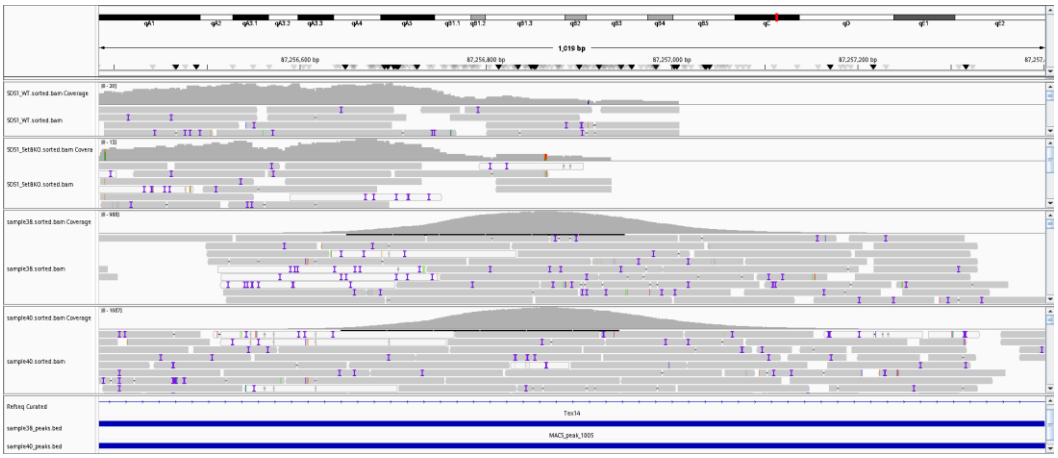


Figure 18: SDS1_WT vs SDS1_Set8KO, chr11: 87,261,757 - 87,262,847

- A similar trend is observed where WT shows a distinct Pol II peak, but Set8KO has a markedly lower signal.
- The WT-specific peak suggests active transcription that is compromised in the knockout.
- Biological Interpretation: This pattern implies that proper H4K20me1 levels are crucial for sustaining Pol II dynamics at this locus, which may regulate key genes through nearby enhancers or promoter elements.

chr11: 87,256,186 - 87,257,543

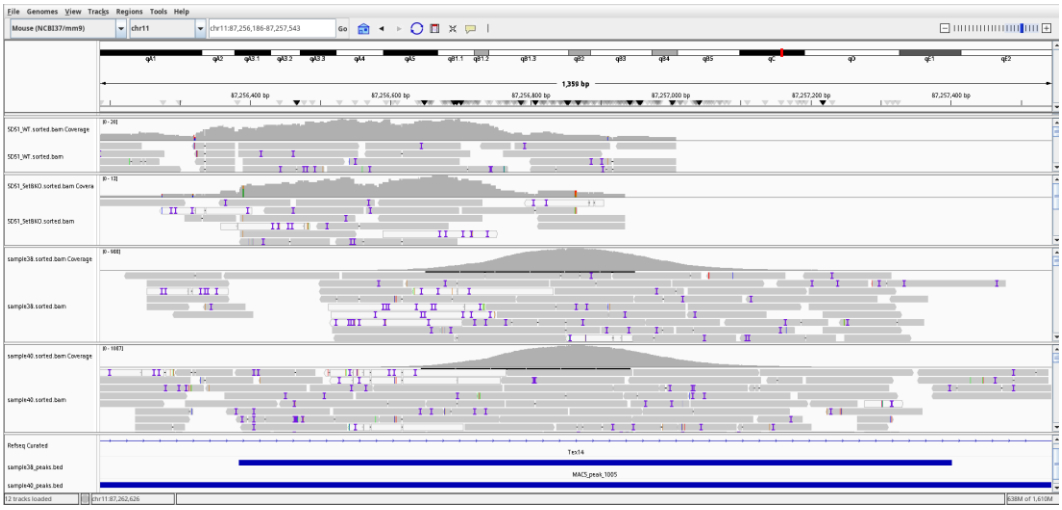


Figure 19: SDS1_WT vs SDS1_Set8KO, chr11: 87,256,186 - 87,257,543

- WT again shows consistent Pol II enrichment, while Set8KO exhibits a loss of signal.
- The clear peak in WT versus the diminished signal in Set8KO highlights an effect on transcriptional engagement.
- Biological Interpretation: This loss of Pol II signal under SDS1 conditions could lead to impaired transcription initiation or elongation at a region likely harboring regulatory sequences.

chr11: 87,235,807 - 87,237,138

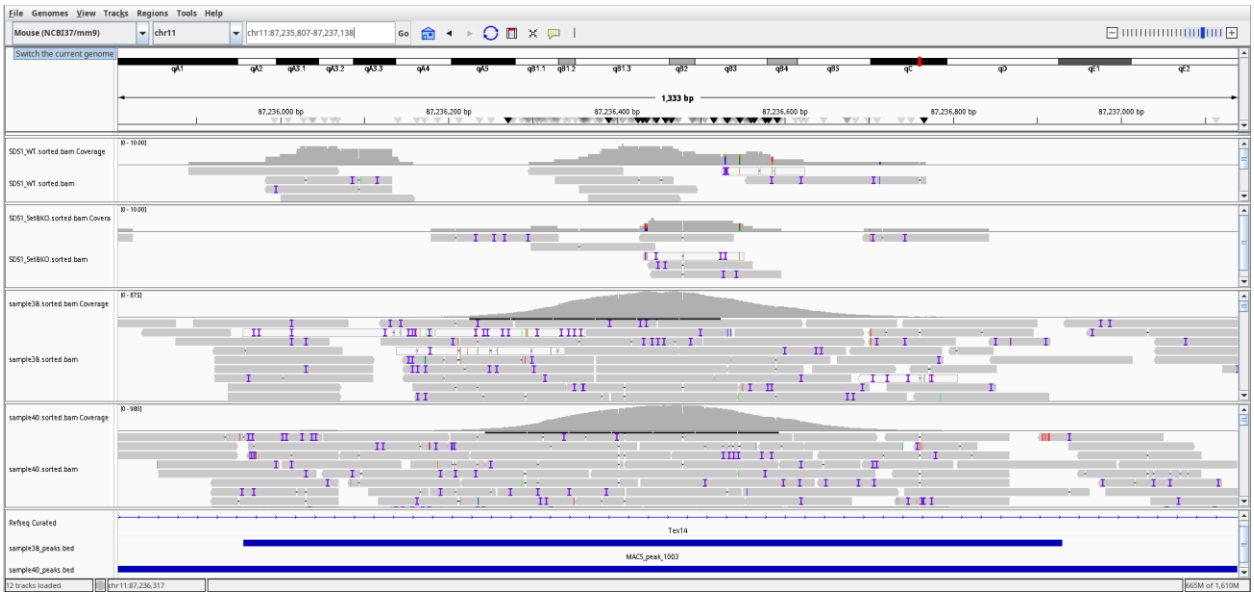


Figure 20: SDS1_WT vs SDS1_Set8KO, chr11: 87,235,807 - 87,237,138

- The WT sample has a strong Pol II peak, while the Set8KO condition shows a substantially lower signal.
- The defined peak in WT compared to its near absence in Set8KO reinforces the impact of H4K20me1 loss on Pol II binding.
- Biological Interpretation: Such differences suggest that transcriptional activation at this locus is compromised in Set8KO, potentially affecting genes that rely on this region's promoter or enhancer activity.

4.4 Motifs

The MEME analysis results provided in the text files (meme22.txt, meme24.txt, meme26.txt, meme28.txt, meme38.txt, and meme40.txt) contain information about discovered motifs in the ChIP-seq samples. These motifs are likely to be associated with transcription factor binding sites or other regulatory elements that play a role in the regulation of gene expression, particularly in the context of H4K20Me1 and RNA Polymerase II (Pol II) dynamics.

meme22.txt (ChIP WT H4K20me1)

- Motif: WGATGCCTKTGBWKC
- Width: 15
- Sites: 13
- E-value: 9.7e-001

Comment: This motif is relatively short and has a moderate E-value, suggesting it may be a common sequence element in the WT H4K20me1 ChIP-seq data. The presence of this motif in WT samples but not in Set8KO samples could indicate a role in maintaining H4K20me1 levels, potentially influencing transcription regulation or chromatin structure in wild-type conditions.

meme24.txt (ChIP Set8KO H4K20me1)

- Motif: GAGABAGGGTYKCWNBTKGCHKCCBTGG
- Width: 2
- Sites:
- E-value: 8.3e-002

Comment: This motif is longer and more complex, with a lower E-value, suggesting it may be more specific to the Set8KO condition. The presence of this motif in Set8KO samples could indicate a compensatory mechanism or a shift in regulatory elements due to the loss of Set8, which is responsible for H4K20me1 methylation. This motif might be associated with altered transcriptional regulation in the absence of Set8.

meme26.txt (ChIP Serin2 WT PolII)

- Motif: KSMATKSMATKSMATKSMATKSMATKSMA
- Width: 29
- Sites: 31
- E-value: 4.6e-174

Comment: This motif is highly significant, with a very low E-value, indicating it is a strong and recurrent sequence element in the WT Pol II ChIP-seq data. The repetitive nature of the motif suggests it may play a role in Pol II binding or progression, potentially influencing transcription elongation or termination in wild-type conditions.

meme28.txt (ChIP Serin2 Set8KO PolII)

- Motif: CWGAGRRGWGGRRGG
- Width: 15
- Sites: 55
- E-value: 1.4e-007

Comment: This motif is also significant, with a low E-value, and is present in a large number of sites. The presence of this motif in Set8KO Pol II samples could indicate changes in Pol II binding or transcriptional regulation in the absence of Set8. The motif may be associated with altered transcriptional dynamics, such as promoter escape or elongation, in Set8KO cells.

meme30.txt (ChIP WT Phf8)

- Motif: WAWRDCRWGAAAAMTGAVMAW
- Width: 21
- Sites: 9
- E-value: 7.6e-016

Comment: This motif is highly significant and may be associated with Phf8 binding sites in WT conditions. Phf8 is a histone demethylase that targets H4K20me1, so this motif could be involved in the regulation of genes that are sensitive to H4K20me1 levels. The presence of this motif in WT Phf8 samples suggests it may play a role in the demethylation process or in the regulation of metabolic genes.

meme32.txt (ChIP Set8KO Phf8)

- Motif: TATCCYYKWGAWSWCYTTTTTGYTRTCS
- Width: 28
- Sites: 5
- E-value: 3.1e-014

Comment: This motif is also highly significant and may represent a regulatory element specific to the Set8KO condition in the context of Phf8 binding. The presence of this motif in Set8KO Phf8 samples could indicate changes in the regulatory landscape due to the loss of Set8, potentially affecting Phf8's ability to demethylate H4K20me1 and regulate gene expression.

Overall Observations:

WT vs. Set8KO: The motifs identified in WT and Set8KO samples differ significantly, suggesting that the loss of Set8 (and consequently H4K20me1) leads to changes in the regulatory elements that control transcription. This is consistent with the study's findings that H4K20me1 is crucial for transcriptional regulation, particularly in metabolic genes.

Pol II Motifs: The motifs identified in Pol II ChIP-seq samples (both WT and Set8KO) are highly significant and may be involved in the regulation of transcription elongation and termination. The differences between WT and Set8KO Pol II motifs could reflect changes in transcriptional dynamics due to the loss of H4K20me1.

Phf8 Motifs: The motifs identified in Phf8 ChIP-seq samples are also highly significant and may be involved in the regulation of genes sensitive to H4K20me1 levels. The differences between

WT and Set8KO Phf8 motifs could indicate changes in the regulatory landscape due to the loss of Set8.

4.5 Peak Analyzer 1.4 – Genes Annotation

Using PeakAnalyzer 1.4 and the "Peak Annotation" choice, it is possible to extract some statistics about the peaks and their overlapping with the gene features.

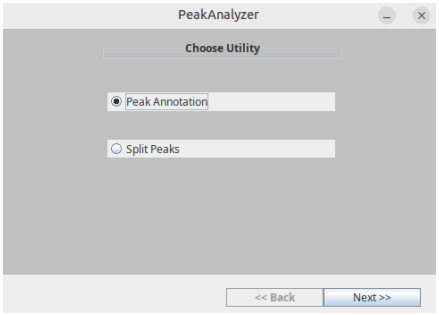


Figure 21: Peak Analyzer GUI

The **pie chart** shows the percentage of sequencing reads that fall within different genomic features. This helps understand where the reads are concentrated relative to genes.

The **bar graph** measures the distribution of distances from the read peaks to the nearest downstream gene. This helps determine if the reads are located near genes and how far away they tend to be.

- Categories: The categories represent distance ranges from the peak to the nearest downstream gene: 0-1kb, 1-3kb, 3-5kb, 5-10kb, 10-100kb, and >100kb.

Sample 22:

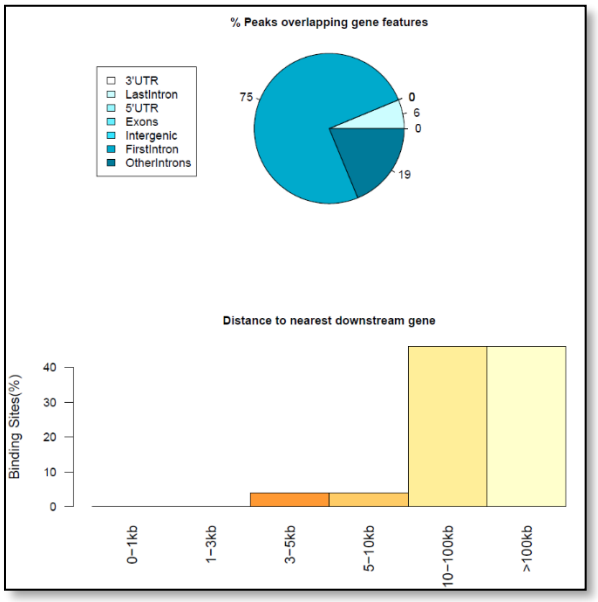


Figure 22: Sample 22 peak analyzer diagrams

This figure shows the distribution of sequencing reads relative to gene features and neighboring genes. A strong enrichment at 5'UTRs (75%) suggests potential roles in transcription initiation or early elongation, while a substantial portion in intergenic regions (19%) hints at other regulatory functions. The bimodal distribution of distances to the nearest downstream gene, with peaks at 0-1kb and >10kb, implies complex relationships with gene regulation, possibly through proximal and distal regulatory elements. Further comparison between WT and Set8 KO samples, along with integration of Pol II and Phf8 data, will be crucial to understand the functional implications of these patterns in the context of H4K20me1.

Sample 24:

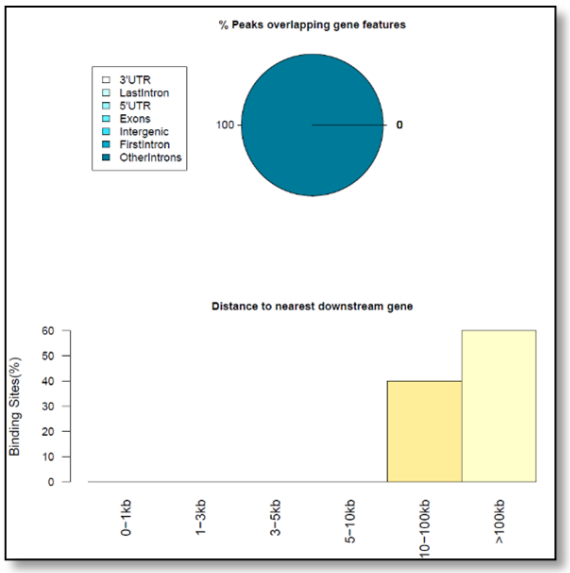


Figure 23: Sample 24 peak analyzer diagrams

This figure, likely showing a different H4K20me1 sample (perhaps Set8 KO), reveals a dramatic shift in distribution compared to the previous one. The complete absence of reads overlapping any gene features (pie chart) suggests a global loss or severe reduction of H4K20me1 at these locations. The distance to nearest downstream gene now shows a strong preference for regions beyond 10kb, with a dominant peak at >100kb. This indicates a potential redistribution of H4K20me1 to very distant sites, possibly reflecting changes in large-scale chromatin organization or heterochromatin spreading in the Set8 KO. Comparing this pattern to the WT distribution will be critical to understand the impact of Set8 loss on H4K20me1 localization and its downstream consequences for transcription and genome integrity.

Sample 26:

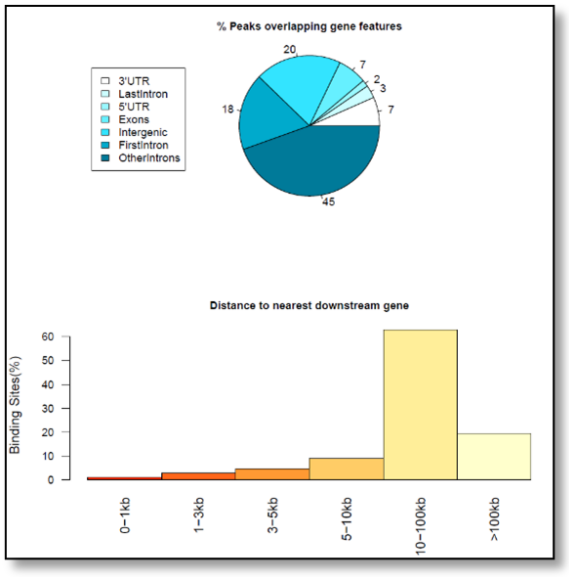


Figure 24: Sample 26 peak analyzer diagrams

This figure, likely representing a WT H4K20me1 sample, shows a more complex distribution than the previous Set8 KO examples. The pie chart reveals H4K20me1 presence across various gene features, with the strongest enrichment in intergenic regions (45%), followed by 5'UTRs (18%) and last introns (20%). This suggests a broader distribution of H4K20me1 in WT cells, not solely focused on 5'UTRs as seen in the first figure, but also significantly present at the 3' end of genes (last introns). The distance to nearest gene plot shows a dominant peak at 10-100kb, indicating that a large proportion of H4K20me1 in WT is located at considerable distances from genes, potentially at regulatory elements or participating in higher-order chromatin structure. While there's still a noticeable presence beyond 100kb, the 10-100kb peak suggests a different distribution compared to the Set8 KO, where the signal was primarily beyond 100kb. This WT profile provides a baseline against which to compare the Set8 KO, highlighting how loss of Set8 dramatically alters H4K20me1 localization. The next crucial step will be to directly compare these WT and Set8 KO distributions to pinpoint specific changes in H4K20me1 marking that contribute to the observed phenotypes.

Sample 28:

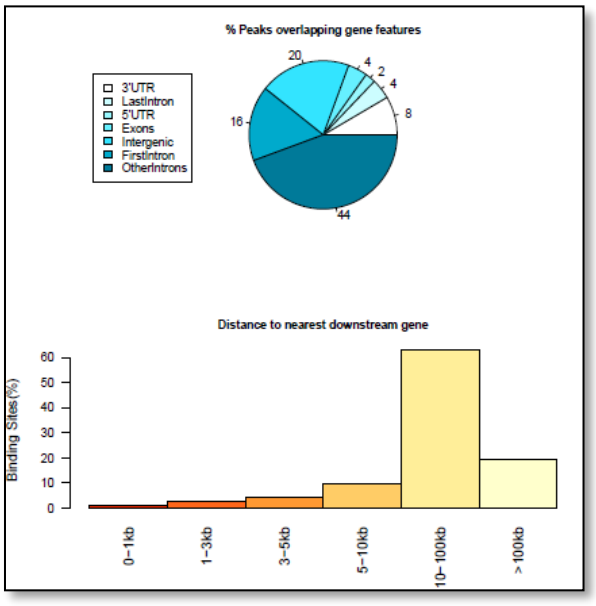


Figure 25: Sample 28 peak analyzer diagrams

This figure, likely depicting a WT H4K20me1 sample, reveals a more nuanced distribution of this histone modification. The pie chart shows a distribution across various gene features, with the highest proportion found in intergenic regions (44%), followed by 5'UTRs (16%) and last introns (20%). This suggests a multifaceted role for H4K20me1 in WT cells, not exclusively concentrated at 5'UTRs but also significantly present at the 3' ends of genes and in the regions between genes. The distance to the nearest gene plot shows a dominant peak at 10-100kb, indicating a preference for H4K20me1 localization at considerable distances from genes, potentially at distal regulatory elements or regions involved in higher-order chromatin structure. While some signal is observed beyond 100kb, the prominent 10-100kb peak suggests a different distribution compared to the Set8 KO, where the signal was primarily beyond 100kb. This WT profile serves as a crucial baseline for comparison with the Set8 KO, allowing for the identification of specific changes in H4K20me1 marking upon Set8 loss, which could then be linked to alterations in gene expression and genome integrity.

Sample 30:

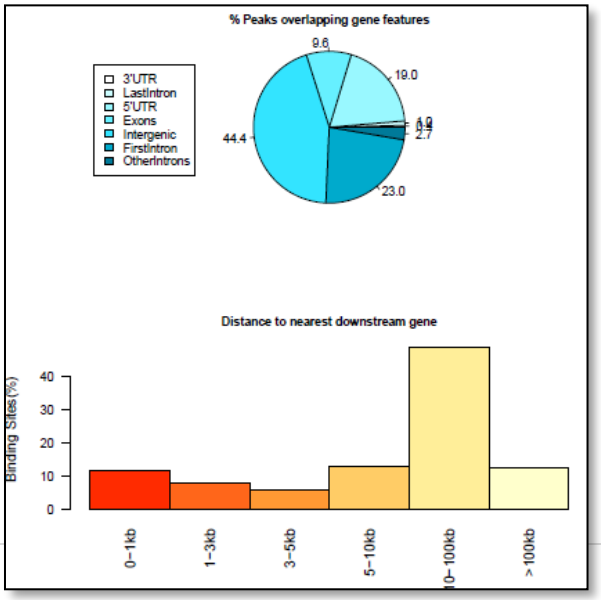


Figure 26: Sample 30 peak diagrams

This figure shows the distribution of a ChIP-seq signal across gene features and distances to the nearest downstream gene. The signal is enriched in intergenic regions (44.4%), exons (23%), and 5'UTRs (19%), suggesting diverse roles in gene regulation and other functions. The distance plot reveals a bimodal distribution with peaks at 0-1kb and 10-100kb, indicating influence on both proximal and distal gene regulation. Further analysis and comparison with other samples are needed to determine the specific function of this mark or protein.

Sample 32:

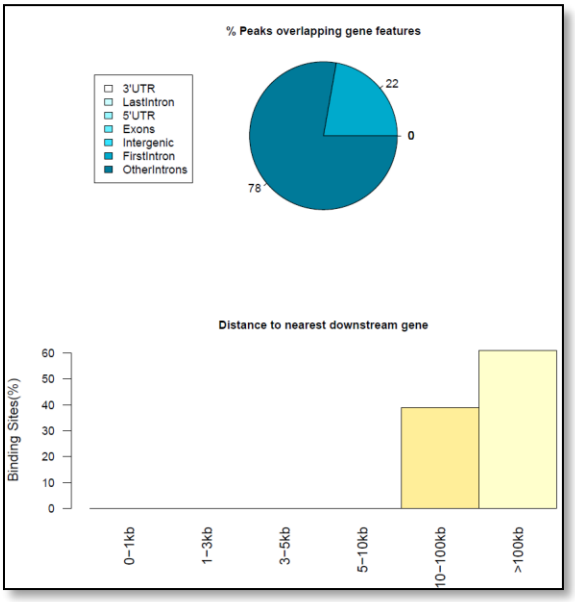


Figure 27: Sample 32 peak diagrams

This figure shows the distribution of a ChIP-seq signal. The pie chart reveals a strong preference for intergenic regions (78%) and 3'UTRs (22%), with no signal at 5'UTRs or exons. The distance plot confirms this, showing localization primarily at regions far from genes (>10kb and >100kb). This suggests a role distinct from typical transcriptional regulation, possibly in long-range chromatin interactions or non-coding RNA regulation.

Sample 38:

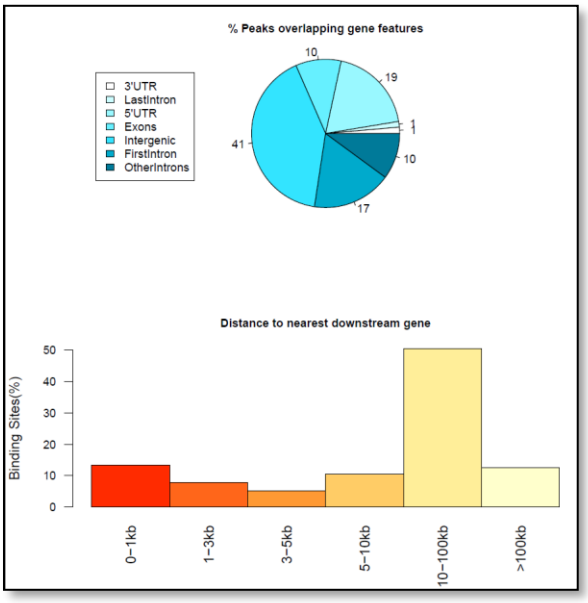


Figure 28: Sample 38 peak diagrams

This figure displays the distribution of a ChIP-seq signal across gene features and distances to the nearest downstream gene. The pie chart shows enrichment in intergenic regions (41%), exons (17%), 5'UTRs (19%), and last introns (10%), suggesting diverse roles. The distance plot reveals a bimodal distribution with peaks at 0-1kb and 10-100kb, indicating influence on both proximal and distal gene regulation. Further investigation is needed to determine the specific function of this mark or protein.

Sample 40:

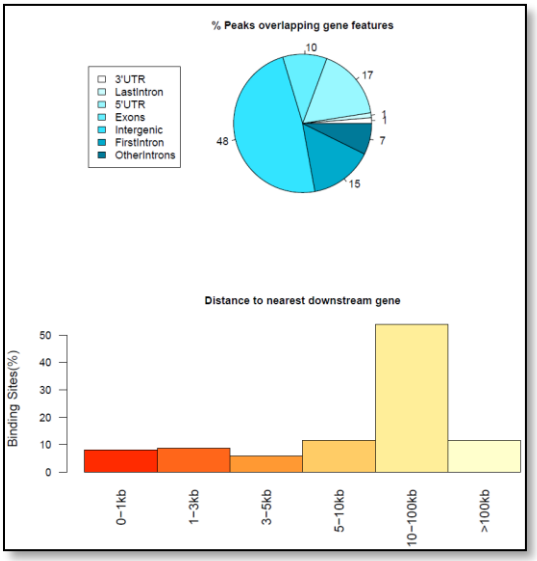


Figure 29: Sample 40 peak diagrams

This figure shows the distribution of a ChIP-seq signal. The pie chart shows the signal distributed across gene features, with the highest proportion in intergenic regions (48%), followed by exons (17%), 5'UTRs (15%), and last introns (10%). This suggests a multifaceted role. The distance plot shows a dominant peak at 10-100kb and a smaller peak at 0-1kb, indicating influence on both distal and proximal gene regulation.

5. References

Paper source:

Nikolaou, K.C., Moulos, P., Harokopos, V., Chalepakis, G. and Talianidis, I., 2017. Kmt5a controls hepatic metabolic pathways by facilitating RNA Pol II release from promoter-proximal regions. *Cell Reports*, 20(4), pp.909-922. Available at: https://doi.org/10.1016/j.celrep.2017.07.003 [Accessed 10 Jan. 2025].

Dataset source:

Nikolaou, K.C., Moulos, P., Harokopos, V., Chalepakis, G. and Talianidis, I., 2017. Kmt5a controls hepatic metabolic pathways by facilitating RNA Pol II release from promoter-proximal regions. Gene Expression Omnibus (GEO), Dataset GSE97338. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97338 [Accessed 10 Jan. 2025].

Websites:

UCSC Genome Browser - mm9 Genome Build:
Available at: https://hgdownload.soe.ucsc.edu/downloads.html#mouse

Tools used:

- Bowtie2
- MACS (Model-based Analysis of ChIP-Seq)
- MEME (Motif Discovery Tool)
- FastQC
- SAMtools
- Integrative Genomics Viewer (IGV)

Supplementary Material

1. **Scripts**
2. **FastQC reports**
3. **Screenshots (log file)**

---

1. Scripts

1.1 ChIP_seq analysis (1st script)

```
#!/bin/bash
# Directory containing the FASTQ files
FASTQ_DIR="Bio_Project"

# Bowtie2 index and genome files
BOWTIE2_INDEX="Bio_Project/bowtie2_index/mm9"
GENOME="Bio_Project/bowtie2_index/mouse.mm9.genome"
GENOME_FA="Bio_Project/bowtie2_index/mm9.fa"

# List of samples
SAMPLES=("SRR5409158.fastq.gz:sample22"
  "SRR5409160.fastq.gz:sample24"
  "SRR5409162.fastq.gz:sample26"
  "SRR5409164.fastq.gz:sample28"
  "SRR5409166.fastq.gz:sample30"
  "SRR5409168.fastq.gz:sample32"
  "SRR5409174.fastq.gz:sample38"
  "SRR5409176.fastq.gz:sample40")
```

```
# Step 1: Build Bowtie2 index for mm9 genome
echo "Building Bowtie2 index for mm9 genome..."
bowtie2-build "$GENOME_FA" "$BOWTIE2_INDEX"
echo "Bowtie2 index for mm9 genome built."

# Step 2: Create a directory for FASTQC reports
mkdir -p "$FASTQ_DIR/fastqc_reports"

# Step 3: Run FASTQC for all samples
echo "Running FASTQC for all samples..."
process_step "fastqc"

# Step 4: Process all steps for all samples
echo "Starting alignment for all samples..."
process_step "align"

echo "Converting SAM to BAM for all samples..."
process_step "convert_bam"

echo "Sorting BAM files for all samples..."
process_step "sort_bam"

echo "Indexing sorted BAM files for all samples..."
process_step "index_bam"

echo "Generating flagstat for all samples..."
process_step "flagstat"

echo "Generating BedGraph for all samples..."
process_step "bedgraph"

echo "Converting BedGraph to BigWig for all samples..."
process_step "bigwig"

echo "All samples processed."
```

```
# Function to process each sample step by step
process_step() {
  local step="$1"
  local file_suffix="$2"

  for sample in "${SAMPLES[@]}"; do
    fastq_file="${sample%%:*}"
    sample_name="${sample##*:}"

    echo "Processing $step for $sample_name..."
    case "$step" in
        "fastqc")

        # Step 0: Run FASTQC on the FASTQ file
        echo "Running FASTQC on $fastq_file for $sample_name..."
        fastqc "$FASTQ_DIR/$fastq_file" -o "$FASTQ_DIR/fastqc_reports"
         ;;
      "align")

        # Step 1: Align with Bowtie2 and create SAM file
        echo "Aligning $fastq_file with Bowtie2 for $sample_name..."
        bowtie2 -p 4 -x "$BOWTIE2_INDEX" -U "$FASTQ_DIR/$fastq_file" -S
"$sample_name.sam"
         ;;

     "convert_bam")

        # Step 2: Convert SAM to BAM
        echo "Converting SAM to BAM for $sample_name..."
        samtools view -bSo "$sample_name.bam" "$sample_name.sam"

        # Remove SAM file after conversion to save space
        rm -f "$sample_name.sam"
        ;;

      "sort_bam")

        # Step 3: Sort BAM file
        echo "Sorting BAM file for $sample_name..."
        samtools sort -o "$sample_name.sorted.bam" "$sample_name.bam"

        # Remove original BAM after sorting
        rm -f "$sample_name.bam"
        ;;
```

```
1
2
3       "index_bam")
4
5         # Step 4: Index sorted BAM file
6         echo "Indexing sorted BAM file for $sample_name..."
7         samtools index "$sample_name.sorted.bam"
8         ;;
9
10      "flagstat")
11
12        # Step 5: Generate flagstat statistics
13        echo "Generating flagstat for $sample_name..."
14        samtools flagstat "$sample_name.sorted.bam" >
15        "$sample_name.flagstat.txt"
16        ;;
17
18      "bedgraph")
19
20        # Step 6: Generate BedGraph file
21        echo "Generating BedGraph for $sample_name..."
22        genomeCoverageBed -bg -ibam "$sample_name.sorted.bam" -g "$GENOME" >
  "$sample_name.bedgraph"
23        ;;
24
25      "bigwig")
26
27        # Step 7: Convert BedGraph to BigWig
28        echo "Converting BedGraph to BigWig for $sample_name..."
29        bedGraphToBigWig "$sample_name.bedgraph" "$GENOME" "$sample_name.bw"
30        ;;
31
32      esac
33
34    done
35  }
```

1.2 Alignment input control (2nd script)

```
#!/bin bash
# Directories
GENOME="Bio_Project/bowtie2_index/mouse.mm9.genome"
FASTQ_DIR="Bio_Project"
#Samples list
SAMPLES=("SRR5409170.fastq.gz:WT"
"SRR5409172.fastq.gz:Set8KO"
"SRR5409178.fastq.gz:SDS1 WT"
"SRR5409180.fastq.gz:SDS1 Set8KO")
# Align reads, convert, sort, and index for each sample
for SAMPLE in "${SAMPLES[@]}"; do
# Extract filename and sample name
FILE=$(echo "$SAMPLE" | cut -d':' -f1)
NAME=$(echo "$SAMPLE" | cut -d':' -f2)
echo "Processing sample: $NAME ($FILE)..."
# Align reads
bowtie2 -x "$GENOME" -U "$FASTQ_DIR/$FILE" -S "$FASTQ_DIR/$NAME.sam" --very-
sensitive -k 1
# Convert SAM to BAM
samtools view -bS "$FASTQ_DIR/$NAME.sam" > "$FASTQ_DIR/$NAME.bam"
# Sort BAM file
samtools sort -o "$FASTQ_DIR/$NAME.sorted.bam" "$FASTQ_DIR/$NAME.bam"
# Index sorted BAM file
samtools index "$FASTQ_DIR/$NAME.sorted.bam"
echo "Finished processing $NAME."
done
echo "All samples processed successfully."
```

1.3 Peaks (3rd script)

```
#!/bin/bash
# Run MACS for peak calling
echo "Running MACS for peak calling..."
# Loop through the samples
for sample_name in sample22 sample26 sample30 sample24 sample28 sample32 sample38 sample40
do
# Determine the control file based on sample name
case "$sample_name" in
sample22|sample26|sample30)
control="WT.sorted.bam"
;;
sample24|sample28|sample32)
control="Set8KO.sorted.bam"
;;
sample38)
control="SDS1_WT.sorted.bam"
;;
sample40)
control="SDS1_Set8KO.sorted.bam"
;;
*)
echo "No control file defined for $sample_name!"
continue
;;
esac
# Run MACS with the determined control
echo "Running MACS for $sample_name with control $control..."
macs -t "${sample_name}.sorted.bam" -c "$control" --format BAM --name "$sample_name" --gsize 138000000 --tsize 26 --diag --wig
done
```

1.4 Final steps (4th script)

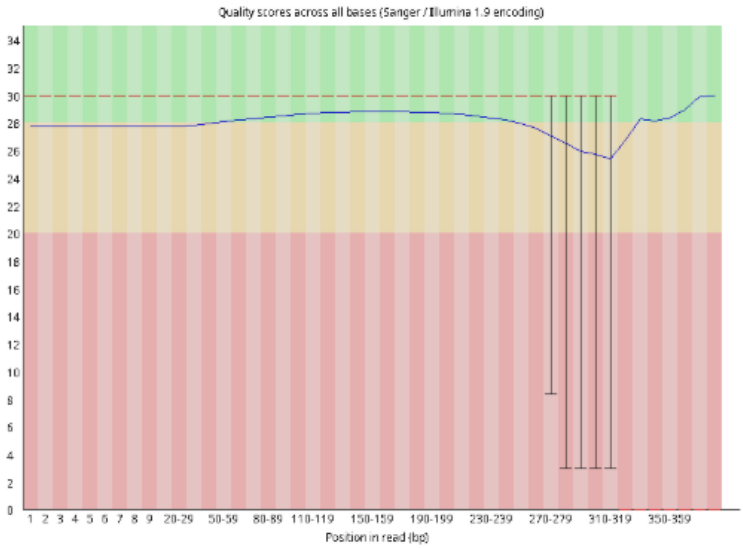
```
#!/bin/bash
# Bowtie2 index and genome files
BOWTIE2_INDEX="Bio_Project/bowtie2_index/mm9"
GENOME="Bio_Project/bowtie2_index/mouse.mm9.genome"
GENOME_FA="Bio_Project/bowtie2_index/mm9.fa"
# List of samples
SAMPLES=("sample22"
"sample24"
"sample26"
"sample28"
"sample30"
"sample32"
"sample38"
"sample40")
# Function to process each sample step by step
process_step() {
local step="$1"
for sample in "${SAMPLES[@]}"; do
sample_name="$sample"
echo "Processing $step for $sample_name..."
case "$step" in
"summits")
# Step 9: Process summits
echo "Processing summits for $sample_name..."
slopBed -i "${sample_name}_summits.bed" -g "$GENOME" -b 20 >
"${sample_name}_summits-b20.bed"
fastaFromBed -fi "$GENOME_FA" -bed "${sample_name}_summits-b20.bed" >
"${sample_name}_summits-b20.fa"
;;
"meme")
# Step 10: Run MEME on summits
echo "Running MEME on summits for $sample_name..."
meme "${sample_name}_summits-b20.fa" -o "${sample_name}_meme" -dna
;;
esac
done
}
# Process all steps for all samples
echo "Processing summits for all samples..."
process_step "summits"
echo "Running MEME for all samples..."
process_step "meme"
echo "All samples processed."
```

2. FastQC reports

Basic Statistics

Measure	Value
Filename	SRR5409158.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	18841588
Total Bases	2.6 Gbp
Sequences flagged as poor quality	0
Sequence length	25-381
%GC	44

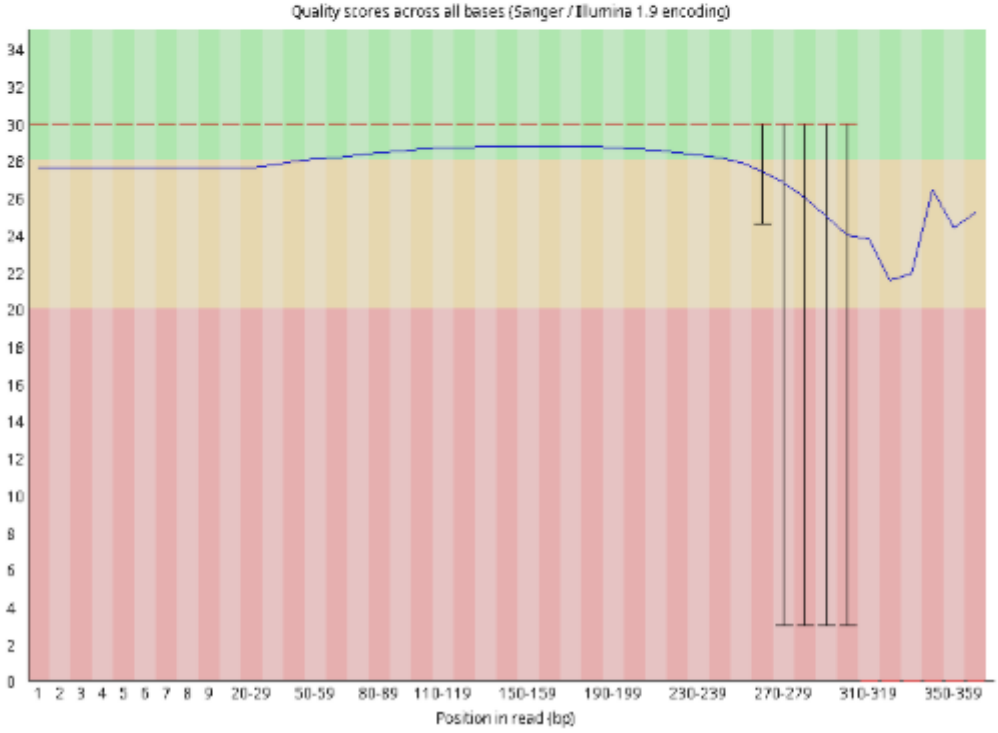
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409160.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	19661157
Total Bases	2.6 Gbp
Sequences flagged as poor quality	0
Sequence length	25-367
%GC	44

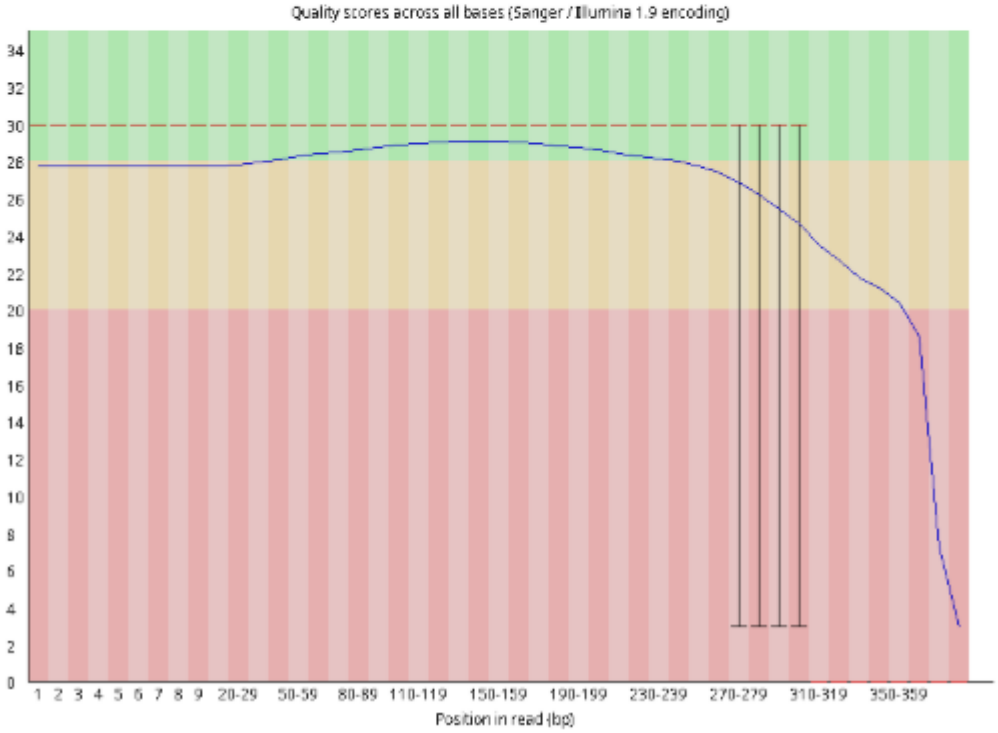
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409162.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	36823752
Total Bases	4.8 Gbp
Sequences flagged as poor quality	0
Sequence length	25-389
%GC	41

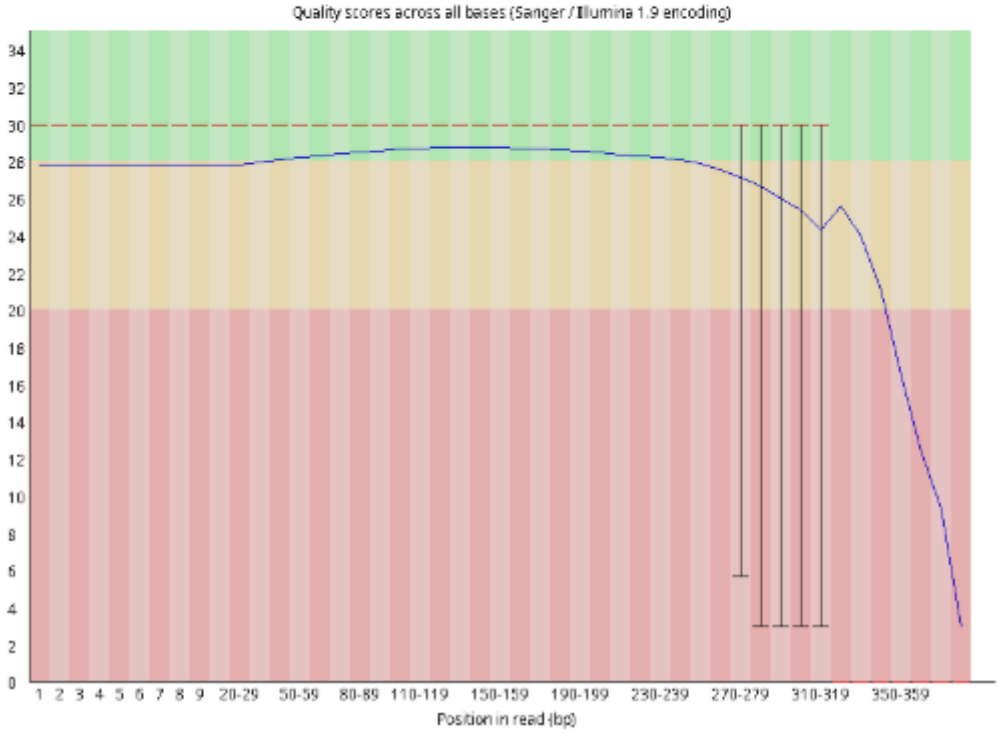
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409164.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	17806682
Total Bases	2.4 Gbp
Sequences flagged as poor quality	0
Sequence length	25-386
%GC	43

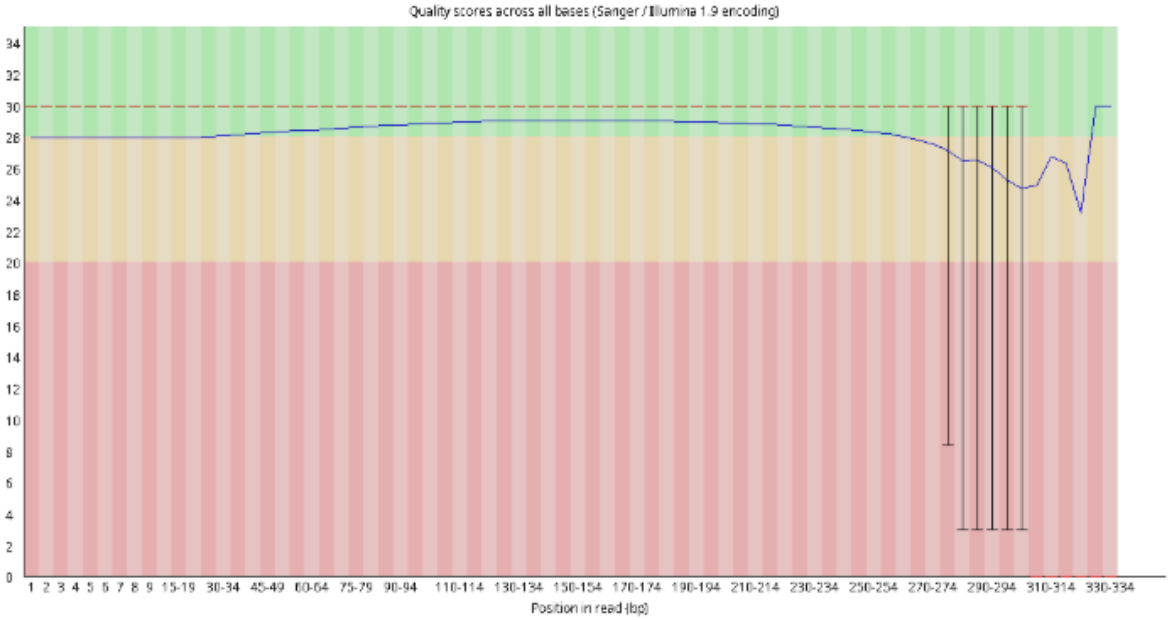
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5489166.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	13286420
Total Bases	1.7 Gbp
Sequences flagged as poor quality	0
Sequence length	25-334
%GC	42

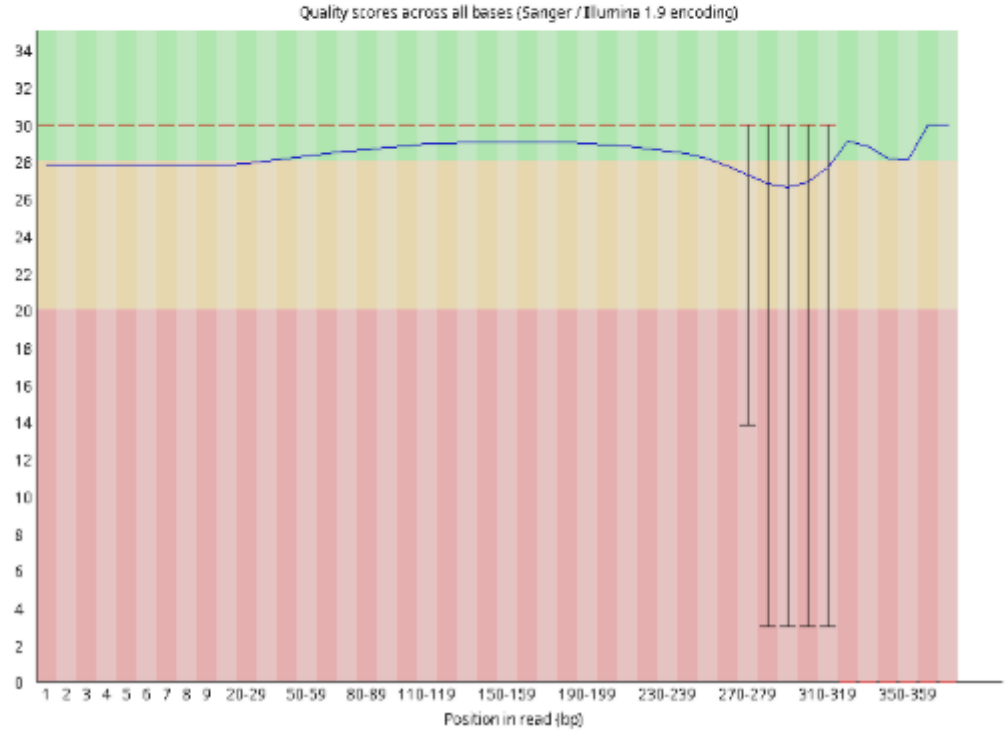
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5489168.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	23616217
Total Bases	3.1 Gbp
Sequences flagged as poor quality	0
Sequence length	25-374
%GC	41

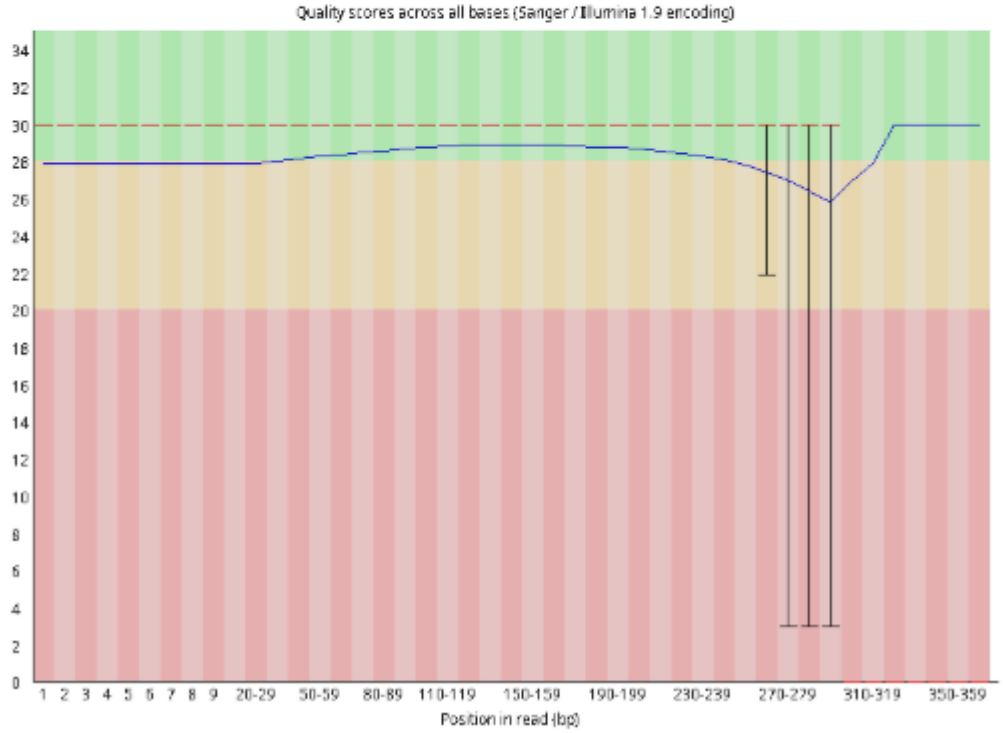
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409170.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	17458147
Total Bases	2.2 Gbp
Sequences flagged as poor quality	0
Sequence length	25-365
%GC	51

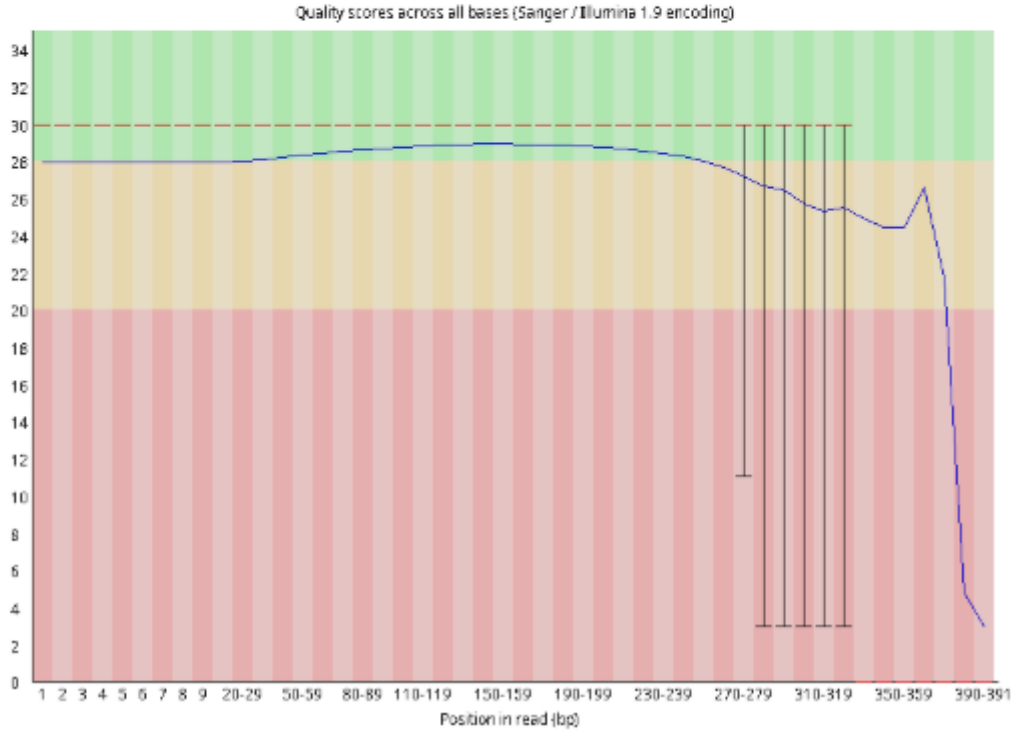
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409172.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	21942513
Total Bases	2.9 Gbp
Sequences flagged as poor quality	0
Sequence length	25-391
%GC	42

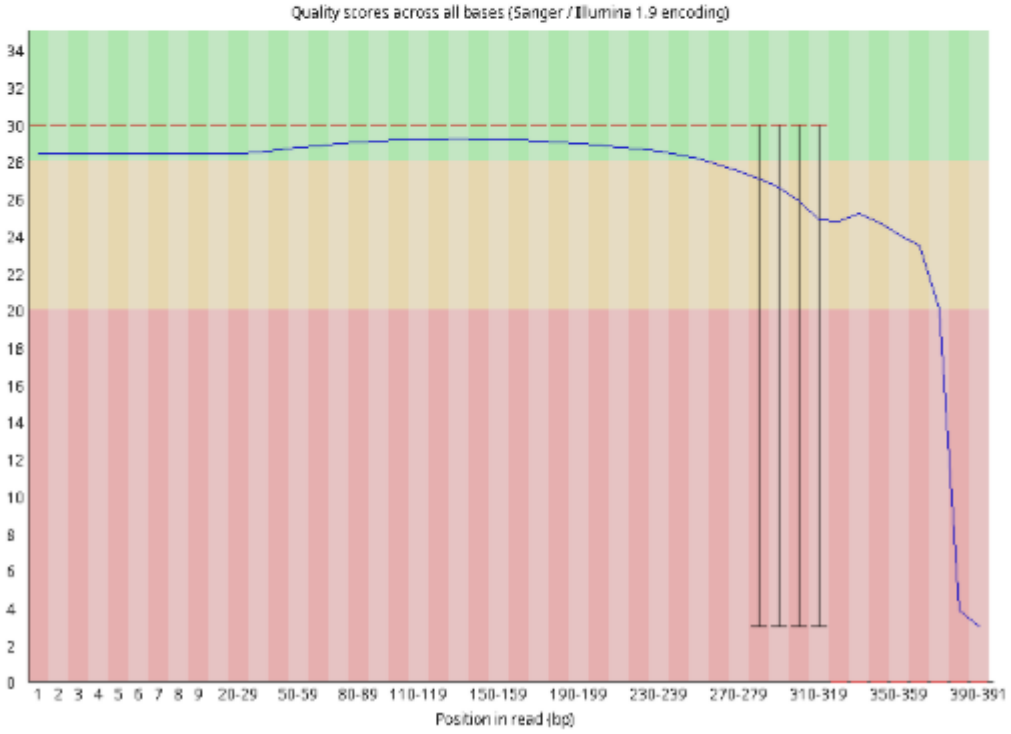
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409174.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	14493540
Total Bases	1.9 Gbp
Sequences flagged as poor quality	0
Sequence length	25-391
%GC	46

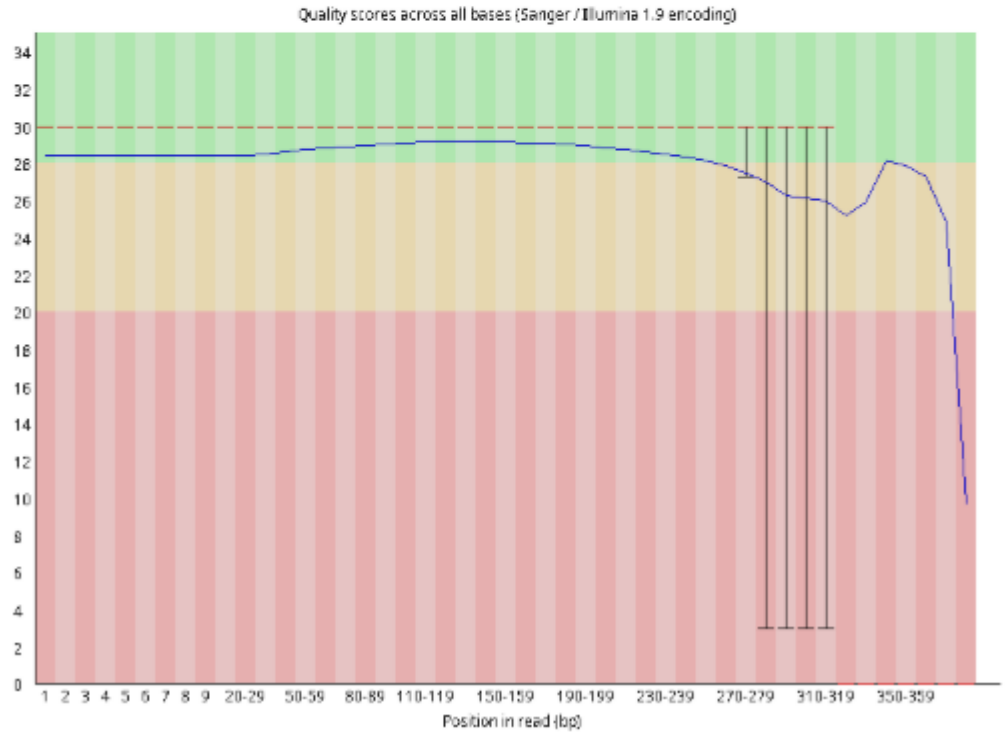
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409176.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	16953247
Total Bases	2.3 Gbp
Sequences flagged as poor quality	0
Sequence length	25-387
%GC	46

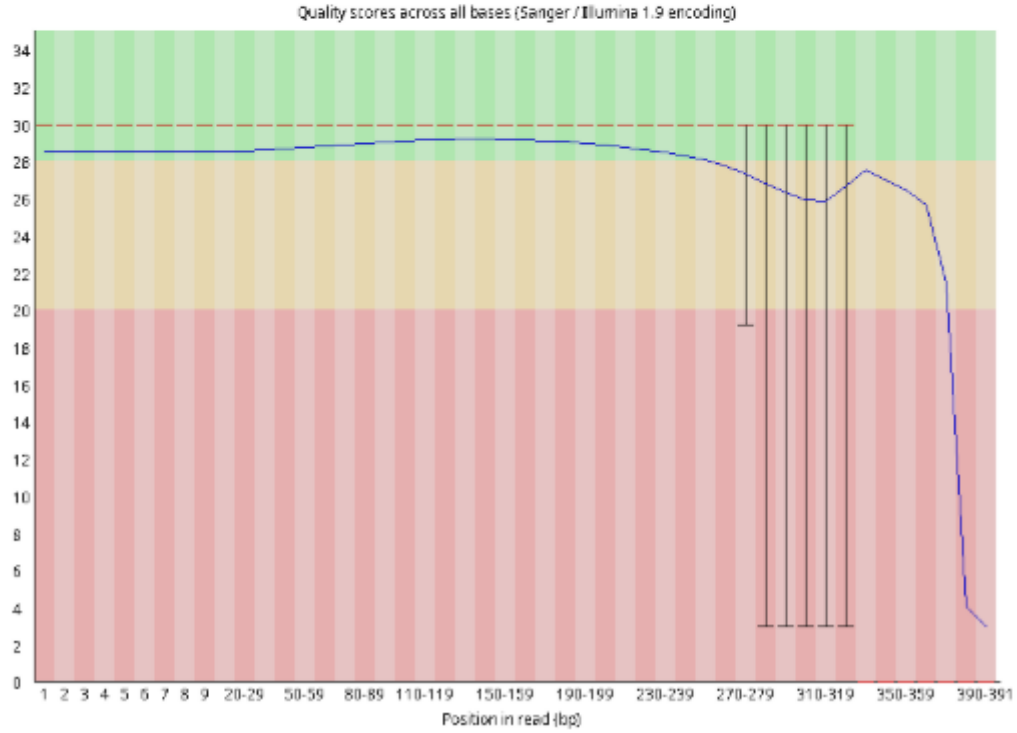
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5489178.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	24498715
Total Bases	3.4 Gbp
Sequences flagged as poor quality	0
Sequence length	25-391
%GC	50

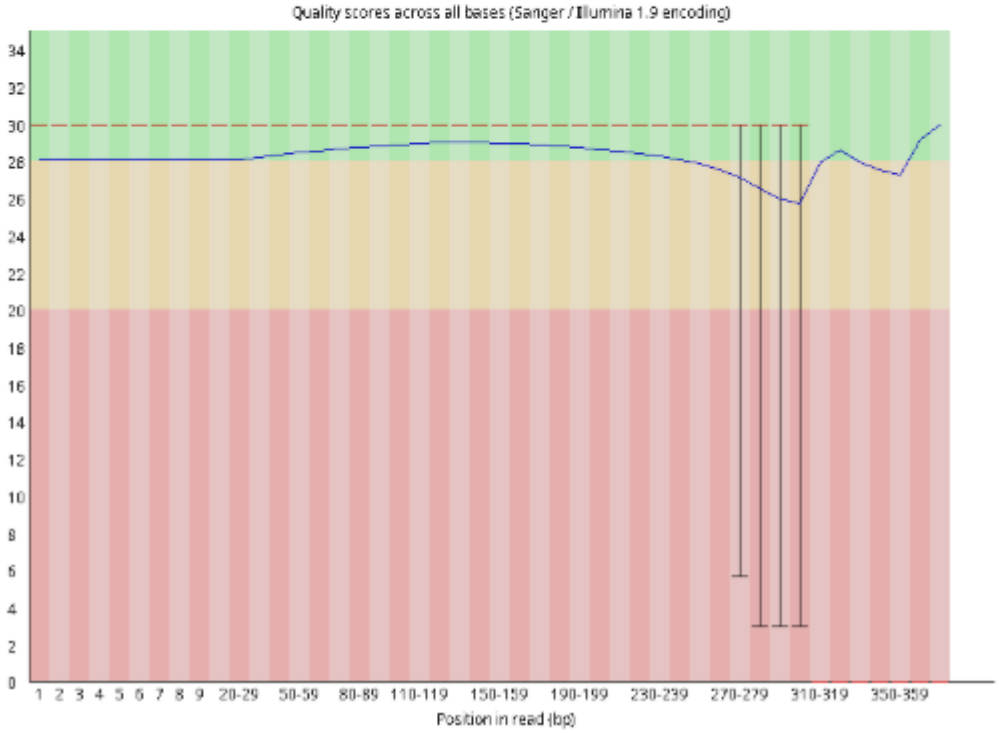
Per base sequence quality



Basic Statistics

Measure	Value
Filename	SRR5409180.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	15081182
Total Bases	2 Gbp
Sequences flagged as poor quality	0
Sequence length	25-374
%GC	48

Per base sequence quality



3. Log files (Screenshots)

Alignments

```
Processing align for sample22...
Aligning SRR5409158.fastq.gz with Bowtie2 for sample22...
18841580 reads; of these:
  18841580 (100.00%) were unpaired; of these:
    1547962 (8.22%) aligned 0 times
    15641711 (83.02%) aligned exactly 1 time
    1651907 (8.77%) aligned >1 times
91.78% overall alignment rate
```

```
Processing align for sample24...
Aligning SRR5409160.fastq.gz with Bowtie2 for sample24...
 19661157 reads; of these:
  19661157 (100.00%) were unpaired; of these:
    1538357 (7.82%) aligned 0 times
    15982084 (81.29%) aligned exactly 1 time
    2140716 (10.89%) aligned >1 times
92.18% overall alignment rate
```

```
Processing align for sample26...
Aligning SRR5409162.fastq.gz with Bowtie2 for sample26...
 36823752 reads; of these:
  36823752 (100.00%) were unpaired; of these:
    24815996 (67.39%) aligned 0 times
    9944208 (27.00%) aligned exactly 1 time
    2063548 (5.60%) aligned >1 times
32.61% overall alignment rate
```

```
Processing align for sample28...
Aligning SRR5409164.fastq.gz with Bowtie2 for sample28...
^[[C^[[C17806682 reads; of these:
  17806682 (100.00%) were unpaired; of these:
    1597825 (8.97%) aligned 0 times
    13277378 (74.56%) aligned exactly 1 time
    2931479 (16.46%) aligned >1 times
91.03% overall alignment rate
```

```
Processing align for sample30...
Aligning SRR5409166.fastq.gz with Bowtie2 for sample30...
13286420 reads; of these:
  13286420 (100.00%) were unpaired; of these:
    1096978 (8.26%) aligned 0 times
    9716963 (73.13%) aligned exactly 1 time
    2472479 (18.61%) aligned >1 times
91.74% overall alignment rate
```

```
Processing align for sample32...
Aligning SRR5409168.fastq.gz with Bowtie2 for sample32...
 23616217 reads; of these:
  23616217 (100.00%) were unpaired; of these:
    1772345 (7.50%) aligned 0 times
    17127666 (72.53%) aligned exactly 1 time
    4716206 (19.97%) aligned >1 times
92.50% overall alignment rate
```

```
Processing align for sample34...
Aligning SRR5409170.fastq.gz with Bowtie2 for sample34...
17458147 reads; of these:
  17458147 (100.00%) were unpaired; of these:
    3942760 (22.58%) aligned 0 times
    11404026 (65.32%) aligned exactly 1 time
    2111361 (12.09%) aligned >1 times
77.42% overall alignment rate
```

```
Aligning SRR5409172.fastq.gz with Bowtie2 for sample36...
21942513 reads; of these:
  21942513 (100.00%) were unpaired; of these:
    1687076 (7.69%) aligned 0 times
    15654081 (71.34%) aligned exactly 1 time
    4601356 (20.97%) aligned >1 times
92.31% overall alignment rate
```

```
Aligning SRR5409174.fastq.gz with Bowtie2 for sample38...
^[[C^[[C^[[C^[[C14493540 reads; of these:
  14493540 (100.00%) were unpaired; of these:
    2597155 (17.92%) aligned 0 times
    9792873 (67.57%) aligned exactly 1 time
    2103512 (14.51%) aligned >1 times
82.08% overall alignment rate
Processing align for sample40...
```

```
Aligning SRR5409178.fastq.gz with Bowtie2 for sample42...
24498715 reads; of these:
  24498715 (100.00%) were unpaired; of these:
    1625333 (6.63%) aligned 0 times
    18866310 (77.01%) aligned exactly 1 time
    4007072 (16.36%) aligned >1 times
93.37% overall alignment rate
```

```
Processing align for sample44...
Aligning SRR5409180.fastq.gz with Bowtie2 for sample44...
15081182 reads; of these:
  15081182 (100.00%) were unpaired; of these:
    4065989 (26.96%) aligned 0 times
    8188445 (54.30%) aligned exactly 1 time
    2826748 (18.74%) aligned >1 times
73.04% overall alignment rate
```

SAM TO BAM CONVERTING

```
Converting SAM to BAM for all samples...
Processing convert_bam for sample22...
Converting SAM to BAM for sample22...
[E::bgzf_flush] File write failed (wrong size)
samtools view: writing to "sample22.bam" failed: Disk quota exceeded
samtools view: error reading file "sample22.sam": Disk quota exceeded
[E::bgzf_close] File write failed
samtools view: error closing "sample22.bam": -1
Processing convert_bam for sample24...
Converting SAM to BAM for sample24...
Processing convert_bam for sample26...
Converting SAM to BAM for sample26...
Processing convert_bam for sample28...
Converting SAM to BAM for sample28...
Processing convert_bam for sample30...
Converting SAM to BAM for sample30...
Processing convert_bam for sample32...
Converting SAM to BAM for sample32...
Processing convert_bam for sample34...
Converting SAM to BAM for sample34...
```

SORTING – FLAGSTAT - BEDGRAPH

```
Sorting BAM files for all samples...
Processing sort_bam for sample22...
Sorting BAM file for sample22...
[W::hts_set_opt] Cannot change block size for this format
samtools sort: failed to read header from "sample22.bam"
Processing sort_bam for sample24...
Sorting BAM file for sample24...
[bam_sort_core] merging from 8 files and 1 in-memory blocks...
Processing sort_bam for sample26...
Sorting BAM file for sample26...
[bam_sort_core] merging from 14 files and 1 in-memory blocks...
Processing sort_bam for sample28...
Sorting BAM file for sample28...
[bam_sort_core] merging from 7 files and 1 in-memory blocks...
Processing sort_bam for sample30...
Sorting BAM file for sample30...
[bam_sort_core] merging from 5 files and 1 in-memory blocks...
Processing sort_bam for sample32...
Sorting BAM file for sample32...
```

```
Converting BedGraph to BigWig for all samples...
Processing bigwig for sample38...
Converting BedGraph to BigWig for sample38...
Processing bigwig for sample40...
Converting BedGraph to BigWig for sample40...
Processing bigwig for sample42...
Converting BedGraph to BigWig for sample42...
Processing bigwig for sample42...
Converting BedGraph to BigWig for sample42...
```

```
Converting SAM to BAM for all samples...
Processing convert_bam for sample22...
Converting SAM to BAM for sample22...
Processing convert_bam for sample36...
Converting SAM to BAM for sample36...
Processing convert_bam for sample42...
Converting SAM to BAM for sample42...
Processing convert_bam for sample44...
Converting SAM to BAM for sample44...
Sorting BAM files for all samples...
Processing sort_bam for sample22...
Sorting BAM file for sample22...
[bam_sort_core] merging from 8 files and 1 in-memory blocks...
Processing sort_bam for sample36...
Sorting BAM file for sample36...
[bam_sort_core] merging from 9 files and 1 in-memory blocks...
Processing sort_bam for sample42...
Sorting BAM file for sample42...
[bam_sort_core] merging from 11 files and 1 in-memory blocks...
Processing sort_bam for sample44...
Sorting BAM file for sample44...
[bam_sort_core] merging from 6 files and 1 in-memory blocks...
Indexing sorted BAM files for all samples...
Processing index_bam for sample22...
Indexing sorted BAM file for sample22...
Processing index_bam for sample36...
Indexing sorted BAM file for sample36...
Processing index_bam for sample42...
Indexing sorted BAM file for sample42...
Processing index_bam for sample44...
Indexing sorted BAM file for sample44...
Generating flagstat for all samples...
Processing flagstat for sample22...
Generating flagstat for sample22...
```

```
Processing sort_bam for sample22...
Sorting BAM file for sample22...
[bam_sort_core] merging from 8 files and 1 in-memory blocks...
Processing sort_bam for sample36...
Sorting BAM file for sample36...
[bam_sort_core] merging from 9 files and 1 in-memory blocks...
Processing sort_bam for sample42...
Sorting BAM file for sample42...
[bam_sort_core] merging from 11 files and 1 in-memory blocks...
Processing sort_bam for sample44...
Sorting BAM file for sample44...
[bam_sort_core] merging from 6 files and 1 in-memory blocks...
Indexing sorted BAM files for all samples...
Processing index_bam for sample22...
Indexing sorted BAM file for sample22...
Processing index_bam for sample36...
Indexing sorted BAM file for sample36...
Processing index_bam for sample42...
Indexing sorted BAM file for sample42...
Processing index_bam for sample44...
Indexing sorted BAM file for sample44...
Generating flagstat for all samples...
Processing flagstat for sample22...
Generating flagstat for sample22...
Processing flagstat for sample36...
Generating flagstat for sample36...
Processing flagstat for sample42...
Generating flagstat for sample42...
Processing flagstat for sample44...
Generating flagstat for sample44...
Generating BedGraph for all samples...
Processing bedgraph for sample22...
Generating BedGraph for sample22...
```

CONVERTING BEDGRAPH TO BIGWIG FILES

```
*****
Converting BedGraph to BigWig for all samples...
Processing bigwig for sample22...
Converting BedGraph to BigWig for sample22...
Processing bigwig for sample36...
Converting BedGraph to BigWig for sample36...
Processing bigwig for sample42...
Converting BedGraph to BigWig for sample42...
Processing bigwig for sample44...
Converting BedGraph to BigWig for sample44...
```

PEAKS

```
Running MACS for peak calling...
Running MACS for sample22 with control WT.sorted.bam...
INFO  @ Sun, 02 Feb 2025 22:30:06:
# ARGUMENTS LIST:
# name = sample22
# format = BAM
# ChIP-seq file = sample22.sorted.bam
# control file = WT.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

MACS

```
Running MACS for sample26 with control WT.sorted.bam...
INFO  @ Sun, 02 Feb 2025 23:01:08:
# ARGUMENTS LIST:
# name = sample26
# format = BAM
# ChIP-seq file = sample26.sorted.bam
# control file = WT.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

```
Running MACS for sample30 with control WT.sorted.bam...
INFO  @ Sun, 02 Feb 2025 23:28:56:
# ARGUMENTS LIST:
# name = sample30
# format = BAM
# ChIP-seq file = sample30.sorted.bam
# control file = WT.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

```
Running MACS for sample24 with control Set8KO.sorted.bam...
INFO  @ Mon, 03 Feb 2025 00:00:35:
# ARGUMENTS LIST:
# name = sample24
# format = BAM
# ChIP-seq file = sample24.sorted.bam
# control file = Set8KO.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

```
Running MACS for sample28 with control Set8KO.sorted.bam...
INFO  @ Mon, 03 Feb 2025 00:38:40:
# ARGUMENTS LIST:
# name = sample28
# format = BAM
# ChIP-seq file = sample28.sorted.bam
# control file = Set8KO.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

```
Running MACS for sample32 with control Set8KO.sorted.bam...
INFO  @ Mon, 03 Feb 2025 01:11:51:
# ARGUMENTS LIST:
# name = sample32
# format = BAM
# ChIP-seq file = sample32.sorted.bam
# control file = Set8KO.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

```
Running MACS for sample38 with control SDS1_WT.sorted.bam...
INFO  @ Mon, 03 Feb 2025 02:00:51:
# ARGUMENTS LIST:
# name = sample38
# format = BAM
# ChIP-seq file = sample38.sorted.bam
# control file = SDS1_WT.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

```
Running MACS for sample40 with control SDS1_Set8KO.sorted.bam...
INFO  @ Mon, 03 Feb 2025 02:32:59:
# ARGUMENTS LIST:
# name = sample40
# format = BAM
# ChIP-seq file = sample40.sorted.bam
# control file = SDS1_Set8KO.sorted.bam
# effective genome size = 1.38e+08
# band width = 300
# model fold = 10,30
# pvalue cutoff = 1.00e-05
# Small dataset will be scaled towards larger dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
```

SUMMITS

```
user18@UoA-Intro2Bio24:/mnt/hdd_b/home/user18/Desktop$ ./finalsteps.sh
Processing summits for all samples...
Processing summits for sample22...
Processing summits for sample22...
Processing summits for sample24...
Processing summits for sample24...
Processing summits for sample26...
Processing summits for sample26...
Processing summits for sample28...
Processing summits for sample28...
Processing summits for sample30...
Processing summits for sample30...
Processing summits for sample32...
Processing summits for sample32...
Processing summits for sample38...
Processing summits for sample38...
Processing summits for sample40...
Processing summits for sample40...
Running MEME for all samples...
Processing meme for sample22...
```

Meme

```
Processing meme for sample22...
Running MEME on summits for sample22...
Writing results to output directory 'sample22_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 26 p0 26 p1 0 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 1066 highwater mark: seq 26 pos 33
Initializing the motif probability tables for 2 to 26 sites...
nsites = 26
Done initializing.

seqs=    26, min_w=  41, max_w=   41, total_size=     1066

motif=1
SEED DEPTHS: 2 4 8 16 26
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=  26, iter=   0
Processing meme for sample24...
Running MEME on summits for sample24...
Writing results to output directory 'sample24_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 20 p0 20 p1 0 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 820 highwater mark: seq 20 pos 33
Initializing the motif probability tables for 2 to 20 sites...
nsites = 20
Done initializing.

seqs=    20, min_w=  41, max_w=   41, total_size=      820

motif=1
SEED DEPTHS: 2 4 8 16 20
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=  20, iter=   0
```

```
Running MEME on summits for sample26...
Writing results to output directory 'sample26_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 1046 p0 1000 p1 46 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 42886 highwater mark: seq 1046 pos 33
Initializing the motif probability tables for 2 to 1000 sites...
nsites = 1000
Done initializing.

seqs=  1046, min_w=  41, max_w=   41, total_size=    42886

motif=1
SEED DEPTHS: 2 4 8 16 32 64 128 256 512 1000
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=1000, iter=   0
Processing meme for sample28...
Running MEME on summits for sample28...
Writing results to output directory 'sample28_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 657 p0 657 p1 0 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 26937 highwater mark: seq 657 pos 33
Initializing the motif probability tables for 2 to 657 sites...
nsites = 657
Done initializing.

seqs=   657, min_w=  41, max_w=   41, total_size=    26937

motif=1
SEED DEPTHS: 2 4 8 16 32 64 128 256 512 657
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites= 657, iter=   0
```

```
Processing meme for sample30...
Running MEME on summits for sample30...
Writing results to output directory 'sample30_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 1085 p0 1000 p1 85 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 44485 highwater mark: seq 1085 pos 33
Initializing the motif probability tables for 2 to 1000 sites...
nsites = 1000
Done initializing.

seqs=  1085, min_w=  41, max_w=   41, total_size=    44485

motif=1
SEED DEPTHS: 2 4 8 16 32 64 128 256 512 1000
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=1000, iter=   0
Processing meme for sample32...
Running MEME on summits for sample32...
Writing results to output directory 'sample32_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 36 p0 36 p1 0 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 1476 highwater mark: seq 36 pos 33
Initializing the motif probability tables for 2 to 36 sites...
nsites = 36
Done initializing.

seqs=    36, min_w=  41, max_w=   41, total_size=     1476

motif=1
SEED DEPTHS: 2 4 8 16 32 36
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=  36, iter=   0
```

```
Processing meme for sample38...
Running MEME on summits for sample38...
Writing results to output directory 'sample38_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 6293 p0 1000 p1 5293 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 60316 highwater mark: seq 1774 pos 26
Initializing the motif probability tables for 2 to 1000 sites...
nsites = 1000
Done initializing.

seqs=  6293, min_w=  41, max_w=   41, total_size=   258013

motif=1
SEED DEPTHS: 2 4 8 16 32 64 128 256 512 1000
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=1000, iter=   0
Processing meme for sample40...
Running MEME on summits for sample40...
Writing results to output directory 'sample40_meme'.
BACKGROUND: using background model of order 0
PRIMARY (classic): n 3357 p0 1000 p1 2357 p2 0
SEQUENCE GROUP USAGE-- Starts/EM: p0; Trim: p0; pvalue: p0; nsites: p0,p1,p2
Option '-maxw' is greather than the length of longest sequence (41).  Setting '-maxw' to 41.
SEEDS: maxwords 60316 highwater mark: seq 1774 pos 26
Initializing the motif probability tables for 2 to 1000 sites...
nsites = 1000
Done initializing.

seqs=  3357, min_w=  41, max_w=   41, total_size=   137637

motif=1
SEED DEPTHS: 2 4 8 16 32 64 128 256 512 1000
SEED WIDTHS: 8 11 15 21 29 41
em: w=  41, psites=1000, iter=   0
All samples processed.
```