Complete ChIP-Seq Analysis Pipeline Report

1. Project Overview

This project involves processing ChIP-Seq data for the transcription factor USF2 from HepG2 cells. The samples include a ChIP sample (USF2 immunoprecipitated DNA) and an Input control. The goal is to perform a complete pipeline: FASTQ processing, alignment, filtering, peak calling, annotation, motif analysis, GREAT enrichment, and visualization.

Biological Background of USF2

USF2 (Upstream Stimulatory Factor 2) is a member of the basic helix-loop-helix leucine zipper family of transcription factors. It binds DNA at canonical E-box motifs (CACGTG) and regulates genes involved in metabolism, stress response, and cell cycle control.

In liver-derived HepG2 cells, USF2 is known to participate in:
- regulation of lipid metabolism genes
- glucose homeostasis
- oxidative stress pathways
- cell proliferation and tumor progression

ChIP-Seq analysis enables identification of its genome-wide binding landscape, revealing both promoter-proximal and enhancer-associated roles.

GEO / SRA Details

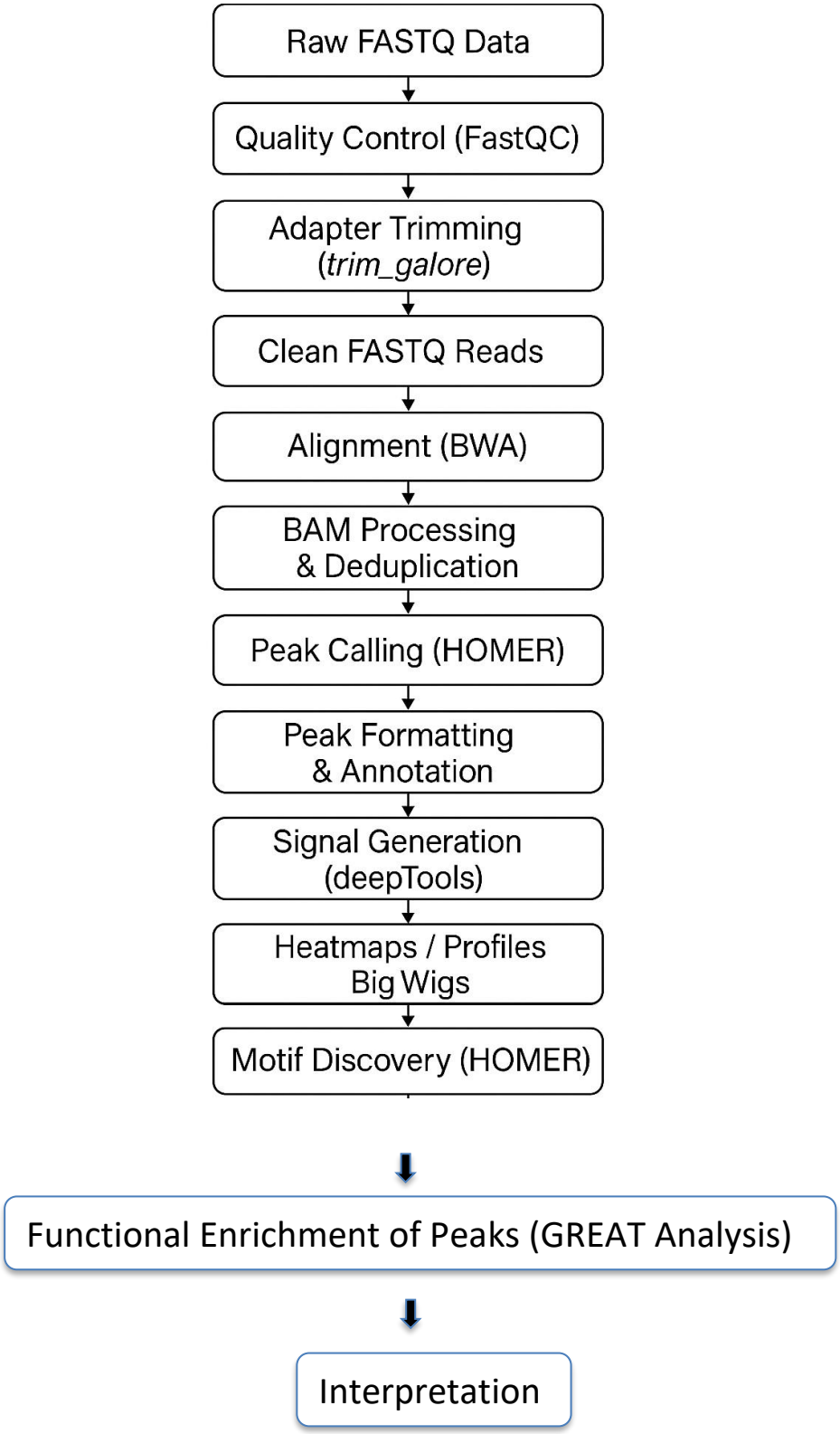
• GEO Series: **GSE104247**
• Organism: Homo sapiens
• **Cell Line: HepG2 (liver cancer)**
• Assay: ChIP-Seq (Genome-wide binding profiling)

Samples used:
- **SRR6117703 → USF2 ChIP sample**
- **SRR6117732 → Input control**

These samples were sequenced on the **Illumina HiSeq 2000 platform**. The Input sample serves as a background reference to distinguish genuine DNA–protein interactions from chromatin accessibility and sequencing noise.

Workflow Diagram

Below is a workflow diagram outlining the complete ChIP-Seq computational pipeline.



2. Setting Up the conda Environment (Ubuntu 24.04.3 LTS, WSL2)

```
conda create -n homer_env python=3.10 -y
conda activate homer_env
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

A dedicated conda environment ensures reproducibility. Adding required channels allows installation of all bioinformatics tools needed throughout the pipeline.

3. Software Installation

```
conda install -y \
  wget \
  samtools \
  sra-tools \
  trim-galore \
  bedtools \
  picard \
  bwa \
  fastqc \
  deeptools
```

These tools form the foundation of a standard ChIP-Seq workflow: SRA download, trimming, alignment, QC, duplicate removal, peak calling, and visualization.

4. HOMER Installation

```
mkdir -p ~/homer
cd ~/homer
wget http://homer.ucsd.edu/homer/configureHomer.pl
perl configureHomer.pl -install

echo 'export PATH="$HOME/homer/bin:$PATH"' >> ~/.bashrc
source ~/.bashrc
perl configureHomer.pl -install hg38
```

HOMER is used for peak calling, motif discovery, and annotation. Installing hg38 ensures the genome reference is available.

5. Reference Genome Preparation

mkdir -p ~/hg38_bwa
cd ~/hg38_bwa
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
gunzip hg38.fa.gz
bwa index hg38.fa

The hg38 genome is downloaded from UCSC and indexed with BWA for alignment.

6. Downloading FASTQ Files

prefetch SRR6117732
cd SRR6117732
fasterq-dump --threads 4 SRR6117732.sra
gzip SRR6117732.fastq
mv SRR6117732.fastq.gz ~/GSE104247/raw

```
spots read      : 41,904,860
reads read      : 83,809,720
reads written   : 41,904,860
reads 0-length  : 41,904,860
```

prefetch SRR6117703
cd SRR6117703
fasterq-dump --threads 4 SRR6117703.sra
gzip SRR6117703.fastq
mv SRR6117703.fastq.gz ~/GSE104247/raw

```
spots read      : 21,541,106
reads read      : 43,082,212
reads written   : 21,541,106
reads 0-length  : 21,541,106
```

The Input and USF2 ChIP samples are downloaded from SRA and converted to FASTQ for downstream analysis.

7. Trimming & Quality Control

trim_galore --fastqc --cores 4 --quality 20 --length 20 -o ~/GSE104247/trim ~/GSE104247/raw/USF2.fastq.gz
trim_galore --fastqc --cores 4 --quality 20 --length 20 -o ~/GSE104247/trim ~/GSE104247/raw/INPUT.fastq.gz

Trim Galore removes low-quality bases and adapters. FastQC reports allow visualizing read quality.

8. Alignment with BWA

bwa mem -t 4 -M -R '@RG\tID:USF2\tSM:USF2\tPL:ILLUMINA' ~/hg38_bwa/hg38.fa ~/GSE104247/trim/USF2_trimmed.fq.gz > ~/GSE104247/bam/USF2.sam

bwa mem -t 4 -M -R '@RG\tID:INPUT\tSM:INPUT\tPL:ILLUMINA' ~/hg38_bwa/hg38.fa ~/GSE104247/trim/INPUT_trimmed.fq.gz > ~/GSE104247/bam/INPUT.sam

BWA-MEM aligns trimmed reads to hg38. Read groups help track sample identity.

9. BAM Processing

(Converting, Sorting, Deduplication, Indexing)

Convert SAM → BAM

samtools view -@ 4 -b USF2.sam > USF2.bam

samtools view -@ 4 -b INPUT.sam > INPUT.bam

Raw Alignment QC (Before Sorting)

samtools flagstat USF2.bam > USF2_flagstat_raw.txt

samtools flagstat INPUT.bam > INPUT_flagstat_raw.txt

```
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$ cat USF2_flagstat_raw.txt
21316852 + 0 in total (QC-passed reads + QC-failed reads)
21316769 + 0 primary
83 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
20836930 + 0 mapped (97.75% : N/A)
20836847 + 0 primary mapped (97.75% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$ cat INPUT_flagstat_raw.txt
40637950 + 0 in total (QC-passed reads + QC-failed reads)
40635594 + 0 primary
2356 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
40099064 + 0 mapped (98.67% : N/A)
40096708 + 0 primary mapped (98.67% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Sort BAM Files

Memory-safe for WSL (4 threads, 1 GB per thread)

samtools sort -@ 4 -m 1G -o USF2_sorted.bam USF2.bam

samtools sort -@ 4 -m 1G -o INPUT_sorted.bam INPUT.bam

```
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$ ls -lh
total 4.9G
-rw-r--r-- 1 shobita shobita 1.8G Nov 20 14:10 INPUT.bam
-rw-r--r-- 1 shobita shobita  480 Nov 20 14:12 INPUT_flagstat_raw.txt
-rw-r--r-- 1 shobita shobita 1.5G Nov 20 14:45 INPUT_sorted.bam
-rw-r--r-- 1 shobita shobita 905M Nov 20 14:07 USF2.bam
-rw-r--r-- 1 shobita shobita  478 Nov 20 14:12 USF2_flagstat_raw.txt
-rw-r--r-- 1 shobita shobita 744M Nov 20 14:43 USF2_sorted.bam
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$
```

Remove PCR Duplicates (Picard MarkDuplicates)

```
picard MarkDuplicates \
   I=USF2_sorted.bam \
   O=USF2_dedup.bam \
   M=USF2_metrics.txt \
   REMOVE_DUPLICATES=true
picard MarkDuplicates \
   I=INPUT_sorted.bam \
   O=INPUT_dedup.bam \
   M=INPUT_metrics.txt \
   REMOVE_DUPLICATES=true
```

Index the Deduplicated BAM files

```
samtools index USF2_dedup.bam
samtools index INPUT_dedup.bam
```

These steps convert, sort, remove PCR duplicates, and index BAM files for further analysis.

Final Alignment QC (After Deduplication)

```
samtools flagstat USF2_dedup.bam > USF2_flagstat_final.txt
```

samtools flagstat INPUT_dedup.bam > INPUT_flagstat_final.txt

```
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$ cat INPUT_flagstat_final.txt
38601434 + 0 in total (QC-passed reads + QC-failed reads)
38599078 + 0 primary
2356 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
38062548 + 0 mapped (98.60% : N/A)
38060192 + 0 primary mapped (98.60% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$ cat  USF2_flagstat_final.txt
19581281 + 0 in total (QC-passed reads + QC-failed reads)
19581198 + 0 primary
83 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
19101359 + 0 mapped (97.55% : N/A)
19101276 + 0 primary mapped (97.55% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$
```

```
(homer_env) shobita@LAPTOP-PCUDG4QI:~/GSE104247/bam$ ls -lh
total 4.5G
-rw-r--r-- 1 shobita shobita 1.6G Nov 20 15:03 INPUT_dedup.bam
-rw-r--r-- 1 shobita shobita 2.2M Nov 20 15:07 INPUT_dedup.bam.bai
-rw-r--r-- 1 shobita shobita  480 Nov 20 15:11 INPUT_flagstat_final.txt
-rw-r--r-- 1 shobita shobita  480 Nov 20 14:12 INPUT_flagstat_raw.txt
-rw-r--r-- 1 shobita shobita 1.6K Nov 20 15:03 INPUT_metrics.txt
-rw-r--r-- 1 shobita shobita 1.5G Nov 20 14:45 INPUT_sorted.bam
-rw-r--r-- 1 shobita shobita 745M Nov 20 14:51 USF2_dedup.bam
-rw-r--r-- 1 shobita shobita 2.6M Nov 20 15:06 USF2_dedup.bam.bai
-rw-r--r-- 1 shobita shobita  478 Nov 20 15:10 USF2_flagstat_final.txt
-rw-r--r-- 1 shobita shobita  478 Nov 20 14:12 USF2_flagstat_raw.txt
-rw-r--r-- 1 shobita shobita 1.6K Nov 20 14:51 USF2_metrics.txt
-rw-r--r-- 1 shobita shobita 744M Nov 20 14:43 USF2_sorted.bam
```

Clean Up Intermediate Large Files (Optional, but recommended)

rm USF2.sam INPUT.sam

rm USF2.bam INPUT.bam

rm USF2_sorted.bam INPUT_sorted.bam

10. Peak Calling with HOMER

(Filtering → Tag Directories → Peak Calling)

Working directory:

cd ~/GSE104247/bam

10.1 Download hg38 ENCODE Blacklist (one-time)

These regions produce artificial ChIP-seq signal and must be removed before peak calling.

wget https://raw.githubusercontent.com/Boyle-Lab/Blacklist/master/lists/hg38-blacklist.v2.bed.gz

gunzip hg38-blacklist.v2.bed.gz

10.2 Remove Blacklisted Regions (bedtools intersect)

-v keeps only the reads that **do NOT** overlap the blacklist

bedtools intersect \

-v \

-abam USF2_dedup.bam \

-b hg38-blacklist.v2.bed \

> USF2_clean.bam

bedtools intersect \

-v \

-abam INPUT_dedup.bam \

-b hg38-blacklist.v2.bed \

> INPUT_clean.bam

10.3 Index the Filtered BAM Files

Required for HOMER, IGV, and region-based tools.

samtools index USF2_clean.bam

samtools index INPUT_clean.bam

10.4 Create HOMER Tag Directories

#Tag directories store optimized read/tag positions. HOMER uses them internally for peak calling.

makeTagDirectory USF2_tags/ USF2_clean.bam

makeTagDirectory INPUT_tags/ INPUT_clean.bam

10.5 Peak Calling Using HOMER (TF-style: narrow peaks)

USF2 is a transcription factor → use -style factor for sharp peaks.

findPeaks USF2_tags/ \

 -style factor \

 -i INPUT_tags/ \

 -o USF2_peaks.txt

Output: USF2_peaks.txt (HOMER TF-peak format)

11. Peak Annotation

(BED Conversion → Peak Annotation)

Now that peaks are called, this section converts them to BED and annotates them.

Convert HOMER Peaks to BED Format

Required for IGV, GREAT, deepTools, bedtools, etc.

pos2bed.pl USF2_peaks.txt > USF2_peaks.bed

Output: USF2_peaks.bed

Peak Annotation Using HOMER (via hg38)

Annotates each peak with the nearest gene, distance to TSS, genomic region (promoter, exon, intron, intergenic).

annotatePeaks.pl USF2_peaks.bed hg38 > USF2_peak_annotation.txt

Output:
USF2_peak_annotation.txt

12 . Visualization Using deepTools

Purpose:
This stage generates genome-wide ChIP-Seq signal tracks, performs background correction, computes peak-centered matrices, and visualizes signal profiles and heatmaps. These files are used for IGV browsing .

12.1 Overview of What Visualization Produces

The following operations will be performed:

- Create **bigWig signal tracks** for USF2 and Input
- Create **$\log_2$(ChIP/Input)** normalized coverage
- Compute **peak-centered matrices** for aggregation plots
- Generate **heatmaps** of enrichment across peaks
- Generate **average signal profiles**
- Prepare visualization files for **IGV genome browser**

All output files will be stored in:

~/GSE104247/visualization/

12.2 Visualization Script (deepTools)

Directory structure note:
The script assumes your BAM & peak files are located at: ~/GSE104247/bam/

Create a directory for scripts:

mkdir -p ~/GSE104247/scripts

Save the following script EXACTLY as written below as:

~/GSE104247/scripts/**chipseq_visualization_full.sh**

Full Script (To automate the Visualisation Part):

```{bash}

#!/bin/bash

set -euo pipefail

# ------------------------------------------------------------------

# Minimal ChIP-seq visualization pipeline (deepTools + IGV-ready)

# Saves outputs to: ~/GSE104247/visualization/

# Designed for your folder layout:

#   ~/GSE104247/bam/USF2_clean.bam

#   ~/GSE104247/bam/INPUT_clean.bam

#   ~/GSE104247/bam/USF2_peaks.bed
```

```
# Requires: deepTools (bamCoverage, bamCompare, computeMatrix, plotHeatmap, plotProfile)
# ------------------------------------------------------------------


WORKDIR="$HOME/GSE104247"
BAMDIR="${WORKDIR}/bam"
OUTDIR="${WORKDIR}/visualization"


# create output directory
mkdir -p "${OUTDIR}"


# move into bam directory for inputs (keeps things predictable)
cd "${BAMDIR}"


echo "Using BAM directory: ${BAMDIR}"
echo "Visualization outputs will be written to: ${OUTDIR}"
echo


# ----------------------------
# Step 1: make BigWig signal tracks
# ----------------------------
# Reason: bigWig files are compact and fast for visualization (IGV & deepTools)
# We create one for ChIP and one for Input. Files are written to OUTDIR.
# ----------------------------


echo "STEP 1: Generating BigWig files (USF2 and INPUT)"
```

```
bamCoverage \
  --bam USF2_clean.bam \
  --outFileName "${OUTDIR}/USF2.bw" \
  --binSize 10 \
  --normalizeUsing RPKM \
  --numberOfProcessors 4

bamCoverage \
  --bam INPUT_clean.bam \
  --outFileName "${OUTDIR}/INPUT.bw" \
  --binSize 10 \
  --normalizeUsing RPKM \
  --numberOfProcessors 4

echo "BigWig files created: ${OUTDIR}/USF2.bw  ${OUTDIR}/INPUT.bw"
echo

# ----------------------------
# Step 2: create log2(ChIP/Input) normalized track
# ----------------------------
# Reason: best single-track representation of true enrichment (background corrected)
# ----------------------------

echo "STEP 2: Creating log2(USF2 / INPUT) BigWig (background-corrected)"
```

```
bamCompare \
 -b1 USF2_clean.bam \
 -b2 INPUT_clean.bam \
 --operation log2 \
 --outFileName "${OUTDIR}/USF2_vs_INPUT_log2.bw" \
 --binSize 10 \
 --numberOfProcessors 4

echo "Log2 track created: ${OUTDIR}/USF2_vs_INPUT_log2.bw"
echo

# ----------------------------
# Step 3: compute matrix around peak centers
# ----------------------------
# Reason: computeMatrix collects numerical signal around each peak center,
# producing a matrix used for heatmap and profile plots.
# -R uses your peak BED
# -S uses the log2 normalized signal track (preferred)
# ----------------------------

PEAKS="${BAMDIR}/USF2_peaks.bed"
MATRIX="${OUTDIR}/matrix_USF2.gz"

echo "STEP 3: Computing matrix around peak centers (±2kb w/ 10bp bins)"
computeMatrix reference-point \
 --referencePoint center \
```

```
 -b 2000 -a 2000 \
 -R "${PEAKS}" \
 -S "${OUTDIR}/USF2_vs_INPUT_log2.bw}" \
 --skipZeros \
 --missingDataAsZero \
 --binSize 10 \
 -o "${MATRIX}" \
 --numberOfProcessors 4

echo "Matrix written to: ${MATRIX}"
echo

# ---------------------------
# Step 4: plot heatmap
# ---------------------------

HEATMAP_PNG="${OUTDIR}/USF2_heatmap.png"
echo "STEP 4: Plotting heatmap -> ${HEATMAP_PNG}"

plotHeatmap \
 -m "${MATRIX}" \
 -out "${HEATMAP_PNG}" \
 --colorMap Reds \
 --regionsLabel "USF2 Peaks" \
 --heatmapHeight 10 \
 --heatmapWidth 4
```

```
echo "Heatmap generated."
echo

# ---------------------------
# Step 5: plot average profile
# ---------------------------

PROFILE_PNG="${OUTDIR}/USF2_profile.png"
echo "STEP 5: Plotting average profile -> ${PROFILE_PNG}"

plotProfile \
 -m "${MATRIX}" \
 -out "${PROFILE_PNG}" \
 --regionsLabel "USF2 Peaks"

echo "Profile plot generated."
echo

# ---------------------------
# Final message
# ---------------------------

echo "Visualization complete. Files in ${OUTDIR}:"
ls -lh "${OUTDIR}" | sed -n '1,200p'
```

```
echo "Minimal IGV upload suggestion (only these files):"
echo "  ${OUTDIR}/USF2_vs_INPUT_log2.bw"
echo "  ${BAMDIR}/USF2_peaks.bed"
echo "Optional files”
echo "  ${BAMDIR}/USF2_clean.bam"
echo "  ${BAMDIR}/USF2_clean.bam.bai"
echo "  ${BAMDIR}/INPUT_clean.bam"
echo "  ${BAMDIR}/INPUT_clean.bam.bai"
# The individual .bw files can also be visualised
```

12.3 Running the Visualization Script

Make executable:

```
chmod +x chipseq_visualization_full.sh
```

Run from anywhere:

```
bash chipseq_visualization_full.sh
```

12.4 Explanations

A. What this script does

- Reads **USF2_clean.bam**, **INPUT_clean.bam**, **USF2_peaks.bed**
- Generates **USF2.bw** & **INPUT.bw**
- Creates **$\log_2$(ChIP/Input)** corrected track
- Computes matrix around peak centers
- Generates **heatmap** & **average profile**
- Stores everything under ~/GSE104247/visualization/

B. Why each step is needed

- **bamCoverage → bigWig** = fast, lightweight tracks
- **Normalization (RPKM)** = compare samples fairly
- **bamCompare ($\log_2$)** = true enrichment vs background
- **computeMatrix** = numerical table for heatmaps/profiles
- **plotHeatmap / plotProfile** = visualization of binding patterns

C. Minimal IGV upload list

Upload only:

- USF2_vs_INPUT_log2.bw
- USF2.bw
- INPUT.bw
- USF2_peaks.bed

16. Motif Discovery using HOMER

Purpose:
Identify enriched DNA sequence motifs found at USF2 binding sites.
This step validates the biological quality of your ChIP-seq experiment by checking whether known USF2 motifs (e.g., **CACGTG**) are enriched.

Working directory:

cd ~/GSE104247/bam

Command:

```
findMotifsGenome.pl \
  USF2_peaks.bed \
  hg38 \
  ../motifs_USF2/ \
  -size 200
```

Parameter Explanation

Parameter	Meaning
USF2_peaks.bed	Your final peak list from HOMER
hg38	Reference genome
../motifs_USF2/	Output directory
-size 200	Extract ±100 bp around the summit (200 bp total)

✓ Why -size 200 Is Biologically Appropriate for Transcription Factor ChIP-Seq

1.TF motifs are small (6–20 bp)

Transcription factors bind short DNA sequences.
For USF2 specifically, the canonical E-box motif is:

CACGTG

= **6 bp**

These short motifs occur very close to the **peak summit**, not across the entire peak region.

2.The peak summit represents the strongest binding point

The summit is the location with **highest ChIP-seq signal**, meaning:

- most reads pile up here
- highest likelihood of direct TF–DNA contact
- strongest sequence conservation

Using a small summit-centered region improves sensitivity and motif discovery accuracy.

Homer *de novo* Motif Results (../motifs_USF2//)

Non-redundant Motif File of Results
Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 20374
Total background sequences = 76143
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-16620	-3.827e+04	70.97%	3.07%	25.7bp (67.4bp)	USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer(0.965) More Information | Similar Motifs Found
2		1e-1193	-2.748e+03	12.43%	1.89%	39.4bp (64.2bp)	NFY(CCAAT)/Promoter/Homer(0.963) More Information | Similar Motifs Found
3		1e-469	-1.082e+03	2.64%	0.14%	39.5bp (59.7bp)	GFY(?)/Promoter/Homer(0.912) More Information | Similar Motifs Found
4		1e-330	-7.600e+02	7.02%	2.09%	50.8bp (65.7bp)	FOS/MA0476.2/Jaspar(0.979) More Information | Similar Motifs Found
5		1e-304	-7.008e+02	26.67%	16.32%	56.3bp (60.7bp)	Erra(NR)/HepG2-Erra-ChIP-Seq(GSE31477)/Homer(0.911) More Information | Similar Motifs Found
6		1e-282	-6.498e+02	13.70%	6.61%	54.9bp (62.5bp)	FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer(0.941) More Information | Similar Motifs Found
7		1e-259	-5.980e+02	2.01%	0.19%	33.3bp (63.6bp)	SP5/MA1965.2/Jaspar(0.664) More Information | Similar Motifs Found
8		1e-165	-3.805e+02	3.01%	0.78%	51.5bp (57.8bp)	CTCF/MA0139.2/Jaspar(0.944) More Information | Similar Motifs Found
9		1e-157	-3.629e+02	0.50%	0.01%	49.3bp (17.5bp)	KLF12/MA0742.2/Jaspar(0.708) More Information | Similar Motifs Found
10		1e-110	-2.551e+02	1.82%	0.43%	40.9bp (57.1bp)	ZNF582/MA1983.2/Jaspar(0.598) More Information | Similar Motifs Found

Homer *de novo* Motif Results (../motifs_USF2//)

Non-redundant Motif File of Results
Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 20374
Total background sequences = 76143
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-16620	-3.827e+04	70.97%	3.07%	25.7bp (67.4bp)	USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer(0.965) More Information | Similar Motifs Found
2		1e-1193	-2.748e+03	12.43%	1.89%	39.4bp (64.2bp)	NFY(CCAAT)/Promoter/Homer(0.963) More Information | Similar Motifs Found
3		1e-469	-1.082e+03	2.64%	0.14%	39.5bp (59.7bp)	GFY(?)/Promoter/Homer(0.912) More Information | Similar Motifs Found
4		1e-330	-7.600e+02	7.02%	2.09%	50.8bp (65.7bp)	FOS/MA0476.2/Jaspar(0.979) More Information | Similar Motifs Found
5		1e-304	-7.008e+02	26.67%	16.32%	56.3bp (60.7bp)	Erra(NR)/HepG2-Erra-ChIP-Seq(GSE31477)/Homer(0.911) More Information | Similar Motifs Found
6		1e-282	-6.498e+02	13.70%	6.61%	54.9bp (62.5bp)	FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer(0.941) More Information | Similar Motifs Found
7		1e-259	-5.980e+02	2.01%	0.19%	33.3bp (63.6bp)	SP5/MA1965.2/Jaspar(0.664) More Information | Similar Motifs Found
8		1e-165	-3.805e+02	3.01%	0.78%	51.5bp (57.8bp)	CTCF/MA0139.2/Jaspar(0.944) More Information | Similar Motifs Found
9		1e-157	-3.629e+02	0.50%	0.01%	49.3bp (17.5bp)	KLF12/MA0742.2/Jaspar(0.708) More Information | Similar Motifs Found
10		1e-110	-2.551e+02	1.82%	0.43%	40.9bp (57.1bp)	ZNF582/MA1983.2/Jaspar(0.598) More Information | Similar Motifs Found

Motif Identified (Strongest Hit)

USF / E-box family motif
Consensus sequence: CACGTG

This is the canonical binding motif for **USF1 and USF2**, confirming that your ChIP-seq successfully captured true TF binding sites.

(A) Motif Enrichment Summary

Metric	Value	Meaning
% of peak sequences containing motif	**~71%**	Strong presence in USF2 binding regions
% in background	**~3%**	Rare in random genome
Fold enrichment	**~25–30×**	Highly enriched signal
P-value	**$\approx 1 \times 10^{-16620}$**	Extremely significant
Best known motif match	**USF1 (similarity = 0.965)**	Excellent match

(B) Why This Matters

✓ Confirms antibody worked
✓ Confirms true USF2 binding events
✓ Peaks reflect real biological binding, not noise
✓ Validates earlier QC steps

This is the **strongest evidence** that your ChIP-seq experiment succeeded.

17 .Functional Enrichment of Peaks (GREAT Analysis)

Purpose:
To determine biological pathways, regulatory programs, and gene networks associated with your USF2 peaks.

Working directory:

cd ~/GSE104247/bam

Before uploading to GREAT, peaks must be converted to **summit-centered 200 bp windows**.

Generate 200 bp Summit-Centered Regions (Unchanged Command)

awk 'BEGIN{OFS="\t"} {summit=$2 + $10; start=summit-100; if(start<0) start=0; end=summit+100; print $1, start, end, $4}' USF2_peaks.bed > USF2_summits_200bp.bed

Explanation

- $1 → chromosome
- $2 → peak start
- $10 → summit offset
- summit = start + summit_offset → absolute summit
- Window = summit −100 bp to +100 bp
- Ensures standard region sizes (200 bp each)

Output File: USF2_summits_200bp.bed

Contains:

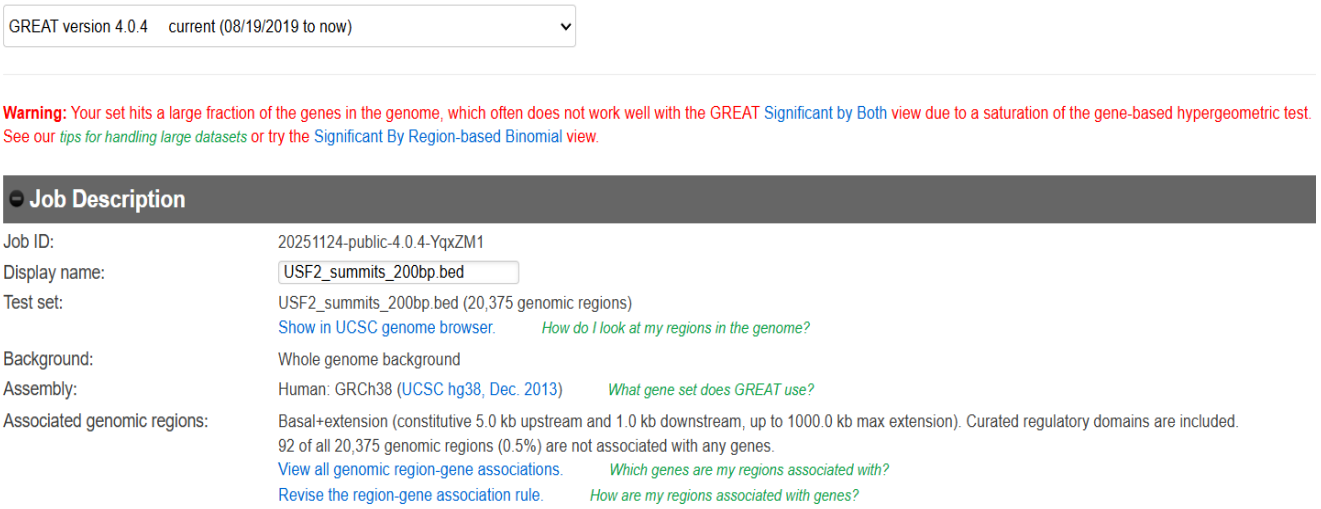
chr summit-100 summit+100 peak_name

A perfect input format for GREAT.

Run GREAT

1. Go to: https://great.stanford.edu
2. Upload: USF2_summits_200bp.bed
3. Select genome: **hg38**
4. Association rule: **Basal + extension** (default)

5. Run the job for GO terms, pathways, and regulatory enrichment.



GREAT displayed a warning because my dataset contains over 20,000 summit-centered peaks, which is typical for transcription factors with broad binding profiles such as USF2. When many peaks map to many genes, the gene-based hypergeometric test used by GREAT becomes saturated. However, this does not indicate an error. GREAT recommends using the "Region-Based Binomial" significance measure for large datasets, which is more appropriate for TF ChIP-seq. The analysis continues normally, and this view provides robust pathway and GO enrichment results.

Why summit-centered regions are required

HOMER peak regions are too broad

Large regions dilute TF motifs and inflate overlap with genes.

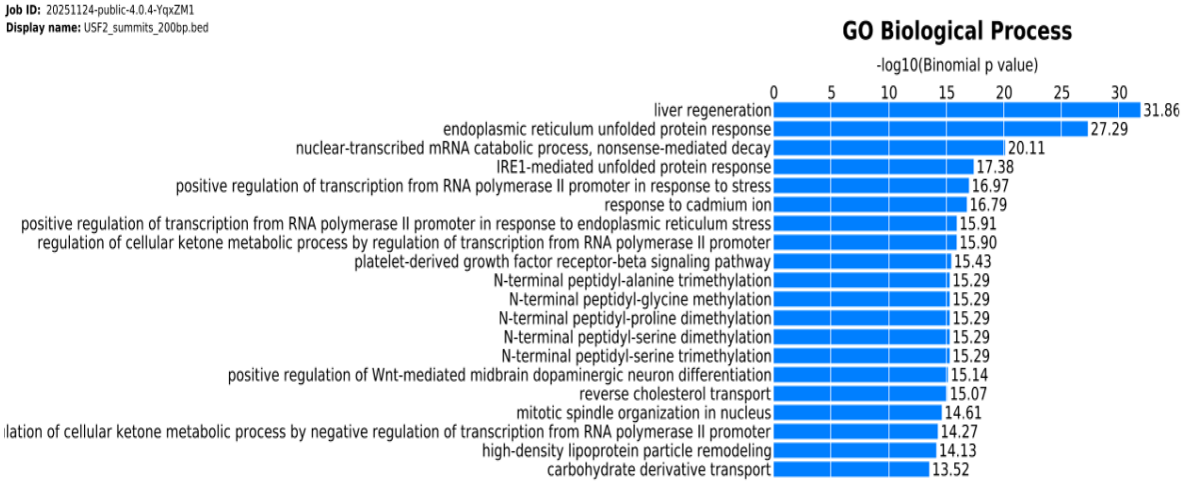
Summits represent the true binding point

This is the exact base where USF2 binds.

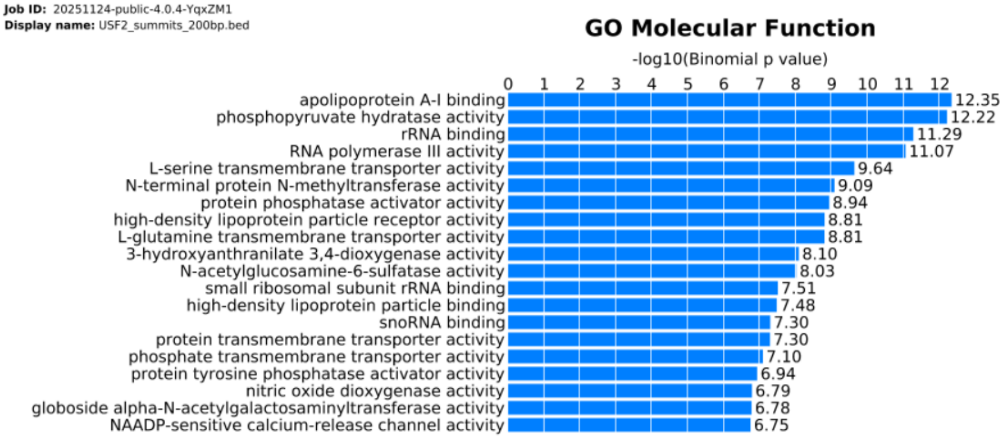
Fixed length (200 bp) ensures fairness

GREAT, HOMER, and motif tools all perform best with equal-sized windows.

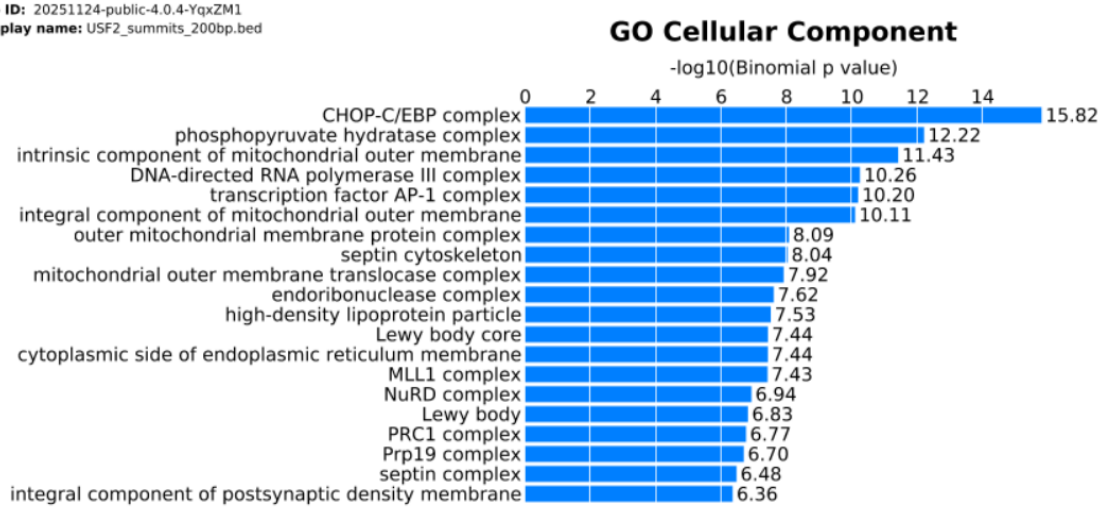
Job ID: 20251124-public-4.0.4-YqxZM1
Display name: USF2_summits_200bp.bed



Job ID: 20251124-public-4.0.4-YqxZM1
Display name: USF2_summits_200bp.bed



Job ID: 20251124-public-4.0.4-YqxZM1
Display name: USF2_summits_200bp.bed



18 . Export Files for Genome Browser Visualization (IGV)

Load the following files into IGV:

1. BAM Files (alignment data - Optional)

USF2_clean.bam

USF2_clean.bam.bai

INPUT_clean.bam

INPUT_clean.bam.bai

2. Signal Tracks (bigWigs)

USF2.bw

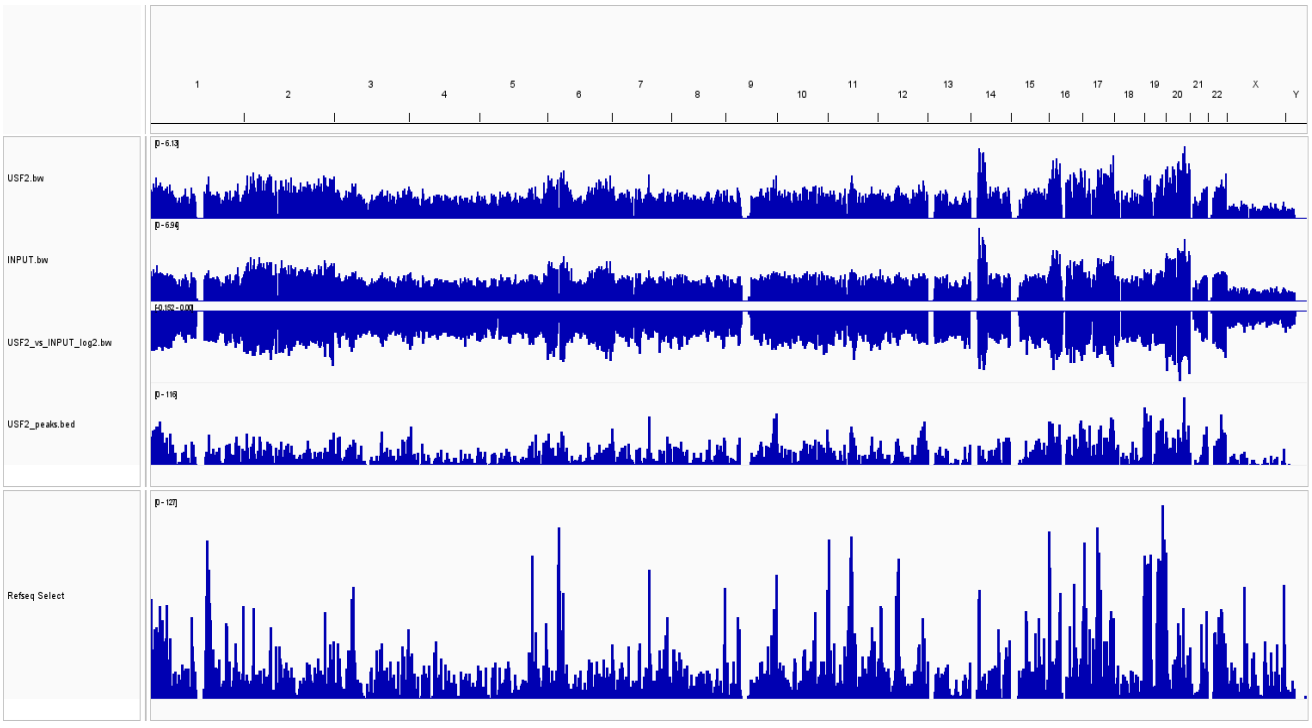
INPUT.bw

USF2_vs_INPUT_log2.bw

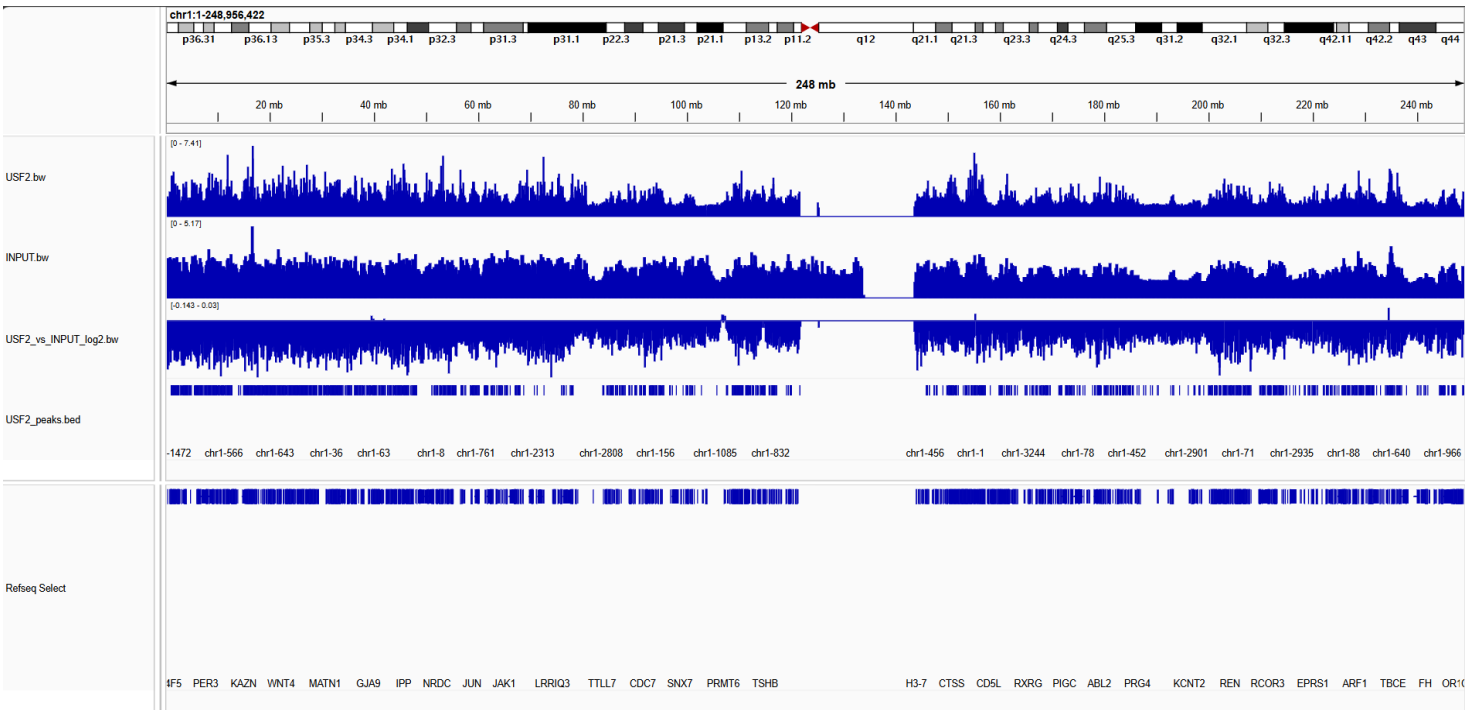
3. Peak File

USF2_peaks.bed

IGV – Whole Genome View



IGV – Whole Chromosome View (eg. Chr1)



Interpretation in IGV

✓ Confirm peak enrichment over Input
✓ Validate summit positions
✓ Inspect known USF2 target genes
✓ Check for artifacts, repetitive regions, or spikes

18 . (Biological Highlight) — Most Important Peak Example

Gene: CDK4
Peak ID: chr12-1
Coordinates: chr12:57752242–57752560
Annotation: promoter-TSS (–91 bp from TSS)
Peak Score: 394

This is biologically the **strongest peak** because:

1. Located at promoter (–91 bp)

Direct regulation, strongest impact on gene transcription.

2.CDK4 regulates the cell cycle

USF2 may regulate proliferation, cancer-related processes, and G1/S checkpoint.

3.High peak score (394)

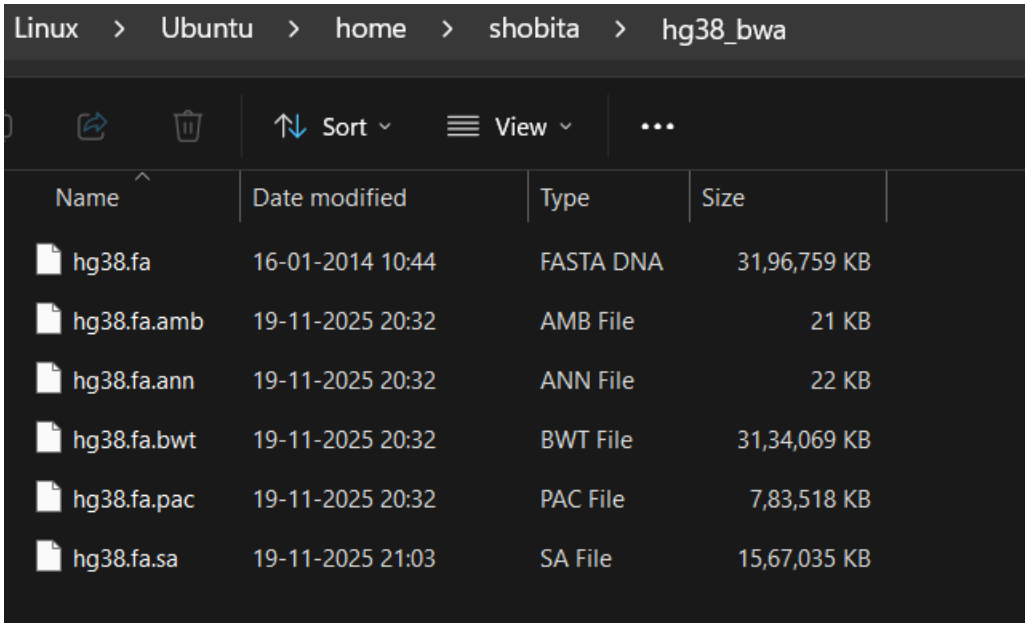
Represents a strong, confident, noise-free TF binding event.

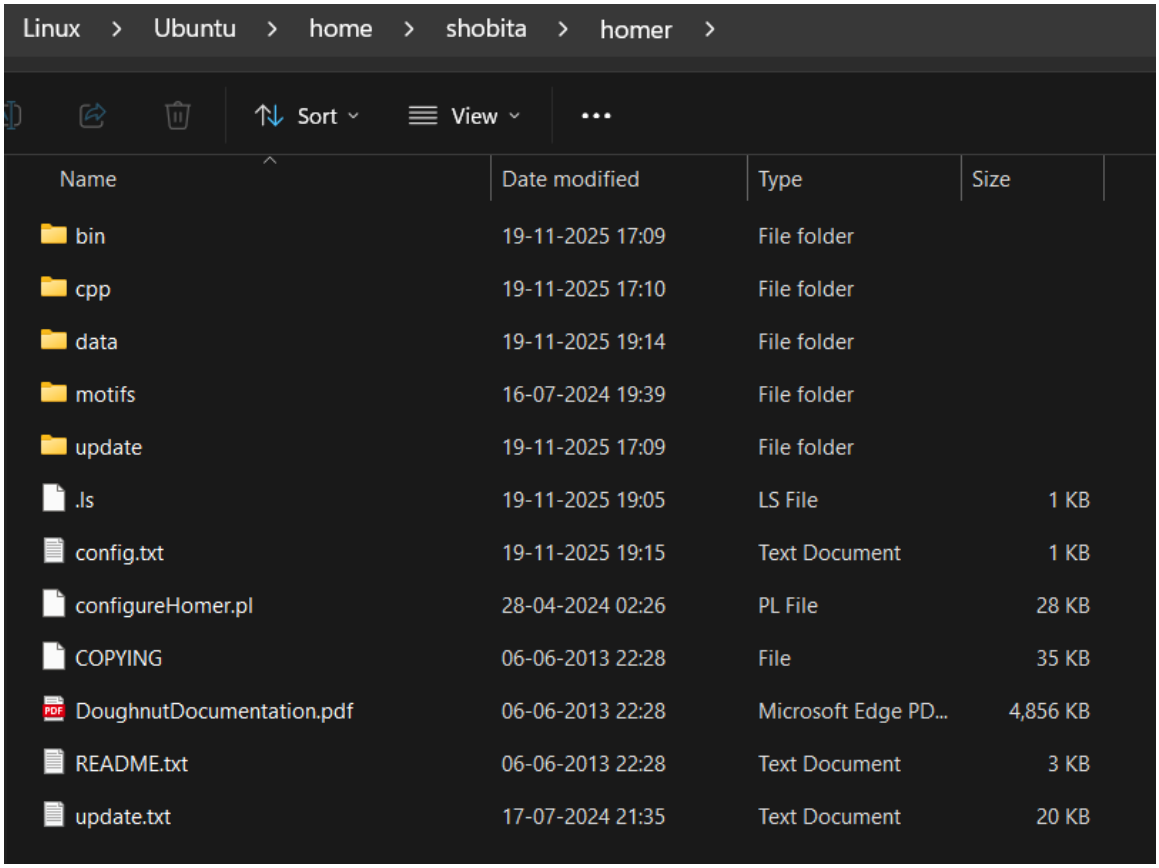
Summary and Conclusion

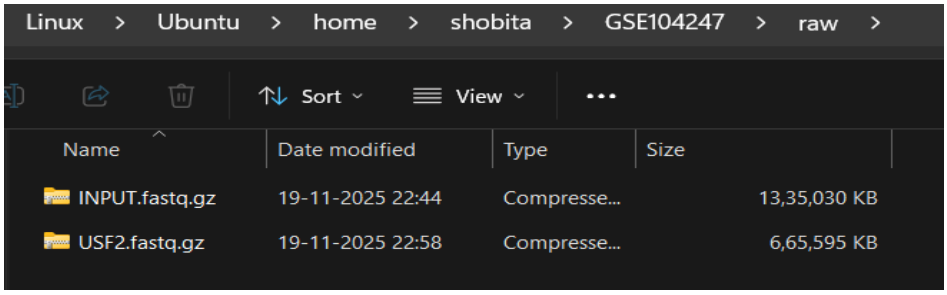
This ChIP-Seq pipeline successfully processed USF2 ChIP and Input samples through quality control, alignment, deduplication, peak calling, annotation, motif discovery, and visualization. The workflow produced biologically meaningful results, including strong USF2 enrichment at E-box motifs and functional associations with metabolic and stress-response pathways.

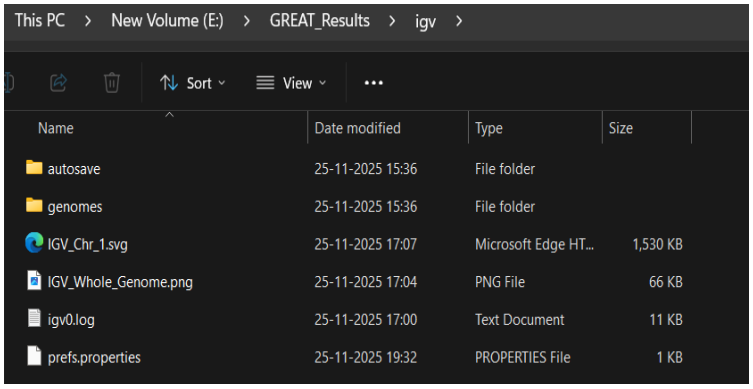
The final outputs—including normalized bigWig tracks, heatmaps, average profiles, and annotated peaks—provide a robust foundation for downstream analysis, such as transcriptional network inference and integration with RNA-Seq datasets.

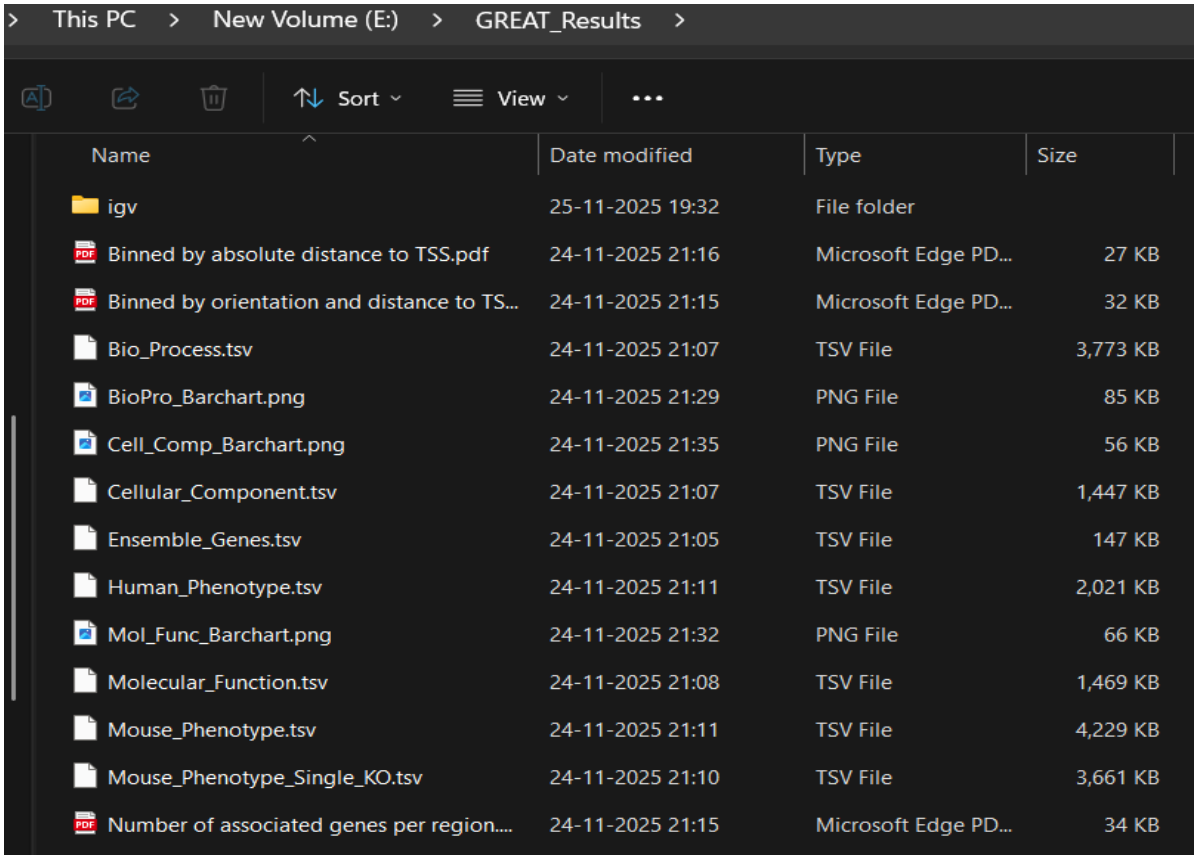
Snapshots of Iportant Folder Contents:






Linux > Ubuntu > home > shobita > GSE104247 > raw >
Sort
View
Name
Date modified
Type
Size
INPUT.fastq.gz
19-11-2025 22:44
Compresse...
13,35,030 KB
USF2.fastq.gz
19-11-2025 22:58
Compresse...
6,65,595 KB


This PC > New Volume (E:) > GREAT_Results > igv >
Sort
View
Name
Date modified
Type
Size
autosave
25-11-2025 15:36
File folder
genomes
25-11-2025 15:36
File folder
IGV_Chr_1.svg
25-11-2025 17:07
Microsoft Edge HT...
1,530 KB
IGV_Whole_Genome.png
25-11-2025 17:04
PNG File
66 KB
igv0.log
25-11-2025 17:00
Text Document
11 KB
prefs.properties
25-11-2025 19:32
PROPERTIES File
1 KB


This PC > New Volume (E:) > GREAT_Results >
Sort
View
Name
Date modified
Type
Size
igv
25-11-2025 19:32
File folder
Binned by absolute distance to TSS.pdf
24-11-2025 21:16
Microsoft Edge PD...
27 KB
Binned by orientation and distance to TS...
24-11-2025 21:15
Microsoft Edge PD...
32 KB
Bio_Process.tsv
24-11-2025 21:07
TSV File
3,773 KB
BioPro_Barchart.png
24-11-2025 21:29
PNG File
85 KB
Cell_Comp_Barchart.png
24-11-2025 21:35
PNG File
56 KB
Cellular_Component.tsv
24-11-2025 21:07
TSV File
1,447 KB
Ensemble_Genes.tsv
24-11-2025 21:05
TSV File
147 KB
Human_Phenotype.tsv
24-11-2025 21:11
TSV File
2,021 KB
Mol_Func_Barchart.png
24-11-2025 21:32
PNG File
66 KB
Molecular_Function.tsv
24-11-2025 21:08
TSV File
1,469 KB
Mouse_Phenotype.tsv
24-11-2025 21:11
TSV File
4,229 KB
Mouse_Phenotype_Single_KO.tsv
24-11-2025 21:10
TSV File
3,661 KB
Number of associated genes per region....
24-11-2025 21:15
Microsoft Edge PD...
34 KB