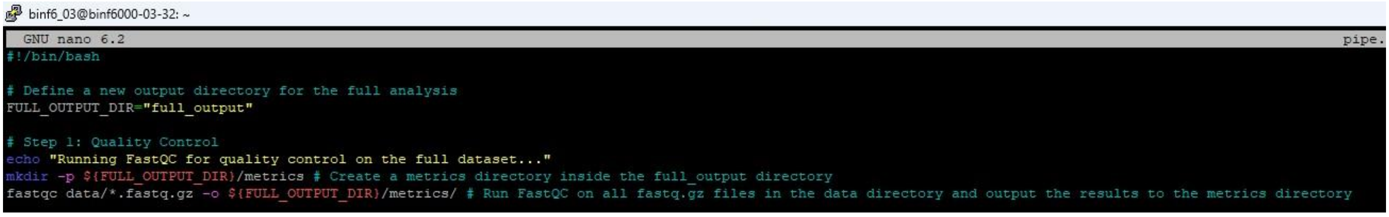
WORKSHOP 1: Analysis of genome-wide transcription factor-binding sites using ChIP-seq data

TASK1: Create an environment using mamba, and name the environment "workshop1"

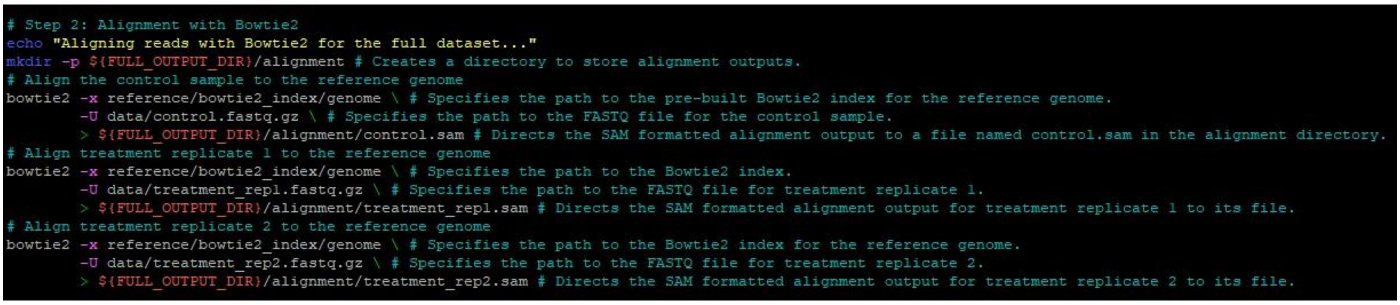
Address: 10.139.1.132 Username: binf6_03 Password: binf-GiantDingo76
mamba create -n workshop1 python=3.7
mamba activate workshop1
mamba install -c bioconda fastqc seqtk bowtie2 samtools macs2 idr

TASK2: PIPELINE

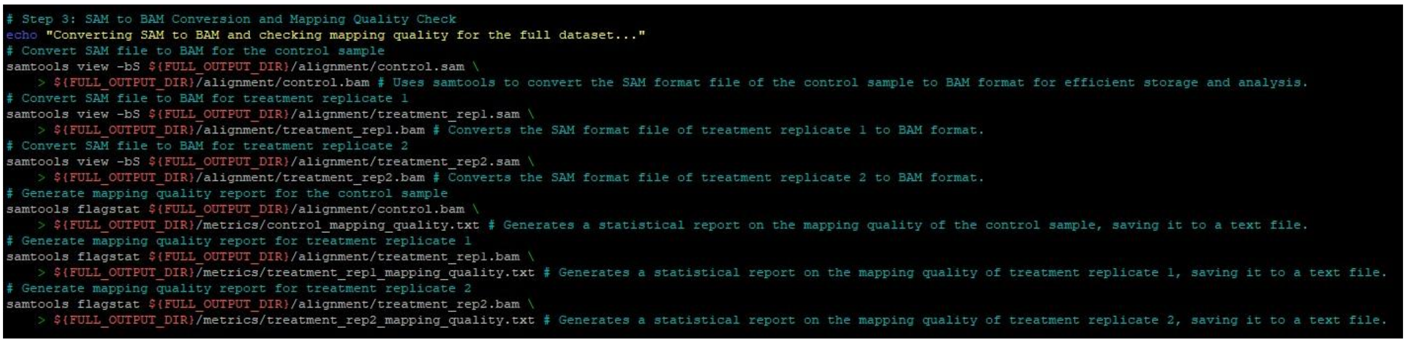


binf6_03@binf6000-03-32: ~

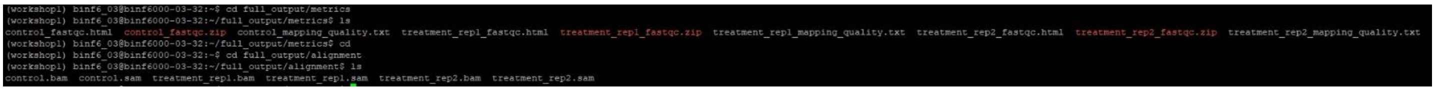
```
  GNU nano 6.2                                                                                                                           pipe.
#!/bin/bash

# Define a new output directory for the full analysis
FULL_OUTPUT_DIR="full_output"

# Step 1: Quality Control
echo "Running FastQC for quality control on the full dataset..."
mkdir -p ${FULL_OUTPUT_DIR}/metrics # Create a metrics directory inside the full_output directory
fastqc data/*.fastq.gz -o ${FULL_OUTPUT_DIR}/metrics/ # Run FastQC on all fastq.gz files in the data directory and output the results to the metrics directory
```



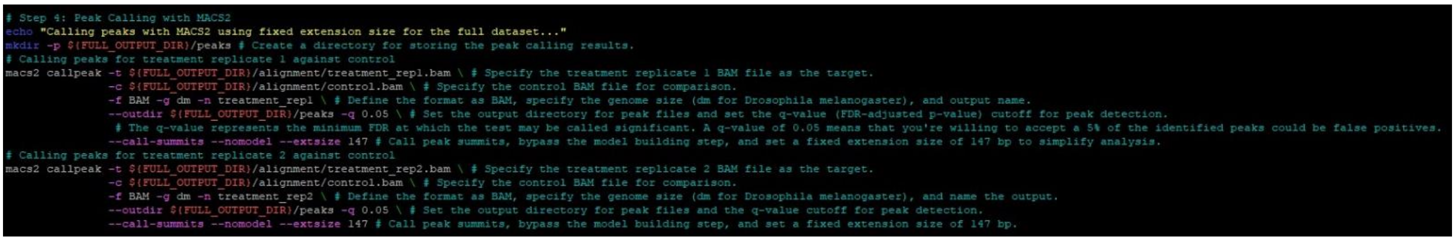
```
# Step 2: Alignment with Bowtie2
echo "Aligning reads with Bowtie2 for the full dataset..."
mkdir -p ${FULL_OUTPUT_DIR}/alignment # Creates a directory to store alignment outputs.
# Align the control sample to the reference genome
bowtie2 -x reference/bowtie2_index/genome \ # Specifies the path to the pre-built Bowtie2 index for the reference genome.
        -U data/control.fastq.gz \ # Specifies the path to the FASTQ file for the control sample.
        > ${FULL_OUTPUT_DIR}/alignment/control.sam # Directs the SAM formatted alignment output to a file named control.sam in the alignment directory.
# Align treatment replicate 1 to the reference genome
bowtie2 -x reference/bowtie2_index/genome \ # Specifies the path to the Bowtie2 index.
        -U data/treatment_rep1.fastq.gz \ # Specifies the path to the FASTQ file for treatment replicate 1.
        > ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.sam # Directs the SAM formatted alignment output for treatment replicate 1 to its file.
# Align treatment replicate 2 to the reference genome
bowtie2 -x reference/bowtie2_index/genome \ # Specifies the path to the Bowtie2 index for the reference genome.
        -U data/treatment_rep2.fastq.gz \ # Specifies the path to the FASTQ file for treatment replicate 2.
        > ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.sam # Directs the SAM formatted alignment output for treatment replicate 2 to its file.
```



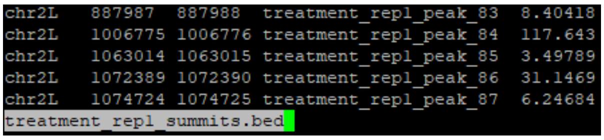
```
# Step 3: SAM to BAM Conversion and Mapping Quality Check
echo "Converting SAM to BAM and checking mapping quality for the full dataset..."
# Convert SAM file to BAM for the control sample
samtools view -bS ${FULL_OUTPUT_DIR}/alignment/control.sam \
    > ${FULL_OUTPUT_DIR}/alignment/control.bam # Uses samtools to convert the SAM format file of the control sample to BAM format for efficient storage and analysis.
# Convert SAM file to BAM for treatment replicate 1
samtools view -bS ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.sam \
    > ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.bam # Converts the SAM format file of treatment replicate 1 to BAM format.
# Convert SAM file to BAM for treatment replicate 2
samtools view -bS ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.sam \
    > ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.bam # Converts the SAM format file of treatment replicate 2 to BAM format.
# Generate mapping quality report for the control sample
samtools flagstat ${FULL_OUTPUT_DIR}/alignment/control.bam \
    > ${FULL_OUTPUT_DIR}/metrics/control_mapping_quality.txt # Generates a statistical report on the mapping quality of the control sample, saving it to a text file.
# Generate mapping quality report for treatment replicate 1
samtools flagstat ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.bam \
    > ${FULL_OUTPUT_DIR}/metrics/treatment_rep1_mapping_quality.txt # Generates a statistical report on the mapping quality of treatment replicate 1, saving it to a text file.
# Generate mapping quality report for treatment replicate 2
samtools flagstat ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.bam \
    > ${FULL_OUTPUT_DIR}/metrics/treatment_rep2_mapping_quality.txt # Generates a statistical report on the mapping quality of treatment replicate 2, saving it to a text file.
```



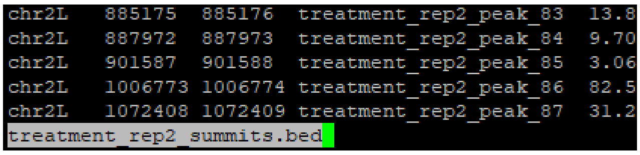
```
(workshop1) binf6_03@binf6000-03-32:~$ cd full_output/metrics
(workshop1) binf6_03@binf6000-03-32:~/full_output/metrics$ ls
control_fastqc.html  control_fastqc.zip  control_mapping_quality.txt  treatment_rep1_fastqc.html  treatment_rep1_fastqc.zip  treatment_rep1_mapping_quality.txt  treatment_rep2_fastqc.html  treatment_rep2_fastqc.zip  treatment_rep2_mapping_quality.txt
(workshop1) binf6_03@binf6000-03-32:~/full_output/metrics$ cd
(workshop1) binf6_03@binf6000-03-32:~$ cd full_output/alignment
(workshop1) binf6_03@binf6000-03-32:~/full_output/alignment$ ls
control.bam  control.sam  treatment_rep1.bam  treatment_rep1.sam  treatment_rep2.bam  treatment_rep2.sam
```



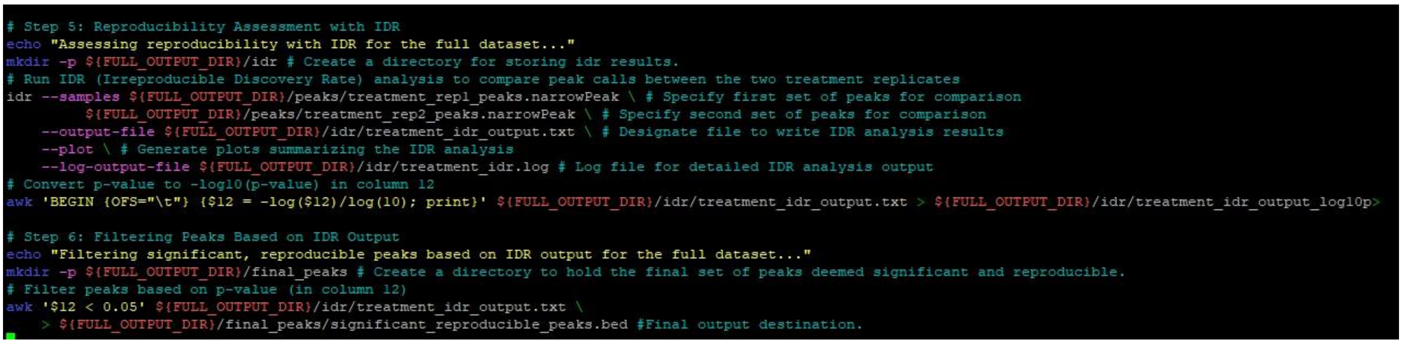
```
# Step 4: Peak Calling with MACS2
echo "Calling peaks with MACS2 using fixed extension size for the full dataset..."
mkdir -p ${FULL_OUTPUT_DIR}/peaks # Create a directory for storing the peak calling results.
# Calling peaks for treatment replicate 1 against control
macs2 callpeak -t ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.bam \ # Specify the treatment replicate 1 BAM file as the target.
               -c ${FULL_OUTPUT_DIR}/alignment/control.bam \ # Specify the control BAM file for comparison.
               -f BAM -g dm -n treatment_rep1 \ # Define the format as BAM, specify the genome size (dm for Drosophila melanogaster), and output name.
               --outdir ${FULL_OUTPUT_DIR}/peaks -q 0.05 \ # Set the output directory for peak files and set the q-value (FDR-adjusted p-value) cutoff for peak detection.
                # The q-value represents the minimum FDR at which the test may be called significant. A q-value of 0.05 means that you're willing to accept a 5% of the identified peaks could be false positives.
               --call-summits --nomodel --extsize 147 # Call peak summits, bypass the model building step, and set a fixed extension size of 147 bp to simplify analysis.
# Calling peaks for treatment replicate 2 against control
macs2 callpeak -t ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.bam \ # Specify the treatment replicate 2 BAM file as the target.
               -c ${FULL_OUTPUT_DIR}/alignment/control.bam \ # Specify the control BAM file for comparison.
               -f BAM -g dm -n treatment_rep2 \ # Define the format as BAM, specify the genome size (dm for Drosophila melanogaster), and name the output.
               --outdir ${FULL_OUTPUT_DIR}/peaks -q 0.05 \ # Set the output directory for peak files and the q-value cutoff for peak detection.
               --call-summits --nomodel --extsize 147 # Call peak summits, bypass the model building step, and set a fixed extension size of 147 bp.
```



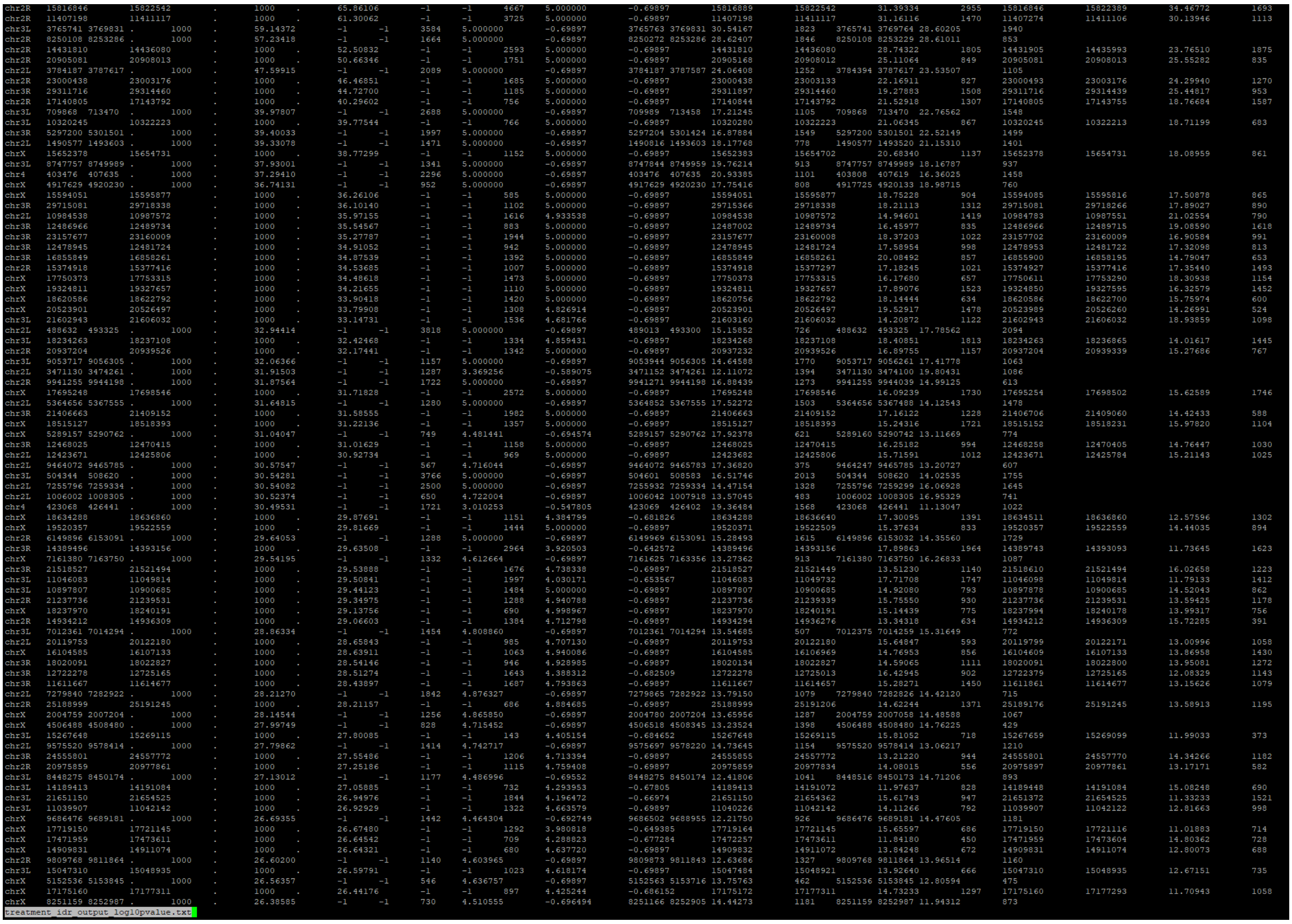
```
chr2L   887987  887988  treatment_rep1_peak_83  8.40418
chr2L   1006775 1006776 treatment_rep1_peak_84  117.643
chr2L   1063014 1063015 treatment_rep1_peak_85  3.49789
chr2L   1072389 1072390 treatment_rep1_peak_86  31.1469
chr2L   1074724 1074725 treatment_rep1_peak_87  6.24684
treatment_rep1_summits.bed
```



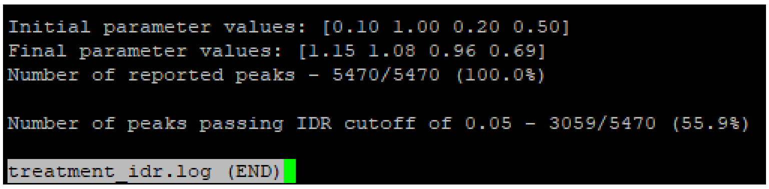
```
chr2L   885175  885176  treatment_rep2_peak_83  13.8
chr2L   887972  887973  treatment_rep2_peak_84  9.70
chr2L   901587  901588  treatment_rep2_peak_85  3.06
chr2L   1006773 1006774 treatment_rep2_peak_86  82.5
chr2L   1072408 1072409 treatment_rep2_peak_87  31.2
treatment_rep2_summits.bed
```



```
# Step 5: Reproducibility Assessment with IDR
echo "Assessing reproducibility with IDR for the full dataset..."
mkdir -p ${FULL_OUTPUT_DIR}/idr # Create a directory for storing idr results.
# Run IDR (Irreproducible Discovery Rate) analysis to compare peak calls between the two treatment replicates
idr --samples ${FULL_OUTPUT_DIR}/peaks/treatment_rep1_peaks.narrowPeak \ # Specify first set of peaks for comparison
         ${FULL_OUTPUT_DIR}/peaks/treatment_rep2_peaks.narrowPeak \ # Specify second set of peaks for comparison
    --output-file ${FULL_OUTPUT_DIR}/idr/treatment_idr_output.txt \ # Designate file to write IDR analysis results
    --plot \ # Generate plots summarizing the IDR analysis
    --log-output-file ${FULL_OUTPUT_DIR}/idr/treatment_idr.log # Log file for detailed IDR analysis output
# Convert p-value to -log10(p-value) in column 12
awk 'BEGIN {OFS="\t"} {$12 = -log($12)/log(10); print}' ${FULL_OUTPUT_DIR}/idr/treatment_idr_output.txt > ${FULL_OUTPUT_DIR}/idr/treatment_idr_output_log10p>

# Step 6: Filtering Peaks Based on IDR Output
echo "Filtering significant, reproducible peaks based on IDR output for the full dataset..."
mkdir -p ${FULL_OUTPUT_DIR}/final_peaks # Create a directory to hold the final set of peaks deemed significant and reproducible.
# Filter peaks based on p-value (in column 12)
awk '$12 < 0.05' ${FULL_OUTPUT_DIR}/idr/treatment_idr_output.txt \
    > ${FULL_OUTPUT_DIR}/final_peaks/significant_reproducible_peaks.bed #Final output destination.
```



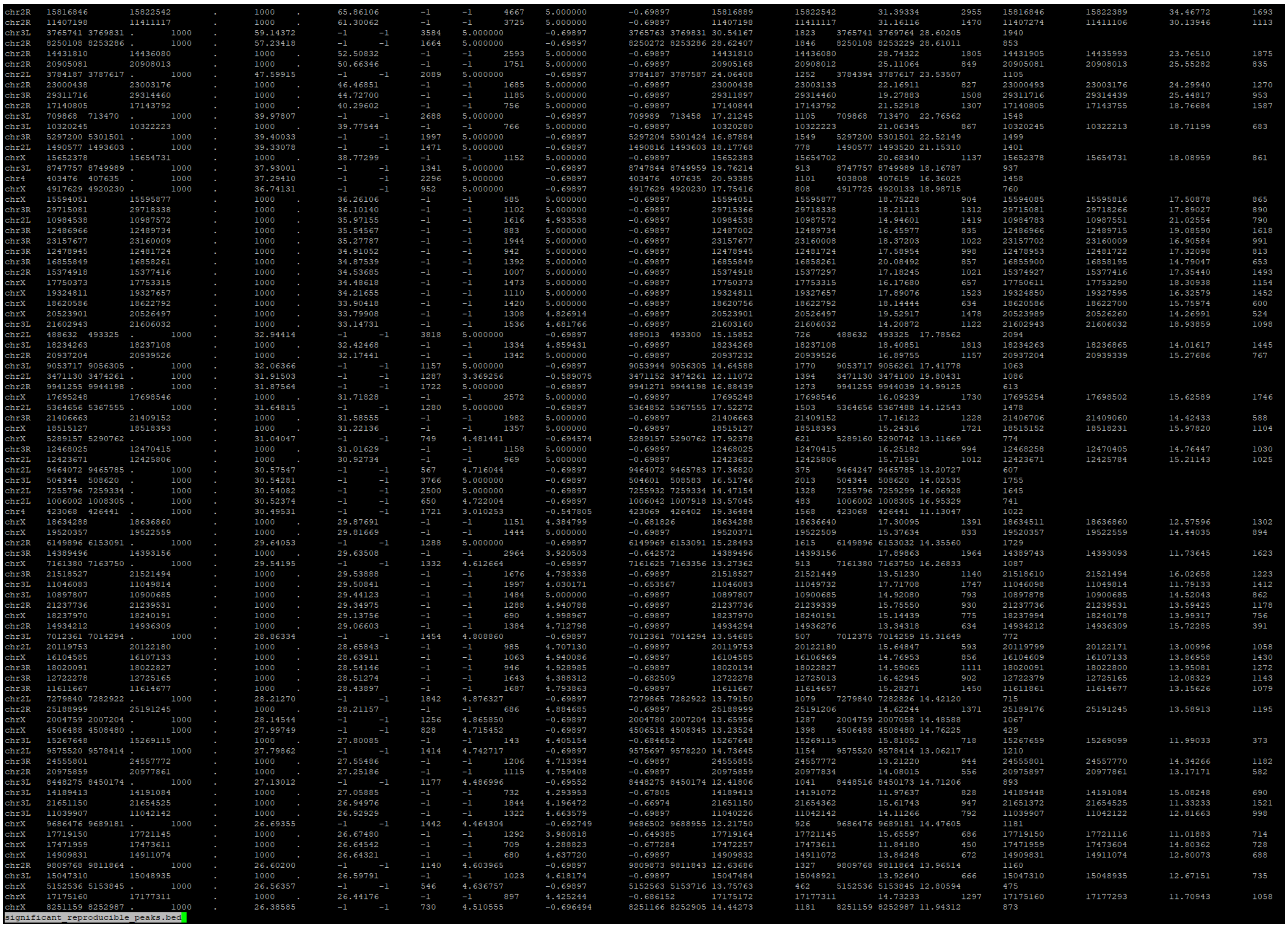
```
chr2R   15816846        15822542        .       1000    .       65.86106        -1      -1      4667    5.000000        -0.69897        15816889        15822542        31.39334        2955    15816846        15822389        34.46772        1693
chr2R   11407198        11411117        .       1000    .       61.30062        -1      -1      3725    5.000000        -0.69897        11407198        11411117        31.16116        1470    11407274        11411106        30.13946        1113
chr3L   3765741 3769831 .       1000    .       59.14372        -1      -1      3584    5.000000        -0.69897        3765763 3769831 30.54167        1823    3765741 3769764 28.60205        1940
chr2R   8250108 8253286 .       1000    .       57.23418        -1      -1      1664    5.000000        -0.69897        8250272 8253286 28.62407        1846    8250108 8253229 28.61011        853
chr2R   14431810        14436080        .       1000    .       52.50832        -1      -1      2593    5.000000        -0.69897        14431810        14436080        28.74322        1805    14431905        14435993        23.76510        1875
chr2R   20905081        20908013        .       1000    .       50.66346        -1      -1      1751    5.000000        -0.69897        20905168        20908012        25.11064        849     20905081        20908013        25.55282        835
chr2L   3784187 3787617 .       1000    .       47.59915        -1      -1      2089    5.000000        -0.69897        3784187 3787587 24.06408        1252    3784394 3787617 23.53507        1105
chr2R   23000438        23003176        .       1000    .       46.46851        -1      -1      1685    5.000000        -0.69897        23000438        23003133        22.16911        827     23000493        23003176        24.29940        1270
chr3R   29311716        29314460        .       1000    .       44.72700        -1      -1      1185    5.000000        -0.69897        29311897        29314460        19.27883        1508    29311716        29314439        25.44817        953
chr2R   17140805        17143792        .       1000    .       40.29602        -1      -1      756     5.000000        -0.69897        17140844        17143792        21.52918        1307    17140805        17143755        18.76684        1587
chr3L   709868  713470  .       1000    .       39.97807        -1      -1      2688    5.000000        -0.69897        709989  713458  17.21245        1105    709868  713470  22.76562        1548
chr3L   10320245        10322223        .       1000    .       39.77544        -1      -1      766     5.000000        -0.69897        10320280        10322223        21.06345        867     10320245        10322213        18.71199        683
chr3R   5297200 5301501 .       1000    .       39.40033        -1      -1      1997    5.000000        -0.69897        5297204 5301424 16.87884        1549    5297200 5301501 22.52149        1499
chr2L   1490577 1493603 .       1000    .       39.33078        -1      -1      1471    5.000000        -0.69897        1490816 1493603 18.17768        778     1490577 1493520 21.15310        1401
chrX    15652378        15654731        .       1000    .       38.77299        -1      -1      1152    5.000000        -0.69897        15652383        15654702        20.68340        1137    15652378        15654731        18.08959        861
chr3L   8747757 8749989 .       1000    .       37.93001        -1      -1      1341    5.000000        -0.69897        8747844 8749959 19.76214        913     8747757 8749989 18.16787        937
chr4    403476  407635  .       1000    .       37.29410        -1      -1      2296    5.000000        -0.69897        403476  407635  20.93385        1101    403808  407619  16.36025        1458
chrX    4917629 4920230 .       1000    .       36.74131        -1      -1      952     5.000000        -0.69897        4917629 4920230 17.75416        808     4917725 4920133 18.98715        760
chrX    15594051        15595877        .       1000    .       36.26106        -1      -1      585     5.000000        -0.69897        15594051        15595877        18.75228        904     15594085        15595816        17.50878        865
chr3R   29715081        29718338        .       1000    .       36.10140        -1      -1      1102    5.000000        -0.69897        29715366        29718338        18.21113        1312    29715081        29718266        17.89027        890
chr2L   10984538        10987572        .       1000    .       35.97155        -1      -1      1616    4.933538        -0.69897        10984538        10987572        14.94601        1419    10984783        10987551        21.02554        790
chr3R   12486966        12489734        .       1000    .       35.54567        -1      -1      883     5.000000        -0.69897        12487002        12489734        16.45977        835     12486966        12489715        19.08590        1618
chr3R   23157677        23160009        .       1000    .       35.27787        -1      -1      1944    5.000000        -0.69897        23157677        23160008        18.37203        1022    23157702        23160009        16.90584        991
chr3R   12478945        12481724        .       1000    .       34.91052        -1      -1      942     5.000000        -0.69897        12478945        12481724        17.58954        998     12478953        12481722        17.32098        813
chr3R   16855849        16858261        .       1000    .       34.87539        -1      -1      1392    5.000000        -0.69897        16855849        16858261        20.08492        857     16855900        16858195        14.79047        653
chr2R   15374918        15377416        .       1000    .       34.53685        -1      -1      1007    5.000000        -0.69897        15374918        15377297        17.18245        1021    15374927        15377416        17.35440        1493
chrX    17750373        17753315        .       1000    .       34.48618        -1      -1      1473    5.000000        -0.69897        17750373        17753315        16.17680        657     17750611        17753290        18.30938        1154
chrX    19324811        19327657        .       1000    .       34.21655        -1      -1      1110    5.000000        -0.69897        19324811        19327657        17.89076        1523    19324850        19327595        16.32579        1452
chrX    18620586        18622792        .       1000    .       33.90418        -1      -1      1420    5.000000        -0.69897        18620756        18622792        18.14444        634     18620586        18622700        15.75974        600
chrX    20523901        20526497        .       1000    .       33.79908        -1      -1      1308    4.826914        -0.69897        20523901        20526497        19.52917        1478    20523989        20526260        14.26991        524
chr3L   21602943        21606032        .       1000    .       33.14731        -1      -1      1536    4.681766        -0.69897        21603160        21606032        14.20872        1122    21602943        21606032        18.93859        1098
chr2L   488632  493325  .       1000    .       32.94414        -1      -1      3818    5.000000        -0.69897        489013  493300  15.15852        726     488632  493325  17.78562        2094
chr3L   18234263        18237108        .       1000    .       32.42468        -1      -1      1334    4.859431        -0.69897        18234268        18237108        18.40851        1813    18234263        18236865        14.01617        1445
chr2R   20937204        20939526        .       1000    .       32.17441        -1      -1      1342    5.000000        -0.69897        20937232        20939526        16.89755        1157    20937204        20939339        15.27686        767
chr3L   9053717 9056305 .       1000    .       32.06366        -1      -1      1157    5.000000        -0.69897        9053944 9056305 14.64588        1770    9053717 9056261 17.41778        1063
chr2L   3471130 3474261 .       1000    .       31.91503        -1      -1      1287    3.369256        -0.589075       3471152 3474261 12.11072        1394    3471130 3474100 19.80431        1086
chr2R   9941255 9944198 .       1000    .       31.87564        -1      -1      1722    5.000000        -0.69897        9941271 9944198 16.88439        1273    9941255 9944039 14.99125        613
chrX    17695248        17698546        .       1000    .       31.71828        -1      -1      2572    5.000000        -0.69897        17695248        17698546        16.09239        1730    17695254        17698502        15.62589        1746
chr2L   5364656 5367555 .       1000    .       31.64815        -1      -1      1280    5.000000        -0.69897        5364852 5367555 17.52272        1503    5364656 5367488 14.12543        1478
chr3R   21406663        21409152        .       1000    .       31.58555        -1      -1      1982    5.000000        -0.69897        21406663        21409152        17.16122        1228    21406706        21409060        14.42433        588
chrX    18515127        18518393        .       1000    .       31.22136        -1      -1      1357    5.000000        -0.69897        18515127        18518393        15.24316        1721    18515152        18518231        15.97820        1104
chrX    5289157 5290762 .       1000    .       31.04047        -1      -1      749     4.481441        -0.694574       5289157 5290762 17.92378        621     5289160 5290742 13.11669        774
chr3R   12468025        12470415        .       1000    .       31.01629        -1      -1      1158    5.000000        -0.69897        12468025        12470415        16.25182        994     12468258        12470405        14.76447        1030
chr2L   12423671        12425806        .       1000    .       30.92734        -1      -1      969     5.000000        -0.69897        12423682        12425806        15.71591        1012    12423671        12425784        15.21143        1025
chr2L   9464072 9465785 .       1000    .       30.57547        -1      -1      567     4.716044        -0.69897        9464072 9465783 17.36820        375     9464247 9465785 13.20727        607
chr3L   504344  508620  .       1000    .       30.54281        -1      -1      3766    5.000000        -0.69897        504601  508583  16.51746        2013    504344  508620  14.02535        1755
chr2L   7255796 7259334 .       1000    .       30.54082        -1      -1      2500    5.000000        -0.69897        7255932 7259334 14.47154        1328    7255796 7259299 16.06928        1645
chr2L   1006002 1008305 .       1000    .       30.52374        -1      -1      650     4.722004        -0.69897        1006042 1007918 13.57045        483     1006002 1008305 16.95329        741
chr4    423068  426441  .       1000    .       30.49531        -1      -1      1721    3.010253        -0.547805       423069  426402  19.36484        1568    423068  426441  11.13047        1022
chrX    18634288        18636860        .       1000    .       29.87691        -1      -1      1151    4.384799        -0.681826       18634288        18636640        17.30095        1391    18634511        18636860        12.57596        1302
chrX    19520357        19522559        .       1000    .       29.81669        -1      -1      1444    5.000000        -0.69897        19520371        19522509        15.37634        833     19520357        19522559        14.44035        894
chr2R   6149896 6153091 .       1000    .       29.64053        -1      -1      1288    5.000000        -0.69897        6149969 6153091 15.28493        1615    6149896 6153032 14.35560        1729
chr3R   14389496        14393156        .       1000    .       29.63508        -1      -1      2964    3.920503        -0.642572       14389496        14393156        17.89863        1964    14389743        14393093        11.73645        1623
chrX    7161380 7163750 .       1000    .       29.54195        -1      -1      1332    4.612664        -0.69897        7161625 7163356 13.27362        913     7161380 7163750 16.26833        1087
chr3R   21518527        21521494        .       1000    .       29.53888        -1      -1      1676    4.738338        -0.69897        21518527        21521449        13.51230        1140    21518610        21521494        16.02658        1223
chr3L   11046083        11049814        .       1000    .       29.50841        -1      -1      1997    4.030171        -0.653567       11046083        11049732        17.71708        1747    11046098        11049814        11.79133        1412
chr3L   10897807        10900685        .       1000    .       29.44123        -1      -1      1484    5.000000        -0.69897        10897807        10900685        14.92080        793     10897878        10900685        14.52043        862
chr2R   21237736        21239531        .       1000    .       29.34975        -1      -1      1288    4.940788        -0.69897        21237736        21239339        15.75550        930     21237736        21239531        13.59425        1178
chrX    18237970        18240191        .       1000    .       29.13756        -1      -1      690     4.998967        -0.69897        18237970        18240191        15.14439        775     18237994        18240178        13.99317        756
chr2R   14934212        14936309        .       1000    .       29.06603        -1      -1      1384    4.712798        -0.69897        14934294        14936276        13.34318        634     14934212        14936309        15.72285        391
chr3L   7012361 7014294 .       1000    .       28.86334        -1      -1      1454    4.808860        -0.69897        7012361 7014294 13.54685        507     7012375 7014259 15.31649        772
chr2L   20119753        20122180        .       1000    .       28.65843        -1      -1      985     4.707130        -0.69897        20119753        20122180        15.64847        593     20119799        20122171        13.00996        1058
chrX    16104585        16107133        .       1000    .       28.63911        -1      -1      1063    4.940086        -0.69897        16104585        16106969        14.76953        856     16104609        16107133        13.86958        1430
chr3R   18020091        18022827        .       1000    .       28.54146        -1      -1      946     4.928985        -0.69897        18020134        18022827        14.59065        1111    18020091        18022800        13.95081        1272
chr3R   12722278        12725165        .       1000    .       28.51274        -1      -1      1643    4.388312        -0.682509       12722278        12725013        16.42945        902     12722379        12725165        12.08329        1143
chr3R   11611667        11614677        .       1000    .       28.43897        -1      -1      1687    4.793863        -0.69897        11611667        11614657        15.28271        1450    11611861        11614677        13.15626        1079
chr2L   7279840 7282922 .       1000    .       28.21270        -1      -1      1842    4.876327        -0.69897        7279865 7282922 13.79150        1079    7279840 7282826 14.42120        715
chr2R   25188999        25191245        .       1000    .       28.21157        -1      -1      686     4.884685        -0.69897        25188999        25191206        14.62244        1371    25189176        25191245        13.58913        1195
chrX    2004759 2007204 .       1000    .       28.14544        -1      -1      1256    4.865850        -0.69897        2004780 2007204 13.65956        1287    2004759 2007058 14.48588        1067
chrX    4506488 4508480 .       1000    .       27.99749        -1      -1      828     4.715452        -0.69897        4506518 4508345 13.23524        1398    4506488 4508480 14.76225        429
chr3L   15267648        15269115        .       1000    .       27.80085        -1      -1      143     4.405154        -0.684652       15267648        15269115        15.81052        718     15267659        15269099        11.99033        373
chr2L   9575520 9578414 .       1000    .       27.79862        -1      -1      1414    4.742717        -0.69897        9575697 9578220 14.73645        1154    9575520 9578414 13.06217        1210
chr3R   24555801        24557772        .       1000    .       27.55486        -1      -1      1206    4.713394        -0.69897        24555855        24557772        13.21220        944     24555801        24557770        14.34266        1182
chr2R   20975859        20977861        .       1000    .       27.25186        -1      -1      1115    4.759408        -0.69897        20975859        20977834        14.08015        556     20975897        20977861        13.17171        582
chr3L   8448275 8450174 .       1000    .       27.13012        -1      -1      1177    4.486996        -0.69552        8448275 8450174 12.41806        1041    8448516 8450173 14.71206        893
chr3L   14189413        14191084        .       1000    .       27.05885        -1      -1      732     4.293953        -0.67805        14189413        14191072        11.97637        828     14189448        14191084        15.08248        690
chr3L   21651150        21654525        .       1000    .       26.94976        -1      -1      1844    4.196472        -0.66974        21651150        21654362        15.61743        947     21651372        21654525        11.33233        1521
chr3L   11039907        11042142        .       1000    .       26.92929        -1      -1      1322    4.663579        -0.69897        11040226        11042142        14.11266        792     11039907        11042122        12.81663        998
chrX    9686476 9689181 .       1000    .       26.69355        -1      -1      1442    4.464304        -0.692749       9686502 9688955 12.21750        926     9686476 9689181 14.47605        1181
chrX    17719150        17721145        .       1000    .       26.67480        -1      -1      1292    3.980818        -0.649385       17719164        17721145        15.65597        686     17719150        17721116        11.01883        714
chrX    17471959        17473611        .       1000    .       26.64542        -1      -1      709     4.288823        -0.677284       17472257        17473611        11.84180        450     17471959        17473604        14.80362        728
chrX    14909831        14911074        .       1000    .       26.64321        -1      -1      680     4.637720        -0.69897        14909832        14911072        13.84248        672     14909831        14911074        12.80073        688
chr2R   9809768 9811864 .       1000    .       26.60200        -1      -1      1140    4.603965        -0.69897        9809873 9811843 12.63686        1327    9809768 9811864 13.96514        1160
chr3L   15047310        15048935        .       1000    .       26.59791        -1      -1      1023    4.618174        -0.69897        15047484        15048921        13.92640        666     15047310        15048935        12.67151        735
chrX    5152536 5153845 .       1000    .       26.56357        -1      -1      546     4.636757        -0.69897        5152563 5153716 13.75763        462     5152536 5153845 12.80594        475
chrX    17175160        17177311        .       1000    .       26.44176        -1      -1      897     4.425244        -0.686152       17175172        17177311        14.73233        1297    17175160        17177293        11.70943        1058
chrX    8251159 8252987 .       1000    .       26.38585        -1      -1      730     4.510555        -0.696494       8251166 8252905 14.44273        1181    8251159 8252987 11.94312        873
treatment_idr_output_log10pvalue.txt
```



```
Initial parameter values: [0.10 1.00 0.20 0.50]
Final parameter values: [1.15 1.08 0.96 0.69]
Number of reported peaks - 5470/5470 (100.0%)

Number of peaks passing IDR cutoff of 0.05 - 3059/5470 (55.9%)

treatment_idr.log (END)
```

TASK3: Identify a gene bound by the transcription factor CTCF

less significant_reproducible_peaks.bed.



```
chr2R	15816846	15822542	.	1000	.	65.86106	-1	-1	4667	5.000000	-0.69897	15816889	15822542	31.39334	2955	15816846	15822389	34.46772	1693
chr2R	11407198	11411117	.	1000	.	61.30062	-1	-1	3725	5.000000	-0.69897	11407198	11411117	31.16116	1470	11407274	11411106	30.13946	1113
chr3L	3765741	3769831	.	1000	.	59.14372	-1	-1	3584	5.000000	-0.69897	3765763	3769831	30.54167	1823	3765741	3769764	28.60205	1940
chr2R	8250108	8253286	.	1000	.	57.23418	-1	-1	1664	5.000000	-0.69897	8250272	8253286	28.62407	1846	8250108	8253229	28.61011	853
chr2R	14431810	14436080	.	1000	.	52.50832	-1	-1	2593	5.000000	-0.69897	14431810	14436080	28.74322	1805	14431905	14435993	23.76510	1875
chr2R	20905081	20908013	.	1000	.	50.66346	-1	-1	1751	5.000000	-0.69897	20905168	20908012	25.11064	849	20905081	20908013	25.55282	835
chr2L	3784187	3787617	.	1000	.	47.59915	-1	-1	2089	5.000000	-0.69897	3784187	3787587	24.06408	1252	3784394	3787617	23.53507	1105
chr2R	23000438	23003176	.	1000	.	46.46851	-1	-1	1685	5.000000	-0.69897	23000438	23003133	22.16911	827	23000493	23003176	24.29940	1270
chr3R	29311716	29314460	.	1000	.	44.72700	-1	-1	1185	5.000000	-0.69897	29311897	29314460	19.27883	1508	29311716	29314439	25.44817	953
chr2R	17140805	17143792	.	1000	.	40.29602	-1	-1	756	5.000000	-0.69897	17140844	17143792	21.52918	1307	17140805	17143755	18.76684	1587
chr3L	709868	713470	.	1000	.	39.97807	-1	-1	2688	5.000000	-0.69897	709989	713458	17.21245	1105	709868	713470	22.76562	1548
chr3L	10320245	10322223	.	1000	.	39.77544	-1	-1	766	5.000000	-0.69897	10320280	10322223	21.06345	867	10320245	10322213	18.71199	683
chr3R	5297200	5301501	.	1000	.	39.40033	-1	-1	1997	5.000000	-0.69897	5297204	5301424	16.87884	1549	5297200	5301501	22.52149	1499
chr2L	1490577	1493603	.	1000	.	39.33078	-1	-1	1471	5.000000	-0.69897	1490816	1493603	18.17768	778	1490577	1493520	21.15310	1401
chrX	15652378	15654731	.	1000	.	38.77299	-1	-1	1152	5.000000	-0.69897	15652383	15654702	20.68340	1137	15652378	15654731	18.08959	861
chr3L	8747757	8749989	.	1000	.	37.93001	-1	-1	1341	5.000000	-0.69897	8747844	8749959	19.76214	913	8747757	8749989	18.16787	937
chr4	403476	407635	.	1000	.	37.29410	-1	-1	2296	5.000000	-0.69897	403476	407635	20.93385	1101	403808	407619	16.36025	1458
chrX	4917629	4920230	.	1000	.	36.74131	-1	-1	952	5.000000	-0.69897	4917629	4920230	17.75416	808	4917725	4920133	18.98715	760
chrX	15594051	15595877	.	1000	.	36.26106	-1	-1	585	5.000000	-0.69897	15594051	15595877	18.75228	904	15594085	15595816	17.50878	865
chr3R	29715081	29718338	.	1000	.	36.10140	-1	-1	1102	5.000000	-0.69897	29715366	29718338	18.21113	1312	29715081	29718266	17.89027	890
chr2L	10984538	10987572	.	1000	.	35.97155	-1	-1	1616	4.933538	-0.69897	10984538	10987572	14.94601	1419	10984783	10987551	21.02554	790
chr3R	12486966	12489734	.	1000	.	35.54567	-1	-1	883	5.000000	-0.69897	12487002	12489734	16.45977	835	12486966	12489715	19.08590	1618
chr3R	23157677	23160009	.	1000	.	35.27787	-1	-1	1944	5.000000	-0.69897	23157677	23160008	18.37203	1022	23157702	23160009	16.90584	991
chr3R	12478945	12481724	.	1000	.	34.91052	-1	-1	942	5.000000	-0.69897	12478945	12481724	17.58954	998	12478953	12481722	17.32098	813
chr3R	16855849	16858261	.	1000	.	34.87539	-1	-1	1392	5.000000	-0.69897	16855849	16858261	20.08492	857	16855900	16858195	14.79047	653
chr2R	15374918	15377416	.	1000	.	34.53685	-1	-1	1007	5.000000	-0.69897	15374918	15377297	17.18245	1021	15374927	15377416	17.35440	1493
chrX	17750373	17753315	.	1000	.	34.48618	-1	-1	1473	5.000000	-0.69897	17750373	17753315	16.17680	657	17750611	17753290	18.30938	1154
chrX	19324811	19327657	.	1000	.	34.21655	-1	-1	1110	5.000000	-0.69897	19324811	19327657	17.89076	1523	19324850	19327595	16.32579	1452
chrX	18620586	18622792	.	1000	.	33.90418	-1	-1	1420	5.000000	-0.69897	18620756	18622792	18.14444	634	18620586	18622700	15.75974	600
chrX	20523901	20526497	.	1000	.	33.79908	-1	-1	1308	4.826914	-0.69897	20523901	20526497	19.52917	1478	20523989	20526260	14.26991	524
chr3L	21602943	21606032	.	1000	.	33.14731	-1	-1	1536	4.681766	-0.69897	21603160	21606032	14.20872	1122	21602943	21606032	18.93859	1098
chr2L	488632	493325	.	1000	.	32.94414	-1	-1	3818	5.000000	-0.69897	489013	493300	15.15852	726	488632	493325	17.78562	2094
chr3L	18234263	18237108	.	1000	.	32.42468	-1	-1	1334	4.859431	-0.69897	18234268	18237108	18.40851	1813	18234263	18236865	14.01617	1445
chr2R	20937204	20939526	.	1000	.	32.17441	-1	-1	1342	5.000000	-0.69897	20937232	20939526	16.89755	1157	20937204	20939339	15.27686	767
chr3L	9053717	9056305	.	1000	.	32.06366	-1	-1	1157	5.000000	-0.69897	9053944	9056305	14.64588	1770	9053717	9056261	17.41778	1063
chr2L	3471130	3474261	.	1000	.	31.91503	-1	-1	1287	3.369256	-0.589075	3471152	3474261	12.11072	1394	3471130	3474100	19.80431	1086
chr2R	9941255	9944198	.	1000	.	31.87564	-1	-1	1722	5.000000	-0.69897	9941271	9944198	16.88439	1273	9941255	9944039	14.99125	613
chrX	17695248	17698546	.	1000	.	31.71828	-1	-1	2572	5.000000	-0.69897	17695248	17698546	16.09239	1730	17695254	17698502	15.62589	1746
chr2L	5364656	5367555	.	1000	.	31.64815	-1	-1	1280	5.000000	-0.69897	5364852	5367555	17.52272	1503	5364656	5367488	14.12543	1478
chr3R	21406663	21409152	.	1000	.	31.58555	-1	-1	1982	5.000000	-0.69897	21406663	21409152	17.16122	1228	21406706	21409060	14.42433	588
chrX	18515127	18518393	.	1000	.	31.22136	-1	-1	1357	5.000000	-0.69897	18515127	18518393	15.24316	1721	18515152	18518231	15.97820	1104
chrX	5289157	5290762	.	1000	.	31.04047	-1	-1	749	4.481441	-0.694574	5289157	5290762	17.92378	621	5289160	5290742	13.11669	774
chr3R	12468025	12470415	.	1000	.	31.01629	-1	-1	1158	5.000000	-0.69897	12468025	12470415	16.25182	994	12468258	12470405	14.76447	1030
chr2L	12423671	12425806	.	1000	.	30.92734	-1	-1	969	5.000000	-0.69897	12423682	12425806	15.71591	1012	12423671	12425784	15.21143	1025
chr2L	9464072	9465785	.	1000	.	30.57547	-1	-1	567	4.716044	-0.69897	9464072	9465783	17.36820	375	9464247	9465785	13.20727	607
chr3L	504344	508620	.	1000	.	30.54281	-1	-1	3766	5.000000	-0.69897	504601	508583	16.51746	2013	504344	508620	14.02535	1755
chr2L	7255796	7259334	.	1000	.	30.54082	-1	-1	2500	5.000000	-0.69897	7255932	7259334	14.47154	1328	7255796	7259299	16.06928	1645
chr2L	1006002	1008305	.	1000	.	30.52374	-1	-1	650	4.722004	-0.69897	1006042	1007918	13.57045	483	1006002	1008305	16.95329	741
chr4	423068	426441	.	1000	.	30.49531	-1	-1	1721	3.010253	-0.547805	423069	426402	19.36484	1568	423068	426441	11.13047	1022
chrX	18634288	18636860	.	1000	.	29.87691	-1	-1	1151	4.384799	-0.681826	18634288	18636640	17.30095	1391	18634511	18636860	12.57596	1302
chrX	19520357	19522559	.	1000	.	29.81669	-1	-1	1444	5.000000	-0.69897	19520371	19522509	15.37634	833	19520357	19522559	14.44035	894
chr2R	6149896	6153091	.	1000	.	29.64053	-1	-1	1288	5.000000	-0.69897	6149969	6153091	15.28493	1615	6149896	6153032	14.35560	1729
chr3R	14389496	14393156	.	1000	.	29.63508	-1	-1	2964	3.920503	-0.642572	14389496	14393156	17.89863	1964	14389743	14393093	11.73645	1623
chrX	7161380	7163750	.	1000	.	29.54195	-1	-1	1332	4.612664	-0.69897	7161625	7163356	13.27362	913	7161380	7163750	16.26833	1087
chr3R	21518527	21521494	.	1000	.	29.53888	-1	-1	1676	4.738338	-0.69897	21518527	21521449	13.51230	1140	21518610	21521494	16.02658	1223
chr3L	11046083	11049814	.	1000	.	29.50841	-1	-1	1997	4.030171	-0.653567	11046083	11049732	17.71708	1747	11046098	11049814	11.79133	1412
chr3L	10897807	10900685	.	1000	.	29.44123	-1	-1	1484	5.000000	-0.69897	10897807	10900685	14.92080	793	10897878	10900685	14.52043	862
chr2R	21237736	21239531	.	1000	.	29.34975	-1	-1	1288	4.940788	-0.69897	21237736	21239339	15.75550	930	21237736	21239531	13.59425	1178
chrX	18237970	18240191	.	1000	.	29.13756	-1	-1	690	4.998967	-0.69897	18237970	18240191	15.14439	775	18237994	18240178	13.99317	756
chr2R	14934212	14936309	.	1000	.	29.06603	-1	-1	1384	4.712798	-0.69897	14934294	14936276	13.34318	634	14934212	14936309	15.72285	391
chr3L	7012361	7014294	.	1000	.	28.86334	-1	-1	1454	4.808860	-0.69897	7012361	7014294	13.54685	507	7012375	7014259	15.31649	772
chr2L	20119753	20122180	.	1000	.	28.65843	-1	-1	985	4.707130	-0.69897	20119753	20122180	15.64847	593	20119799	20122171	13.00996	1058
chrX	16104585	16107133	.	1000	.	28.63911	-1	-1	1063	4.940086	-0.69897	16104585	16106969	14.76953	856	16104609	16107133	13.86958	1430
chr3R	18020091	18022827	.	1000	.	28.54146	-1	-1	946	4.928985	-0.69897	18020134	18022827	14.59065	1111	18020091	18022800	13.95081	1272
chr3R	12722278	12725165	.	1000	.	28.51274	-1	-1	1643	4.388312	-0.682509	12722278	12725013	16.42945	902	12722379	12725165	12.08329	1143
chr3R	11611667	11614677	.	1000	.	28.43897	-1	-1	1687	4.793863	-0.69897	11611667	11614657	15.28271	1450	11611861	11614677	13.15626	1079
chr2L	7279840	7282922	.	1000	.	28.21270	-1	-1	1842	4.876327	-0.69897	7279865	7282922	13.79150	1079	7279840	7282826	14.42120	715
chr2R	25188999	25191245	.	1000	.	28.21157	-1	-1	686	4.884685	-0.69897	25188999	25191206	14.62244	1371	25189176	25191245	13.58913	1195
chrX	2004759	2007204	.	1000	.	28.14544	-1	-1	1256	4.865850	-0.69897	2004780	2007204	13.65956	1287	2004759	2007058	14.48588	1067
chrX	4506488	4508480	.	1000	.	27.99749	-1	-1	828	4.715452	-0.69897	4506518	4508345	13.23524	1398	4506488	4508480	14.76225	429
chr3L	15267648	15269115	.	1000	.	27.80085	-1	-1	143	4.405154	-0.684652	15267648	15269115	15.81052	718	15267659	15269099	11.99033	373
chr2L	9575520	9578414	.	1000	.	27.79862	-1	-1	1414	4.742717	-0.69897	9575697	9578220	14.73645	1154	9575520	9578414	13.06217	1210
chr3R	24555801	24557772	.	1000	.	27.55486	-1	-1	1206	4.713394	-0.69897	24555855	24557772	13.21220	944	24555801	24557770	14.34266	1182
chr2R	20975859	20977861	.	1000	.	27.25186	-1	-1	1115	4.759408	-0.69897	20975859	20977834	14.08015	556	20975897	20977861	13.17171	582
chr3L	8448275	8450174	.	1000	.	27.13012	-1	-1	1177	4.486996	-0.69552	8448275	8450174	12.41806	1041	8448516	8450173	14.71206	893
chr3L	14189413	14191084	.	1000	.	27.05885	-1	-1	732	4.293953	-0.67805	14189413	14191072	11.97637	828	14189448	14191084	15.08248	690
chr3L	21651150	21654525	.	1000	.	26.94976	-1	-1	1844	4.196472	-0.66974	21651150	21654362	15.61743	947	21651372	21654525	11.33233	1521
chr3L	11039907	11042142	.	1000	.	26.92929	-1	-1	1322	4.663579	-0.69897	11040226	11042142	14.11266	792	11039907	11042122	12.81663	998
chrX	9686476	9689181	.	1000	.	26.69355	-1	-1	1442	4.464304	-0.692749	9686502	9688955	12.21750	926	9686476	9689181	14.47605	1181
chrX	17719150	17721145	.	1000	.	26.67480	-1	-1	1292	3.980818	-0.649385	17719164	17721145	15.65597	686	17719150	17721116	11.01883	714
chrX	17471959	17473611	.	1000	.	26.64542	-1	-1	709	4.288823	-0.677284	17472257	17473611	11.84180	450	17471959	17473604	14.80362	728
chrX	14909831	14911074	.	1000	.	26.64321	-1	-1	680	4.637720	-0.69897	14909832	14911072	13.84248	672	14909831	14911074	12.80073	688
chr2R	9809768	9811864	.	1000	.	26.60200	-1	-1	1140	4.603965	-0.69897	9809873	9811843	12.63686	1327	9809768	9811864	13.96514	1160
chr3L	15047310	15048935	.	1000	.	26.59791	-1	-1	1023	4.618174	-0.69897	15047484	15048921	13.92640	666	15047310	15048935	12.67151	735
chrX	5152536	5153845	.	1000	.	26.56357	-1	-1	546	4.636757	-0.69897	5152563	5153716	13.75763	462	5152536	5153845	12.80594	475
chrX	17175160	17177311	.	1000	.	26.44176	-1	-1	897	4.425244	-0.686152	17175172	17177311	14.73233	1297	17175160	17177293	11.70943	1058
chrX	8251159	8252987	.	1000	.	26.38585	-1	-1	730	4.510555	-0.696494	8251166	8252905	14.44273	1181	8251159	8252987	11.94312	873
significant_reproducible_peaks.bed
```

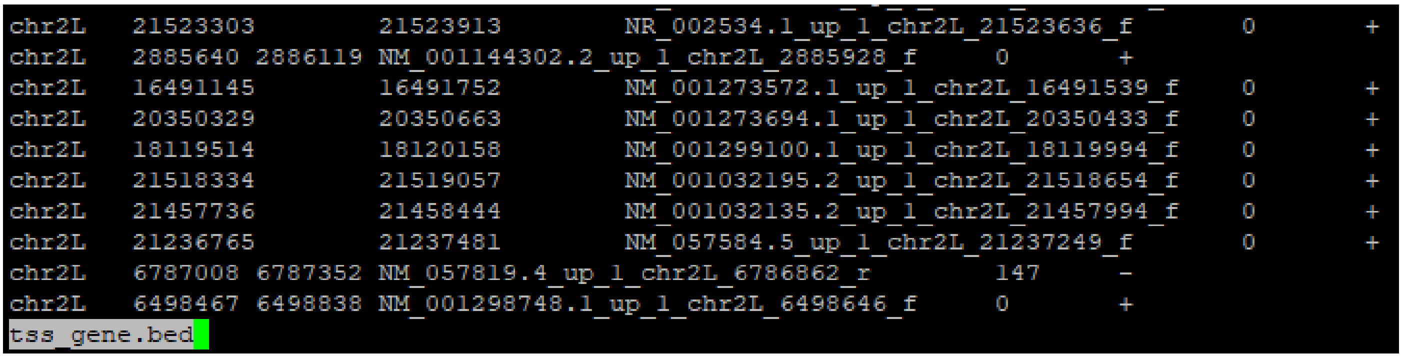
Given these locations a genome annotation file recording gene locations (dm6_tss.bed), and a getclosestgene.py script, I can identify a gene which has a CTCF binding site in its promoter region.

```
$ python binfpy/getclosestgene.py home/binf6_03/significant_reproducible_peaks.bed dm6_tss.bed
```

This will output a bed file `tss_gene.bed` detailing:

```
<CTCF bind chrom> <CTCF bind start> <CTCF bind end> <Gene name> <distance between CTCF and gene> <strand>
```

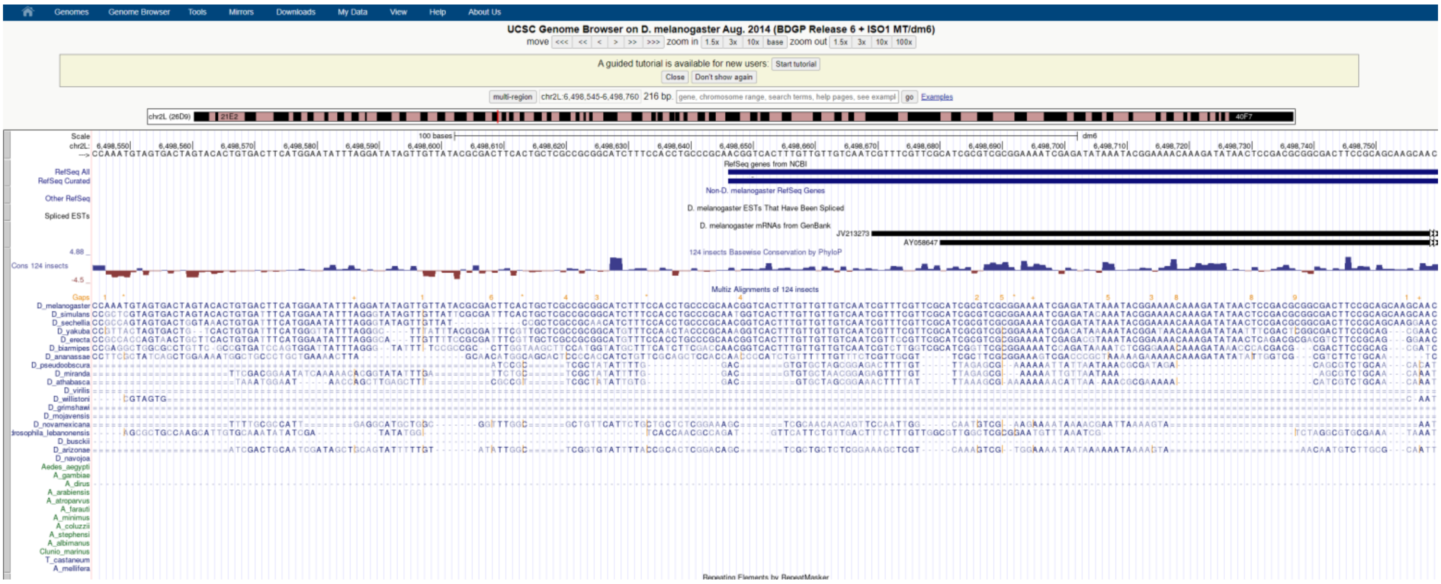
```
less tss_gene.bed
```



```
chr2L	21523303	21523913	NR_002534.1_up_1_chr2L_21523636_f	0	+
chr2L	2885640	2886119	NM_001144302.2_up_1_chr2L_2885928_f	0	+
chr2L	16491145	16491752	NM_001273572.1_up_1_chr2L_16491539_f	0	+
chr2L	20350329	20350663	NM_001273694.1_up_1_chr2L_20350433_f	0	+
chr2L	18119514	18120158	NM_001299100.1_up_1_chr2L_18119994_f	0	+
chr2L	21518334	21519057	NM_001032195.2_up_1_chr2L_21518654_f	0	+
chr2L	21457736	21458444	NM_001032135.2_up_1_chr2L_21457994_f	0	+
chr2L	21236765	21237481	NM_057584.5_up_1_chr2L_21237249_f	0	+
chr2L	6787008	6787352	NM_057819.4_up_1_chr2L_6786862_r	147	-
chr2L	6498467	6498838	NM_001298748.1_up_1_chr2L_6498646_f	0	+
tss_gene.bed
```

chr2L is the chromosome where the binding site is located. "chr2L" refers to the left arm of the second chromosome in Drosophila melanogaster. 6498467 is the start position of the CTCF binding site on the chromosome. BED file coordinates are 0-based, meaning that the first base of the chromosome is considered position 0. Thus, this binding site starts at the 6,498,467th base of chr2L. 6498838 is the end position of the CTCF binding site on the chromosome. In BED format, the end position is exclusive, meaning the actual binding site extends up to but does not include this position. Therefore, the binding site spans from base 6,498,467 to base 6,498,837, making it 371 bases long. NM_001298748.1_up_1_chr2L_6498646_f is a unique identifier for the binding site or the peak. It includes the gene name with which the site is associated, in this case, "NM_001298748.1", which could be a gene identifier in a specific database. The additional details (up_1_chr2L_6498646_f) provide context about the binding site's location, such as it being upstream of the gene, its chromosome, a specific base position, and the direction ("f" for forward strand). The “0” represents the score of the peak or binding site, which can indicate the strength or confidence in the site's identification. A score of "0" might suggest a default value in this context. The "+" means that the binding site is on the forward strand, which has implications for the directionality of any genes or regulatory elements associated with this site.

Discovering a CTCF binding site located on the left arm of chromosome 2L, at the precise coordinates of 6,498,467 to 6,498,837, unveils a fascinating glimpse into the intricate regulatory networks within Drosophila melanogaster. This sequence lies in proximity to the gene tagged as NM_001298748.1. Given CTCF's renowned role as an architectural protein, shaping the 3D organization of chromatin and dictating the rhythm of gene expression, this association is more than mere coincidence. It suggests a targeted regulatory influence, where CTCF could be modulating the expression of NM_001298748.1, thereby impacting fundamental biological processes from development and cell differentiation to the safeguarding of chromosomal architecture. This discovery not only highlights the complexity and precision of genetic regulation in D. melanogaster but also opens up avenues for exploring how such regulatory mechanisms contribute to the organism's biology and evolution.



UCSC Genome Browser on D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)

A guided tutorial is available for new users:

chr2L:6,498,545-6,498,760 216 bp.

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
Start tutorial
Close
Don't show again
multi-region
gene, chromosome range, search terms, help pages, see exampl
go
Examples
chr2L (26D9)
21E2
40F7
Scale
100 bases
dm6
chr2L: 6,498,550 6,498,560 6,498,570 6,498,580 6,498,590 6,498,600 6,498,610 6,498,620 6,498,630 6,498,640 6,498,650 6,498,660 6,498,670 6,498,680 6,498,690 6,498,700 6,498,710 6,498,720 6,498,730 6,498,740 6,498,750
CCAAATGTAGTGACTAGTACACTGTGACTTCATGGAATATTTAGGATATAGTTGTTATACGCGACTTCACTGCTCGCCGCGGCATCTTTCCACCTGCCCGCAACGGTCACTTTGTTGTTGTCAATCGTTTCGTTCGCATCGCGTCGCGGAAAATCGAGATATAAATACGGAAAACAAAGATATAACTCCGACGCGGCGACTTCCGCAGCAAGCAAC
RefSeq genes from NCBI
RefSeq All
RefSeq Curated
Other RefSeq
Non-D. melanogaster RefSeq Genes
Spliced ESTs
D. melanogaster ESTs That Have Been Spliced
D. melanogaster mRNAs from GenBank
JV213273
AY058647
4.88
Cons 124 insects
124 insects Basewise Conservation by PhyloP
-4.5
Multiz Alignments of 124 insects
Gaps
D_melanogaster
D_simulans
D_sechellia
D_yakuba
D_erecta
D_biarmipes
D_ananassae
D_pseudoobscura
D_miranda
D_athabasca
D_virilis
D_willistoni
D_grimshawi
D_mojavensis
D_novamexicana
Drosophila_lebanonensis
D_busckii
D_arizonae
D_navojoa
Aedes_aegypti
A_gambiae
A_dirus
A_arabiensis
A_atroparvus
A_farauti
A_minimus
A_coluzzii
A_stephensi
A_albimanus
Clunio_marinus
T_castaneum
A_mellifera
Repeating Elements by RepeatMasker

```
#!/bin/bash

FULL_OUTPUT_DIR="full_output"

# Step 1: Quality Control

mkdir -p ${FULL_OUTPUT_DIR}/metrics

fastqc data/*.fastq.gz -o ${FULL_OUTPUT_DIR}/metrics/

# Step 2: Alignment with Bowtie2

mkdir -p ${FULL_OUTPUT_DIR}/alignment.

bowtie2 -x reference/bowtie2_index/genome \

        -U data/control.fastq.gz \ # Specifies the path to the FASTQ file for the control sample.

        > ${FULL_OUTPUT_DIR}/alignment/control.sam

bowtie2 -x reference/bowtie2_index/genome \ # Specifies the path to the Bowtie2 index.

        -U data/treatment_rep1.fastq.gz \

        > ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.sam

bowtie2 -x reference/bowtie2_index/genome \

        -U data/treatment_rep2.fastq.gz \ # Specifies the path to the FASTQ file for treatment replicate 2.

        > ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.sam

# Step 3: SAM to BAM Conversion and Mapping Quality Check

samtools view -bS ${FULL_OUTPUT_DIR}/alignment/control.sam \

    > ${FULL_OUTPUT_DIR}/alignment/control.bam

samtools view -bS ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.sam \

    > ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.bam

samtools view -bS ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.sam \

    > ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.bam

samtools flagstat ${FULL_OUTPUT_DIR}/alignment/control.bam \

    > ${FULL_OUTPUT_DIR}/metrics/control_mapping_quality.txt

samtools flagstat ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.bam \

    > ${FULL_OUTPUT_DIR}/metrics/treatment_rep1_mapping_quality.txt

samtools flagstat ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.bam \

    > ${FULL_OUTPUT_DIR}/metrics/treatment_rep2_mapping_quality.txt

# Step 4: Peak Calling with MACS2

mkdir -p ${FULL_OUTPUT_DIR}/peaks

macs2 callpeak -t ${FULL_OUTPUT_DIR}/alignment/treatment_rep1.bam \ # Specify the treatment replicate 1 BAM file as the target.

               -c ${FULL_OUTPUT_DIR}/alignment/control.bam \

               -f BAM -g dm -n treatment_rep1 \

               --outdir ${FULL_OUTPUT_DIR}/peaks -q 0.05

               --call-summits --nomodel --extsize 147

macs2 callpeak -t ${FULL_OUTPUT_DIR}/alignment/treatment_rep2.bam \
        -c ${FULL_OUTPUT_DIR}/alignment/control.bam \
        -f BAM -g dm -n treatment_rep2 \
        --outdir ${FULL_OUTPUT_DIR}/peaks -q 0.05 \
        --call-summits --nomodel --extsize 147
# Step 5: Reproducibility Assessment with IDR
mkdir -p ${FULL_OUTPUT_DIR}/idr # Create a directory for storing idr results.
idr --samples ${FULL_OUTPUT_DIR}/peaks/treatment_rep1_peaks.narrowPeak \
     ${FULL_OUTPUT_DIR}/peaks/treatment_rep2_peaks.narrowPeak \
  --output-file ${FULL_OUTPUT_DIR}/idr/treatment_idr_output.txt \
  --log-output-file ${FULL_OUTPUT_DIR}/idr/treatment_idr.log
awk 'BEGIN {OFS="\t"} {$12 = -log($12)/log(10); print}' ${FULL_OUTPUT_DIR}/idr/treatment_idr_output.txt >
${FULL_OUTPUT_DIR}/idr/treatment_idr_output_log10p>
# Step 6: Filtering Peaks Based on IDR Output
mkdir -p ${FULL_OUTPUT_DIR}/final_peaks
awk '$12 < 0.05' ${FULL_OUTPUT_DIR}/idr/treatment_idr_output.txt \
  > ${FULL_OUTPUT_DIR}/final_peaks/significant_reproducible_peaks.bed
```

SUBSET

```
#!/bin/bash
# Step 0: Create subsets of the original fastq files for testing purposes to speed up the process.
echo "Subsetting data for testing..."
mkdir -p sub_data # Create a directory for the subsetted data.
# Use seqtk to sample 1000 reads from the original fastq.gz files, then gzip the output for FastQC compatibility.
seqtk sample -s100 data/control.fastq.gz 10000 | gzip > sub_data/sub_control.fastq.gz
seqtk sample -s100 data/treatment_rep1.fastq.gz 10000 | gzip > sub_data/sub_treatment_rep1.fastq.gz
seqtk sample -s100 data/treatment_rep2.fastq.gz 10000 | gzip > sub_data/sub_treatment_rep2.fastq.gz

# Step 1: Run FastQC for quality control on the subsetted fastq.gz files.
echo "Running FastQC for quality control..."
mkdir -p output/metrics # Create a directory for FastQC reports.
fastqc sub_data/*.fastq.gz -o output/metrics/ # Run FastQC on all subsetted fastq.gz files and output the reports to the metrics directory.

# Step 2: Align reads to the reference genome with Bowtie2.
mkdir -p output/alignment # Create a directory for the alignment output.
# Align each subsetted fastq.gz file to the reference genome using Bowtie2, outputting SAM files.
bowtie2 -x reference/bowtie2_index/genome -U sub_data/sub_control.fastq.gz > output/alignment/sub_control.sam
bowtie2 -x reference/bowtie2_index/genome -U sub_data/sub_treatment_rep1.fastq.gz > output/alignment/sub_treatment_rep1.sam
bowtie2 -x reference/bowtie2_index/genome -U sub_data/sub_treatment_rep2.fastq.gz > output/alignment/sub_treatment_rep2.sam

# Step 3: Convert SAM files to BAM format and check mapping quality.
echo "Converting SAM to BAM and checking mapping quality..."
# Convert SAM to BAM using samtools for each alignment file.
samtools view -bS output/alignment/sub_control.sam > output/alignment/sub_control.bam
samtools view -bS output/alignment/sub_treatment_rep1.sam > output/alignment/sub_treatment_rep1.bam
samtools view -bS output/alignment/sub_treatment_rep2.sam > output/alignment/sub_treatment_rep2.bam
# Generate mapping quality reports for each BAM file using samtools flagstat.
samtools flagstat output/alignment/sub_control.bam > output/metrics/sub_control_mapping_quality.txt
samtools flagstat output/alignment/sub_treatment_rep1.bam > output/metrics/sub_treatment_rep1_mapping_quality.txt
samtools flagstat output/alignment/sub_treatment_rep2.bam > output/metrics/sub_treatment_rep2_mapping_quality.txt
```

```
# Step 4: Call peaks using MACS2 with a fixed extension size.
mkdir -p output/peaks # Create a directory for peak calling output.
# Use MACS2 to call peaks on each treatment BAM file against the control, specifying parameters like genome size, q-value cutoff, etc.
macs2 callpeak -t output/alignment/sub_treatment_rep1.bam -c output/alignment/sub_control.bam -f BAM -g dm -n sub_treatment_rep1 --outdir output/peaks -q 0.05 --call-summits --nomodel --extsize 147
macs2 callpeak -t output/alignment/sub_treatment_rep2.bam -c output/alignment/sub_control.bam -f BAM -g dm -n sub_treatment_rep2 --outdir output/peaks -q 0.05 --call-summits --nomodel --extsize 147


# Step 5: Use IDR (Irreproducible Discovery Rate) to assess the reproducibility between biological replicates' peak sets.
echo "Assessing reproducibility with IDR..."
idr --samples output/peaks/sub_treatment_rep1_peaks.narrowPeak \
    output/peaks/sub_treatment_rep2_peaks.narrowPeak \
    --output-file output/idr/sub_treatment_idr_output.txt \
    --plot \
    --log-output-file output/idr/sub_treatment_idr.log


# Step 6: Filter for significant, reproducible peaks based on IDR output.
echo "Filtering significant, reproducible peaks based on IDR output..."
# Use awk to filter the IDR output for entries with a column 12 value (adjusted p-value) less than 0.05, indicating significant reproducibility.
awk '$12 < 0.05' output/idr/sub_treatment_idr_output.txt > output/final_peaks/significant_reproducible_peaks.bed


echo "Pipeline execution for subsetted data completed."
```