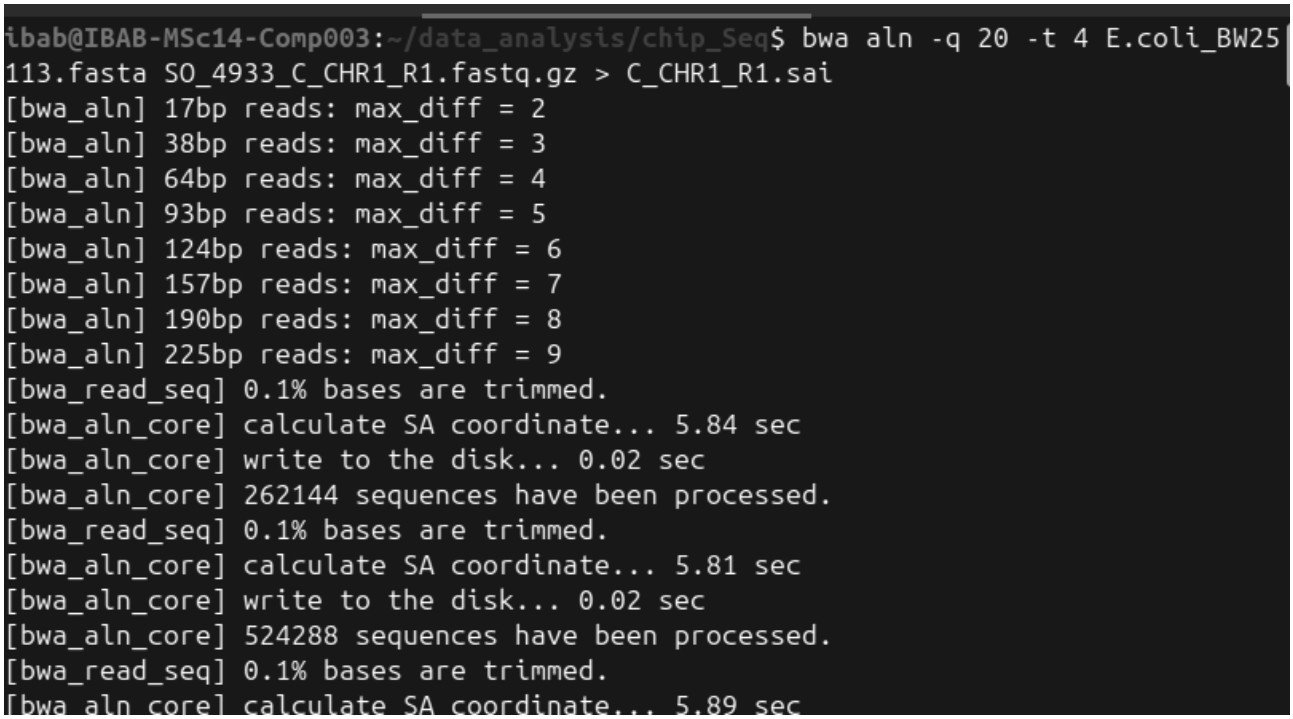
Chip seq

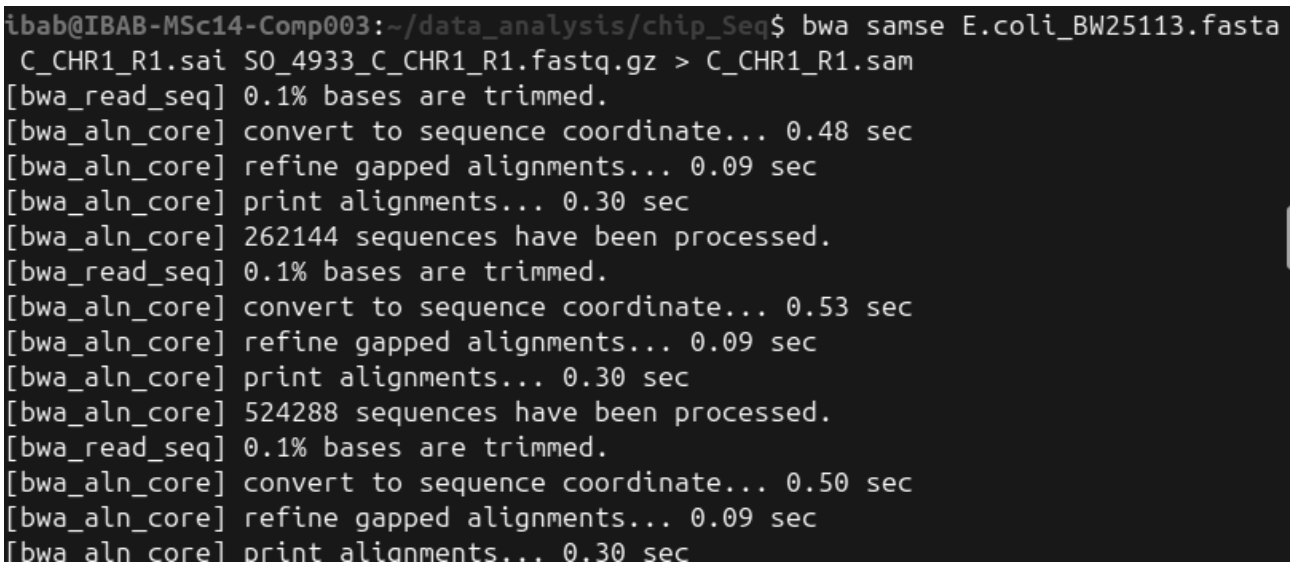
Tools requried

i. Mapping tools- Bowtie2 or BWA ii. Sam to bam conversion – samtools iii. MACS 2 – Peak caller iv. Homer v. Bedtools vi. Meme- motif analysis tools

Workflow

Mapping reads to the reference genome
#used bwa
#Indexing reference genome
bwa index E.coli_BW25113.fasta
#Alignment to the reference genome
bwa aln -q 20 -t 4 E.coli_BW25113.fasta C_CHR1_R1.fastq.gz > C_CHR1_R1.sai
bwa samse E.coli_BW25113.fasta C_CHR1_R1.sai C_CHR1_R1.fastq.gz > C_CHR1_R1.sam
#Converting sam to bam
samtools view -bSq 20 C_CHR1_R1.sam -o C_CHR1_R1.bam -@4
#Sorting the bam file
samtools sort C_CHR1_R1.bam -o C_CHR1_R1_sorted.bam -@4
#Indexing the sorted bam file
samtools index C_CHR1_R1_sorted.bam -@4






```
#Peakcalling using MACS2
macs2 callpeak -t C_CHR1_R1_sorted.bam -c C_INR1_R1_sorted.bam -f BAM -g 4.6e6 -n
C_R1_sorted -B -p 0.001 –nomodel
```

`macs2 callpeak`

This is the **peak-calling function** in MACS2 — it identifies genomic regions where the ChIP sample (treatment) has significantly higher read coverage than the input (control), representing potential **protein-DNA binding sites**.

`-t C_CHR1_R1_sorted.bam`

Treatment file (ChIP sample)
This is your **experimental ChIP-Seq BAM file** (after sorting and indexing).
MACS2 will look for enriched regions (peaks) in this sample.

`-c C_INR1_R1_sorted.bam`

Control or Input file
Represents background signal — DNA fragments not specifically bound by the protein of interest.
MACS2 uses this to estimate noise and correct for biases like open chromatin or GC content.

`-f BAM`

File format
Specifies that your input files (`-t` and `-c`) are in **BAM format** (binary alignment files).

`-g 4.6e6`

Genome size
Approximate number of **mappable bases** in your reference genome.

You used `4.6e6`, which is correct for *E. coli* (~4.6 Mb genome).
For humans, you'd typically use `2.7e9`, and for mouse `1.87e9`.

`-n C_R1_sorted`

Output name prefix
All result files will start with this prefix (e.g., `C_R1_sorted_peaks.xls`, `C_R1_sorted_summits.bed`, `C_R1_sorted_treat_pileup.bdg`).

`-B`

Generate signal tracks
Tells MACS2 to output **BedGraph files**:

- `_treat_pileup.bdg` → ChIP signal track
- `_control_lambda.bdg` → background model
  These can be visualized in genome browsers like IGV or UCSC.

`-p 0.001`

p-value cutoff for peak detection
Only regions with a **p-value < 0.001** are considered significant peaks.
(You can use `-q 0.05` instead for **FDR-based** cutoff if you prefer.)

`--nomodel`

Disables automatic model building
Normally MACS2 tries to infer fragment size from the data, but for bacterial or low-coverage ChIP-Seq (like *E. coli*), this can fail.
`--nomodel` prevents model building and uses raw read coverage directly.

```
sudo apt install macs
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ macs3 callpeak -t D_CHR1_R1_so
rted.bam -c D_INR1_R1_sorted.bam -f BAM -g 4.6e6 -n D_R1_sorted -B -p 0.001 --no
model
INFO  @ 15 Oct 2025 11:04:55: [39 MB]
# Command line: callpeak -t D_CHR1_R1_sorted.bam -c D_INR1_R1_sorted.bam -f BAM
-g 4.6e6 -n D_R1_sorted -B -p 0.001 --nomodel
# ARGUMENTS LIST:
# name = D_R1_sorted
# format = BAM
# ChIP-seq file = ['D_CHR1_R1_sorted.bam']
# control file = ['D_INR1_R1_sorted.bam']
# effective genome size = 4.60e+06
# band width = 300
# model fold = [5, 50]
# pvalue cutoff = 1.00e-03
# qvalue will not be calculated and reported as -1 in the final output.
```

1. narrowpeak with the peak info
2. bedgraph file for visualization
3. summit for the info of the highest intensity region of the peak

#Peak annotation using Homer
annotatePeaks.pl C_R1_sorted_peaks.narrowPeak E.coli_BW25113.fasta -gff E.coli_BW25113_annotation.gff> C_R1_sorted_annotated_peaks.tsv

```
D_INR1_R1.sai                    E.coli_BW25113.fasta.pac
D_INR1_R1.sam                    E.coli_BW25113.fasta.sa
D_INR1_R1_sorted.bam             SO_4933_C_CHR1_R1.fastq.gz
D_INR1_R1_sorted.bam.bai         SO_4933_C_INR1_R1.fastq.gz
D_R1_sorted_annotated_peaks.tsv  SO_4933_D_CHR1_R1.fastq.gz
D_R1_sorted_control_lambda.bdg   SO_4933_D_INR1_R1.fastq.gz
D_R1_sorted_peaks.narrowPeak
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ cat D_R1_sorted_peaks.narrowPe
ak |wc -l
912
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ samtools faidx E.coli_BW25113.
fasta
cut -f1,2 E.coli_BW25113.fasta.fai > genome.chrom.sizes
.coli_BW25113_annotation.gff> C_R1_sorted_annotated_peaks.
```

```
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ annotatePeaks.pl D_R1_sorted_p
eaks.narrowPeak E.coli_BW25113.fasta -gff E.coli_BW25113_annotation.gff> D_R1_so
rted_annotated_peaks.tsv

        Using Custom Genome
        Peak file = D_R1_sorted_peaks.narrowPeak
        Genome = E.coli_BW25113.fasta
        Organism = unknown
        Custom annotation GFF file: E.coli_BW25113_annotation.gff (better to get
 GTF file)
        Peak/BED file conversion summary:
                BED/Header formatted lines: 912
                peakfile formatted lines: 0
                Duplicated Peak IDs: 0

        Peak File Statistics:
```

2145414	+	847	NA	Intergenic
1816167	+	846	NA	Intergenic
281199	+	846	NA	Intergenic
1541144	+	844	NA	Intergenic
1252307	+	843	NA	Intergenic
1578691	+	842	NA	Intergenic
2093487	+	838	NA	Intergenic
2051719	+	836	NA	TTS (ID=exon-IVL04_RS10130-1;Parent=rna-IVL04_RS10130;anticodon=(pos:com
2227770	+	835	NA	Intergenic
1526428	+	833	NA	Intergenic
1847617	+	832	NA	Intergenic
1668343	+	831	NA	Intergenic
2378236	+	828	NA	Intergenic
2198255	+	828	NA	Intergenic
1092385	+	828	NA	TTS (ID=exon-IVL04_RS05230-1;Parent=rna-IVL04_RS05230;Dbxref=RFAM:RF00
1807762	+	826	NA	Intergenic
2853874	+	825	NA	Intergenic
560412	+	823	NA	promoter-TSS (ID=exon-IVL04_RS02710-1;Parent=rna-IVL04_RS02710;anticodon=
1582396	+	823	NA	Intergenic
1337943	+	821	NA	Intergenic
675294	+	820	NA	Intergenic

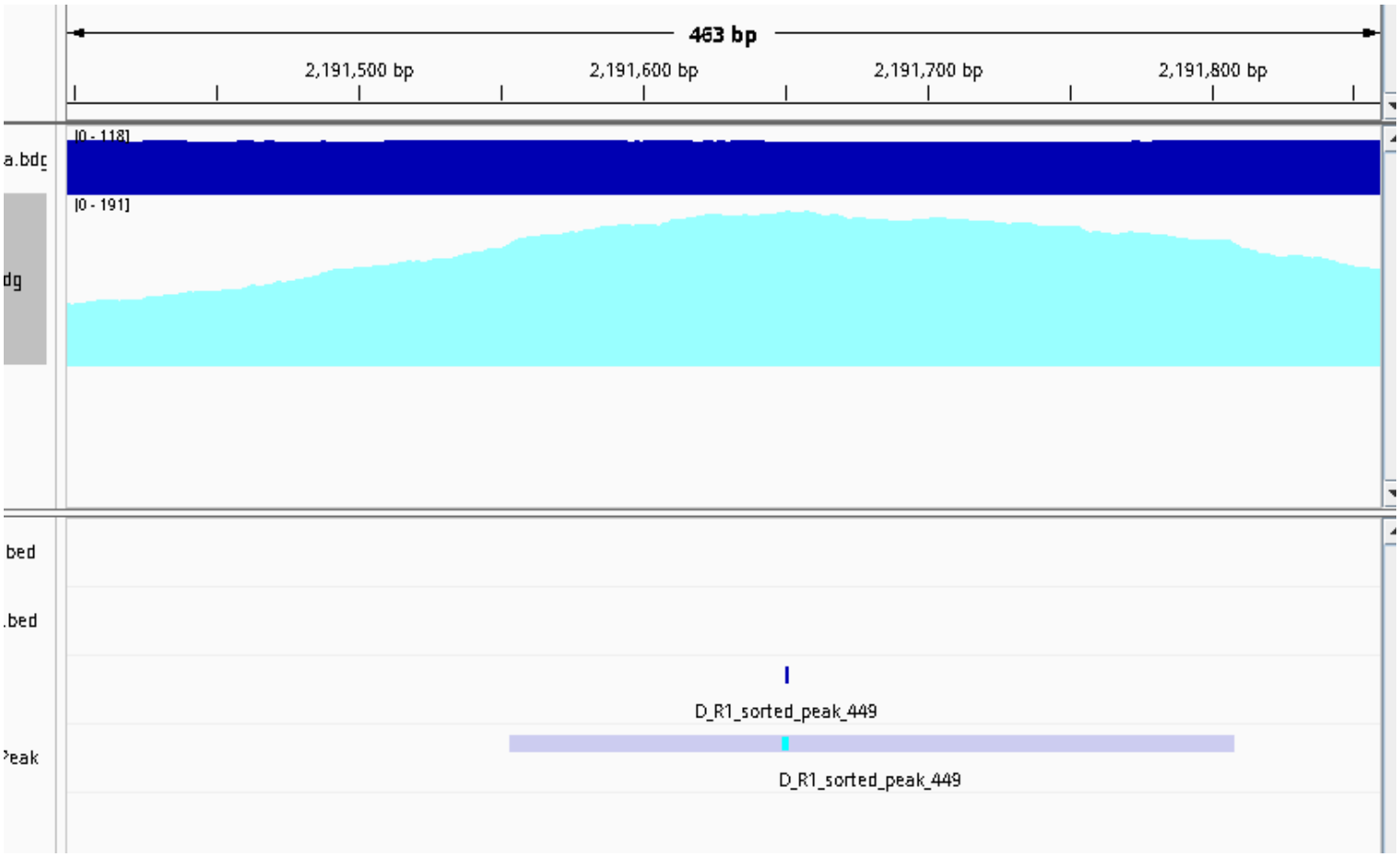
Merging the peaks

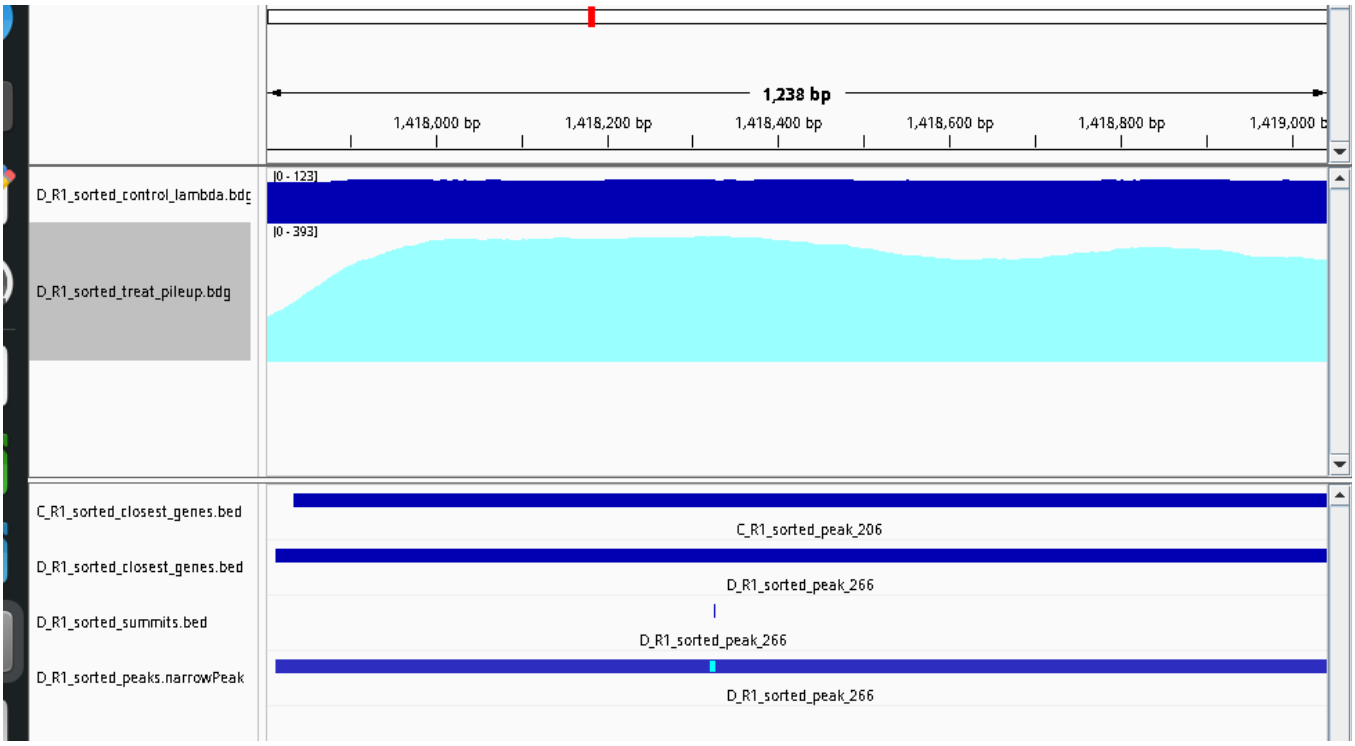
```
mergePeaks D_R1_sorted_peaks.narrowPeak>D_R1_sorted_peaks_merged.txt
```

Finding peaks nearer the features like gene

```
closestBed -a D_R1_sorted_peaks.narrowPeak -b E.coli_BW25113_annotation.bed
```

The results are viewed in IGV





Extract the sequemces from the reference file that corresponds to the narrow peak-calling

Using bed tools we can extract the sequences corresponding to peaks

bedtools getfasta -fi E.coli_BW25113.fasta -bed D_R1_sorted_peaks.narrowPeak -fo D_R1_sorted_peaks.fa

```
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ bedtools getfasta -fi E.coli_BW25113.fasta -
bed D_R1_sorted_peaks.narrowPeak -fo D_R1_sorted_peaks.fa
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ cat D_R1_sorted_peaks.fa
>NZ_CP064677.1:4953-5186
CTTGCTTTCACATAATCTGCCCGCCGATTTTGCTGCGTTGCGTAAATTGATGATGAATCATCAGTAAAATCTATTCATTATCTCAATCAGGCCG
GGTTTGCTTTTATGCAGCCCGGCTTTTTTATGAAGAAATTATGGAGAAAAATGACAGGGAAAAAGGAGAAATTCTCAATAAATGCGGTAACTTA
GAGATTAGGATTGCGGAGAATAACAACCGCCGTTCTCATCGAGTA
>NZ_CP064677.1:10395-10612
AGCGAAATAACCCACGTTGTGCAGGTTAAGCAGAATGGTGGTCATGCCGAAGCCCATCAGGCCCAGCGGTGCCGGATTAGCCAACTTAGTGTTG
CCCATAATTCCTCAAAAATCATCATCGAATGAATGGTGAAATAATTTCCCTGAATAACTGTAGTGTTTTCAGGGCGCGGCATAATAATCAGCCA
GTGGGGCAGTGTCTACGATCTTTTGAGGG
>NZ_CP064677.1:11093-11553
AATTTACCGTGTCCGCGCAGTTTGTTGGCGATACTATCGCCACCAAAATGCTGTAATTCTCCGGCAATCAGCTGCCAGTTGCGGCGATGTTGCT
CGGGATGCCCTTCCATCGATTTAAACAGTTCGTTGCGCATCAGTACGCTGGAGAGGCGAGTTTTGCCTTTTTCATTATGGGTGAGCAATCGGGC
GAAATTTGCCAACTGTTCCTCACTACAATGCTGAAGAAAATCCAGATCTGAATCATTCAGGTAATTAACATTCATTTTTTGTGGCTTCTATATT
CTGGCGTTAGTCGTCGCCGATAATTTTCAGCGTGGCCATATCCGATGAGTTCACCGTATGACCCGAAAAGGTGATTTTTGAGACGCAGCGTTTA
TTGTCGTTATCGCTGTTAATGTTGATCCAGTCAGTGGTTTGCCCTTCTTTTATTTCTGAAGGAATATTCAGGCTCTGACTGGCG
>NZ_CP064677.1:11871-12160
TTCTTTTTATTACAATTTTTTGATGAATGCCTTGGCTGCGATTCATTCTTTATATGAATAAAATTGCTGTCAATTTTACGTCTTGTCCTGCCAT
ATCGCGAAATTTCTGCGCAAAAGCACAAAAAATTTTTGCATCTCCCCCTTGATGACGTGGTTTACGACCCCATTTAGTAGTCAACCGCAGTGAG
TGAGTCTGCAAAAAAATGAAATTGGGCAGTTGAAACCAGACGTTTCGCCCCTATTACAGACTCACAACCACATGATGACCGAATATATAGTGGA
GACGTTT
>NZ_CP064677.1:16961-17609
TAGGAGGCCTCGGGTTGATGGTAAAATATCACTCGGGGCTTTTCTCTATCTGCCGTTCAGCTAATGCCTGAGACAGACAGCCTCAAGCACCCGC
CGCTATTATATCGCTCTCTTTAACCCATTTTGTTTTATCGATTCTAATCCTGAAGACGCCTCGCATTTTTGTGGCGTAATTTTTTAATGATTTA
ATTATTTAACTTTAATTTATCTCTTCATCGCAATTATTGACGACAAGCTGGATTATTTTTGAAATATTGGCCTAACAAGCATCGCCGACTGACA
```

```
CATTTTTTAACAACTCAACCGTTAGTACAGTCAGGAAATAGTTTAGCCTTTTTTAAGCTAAGTAAAGGGCTTTTTCTGCGACTTACGTTAAGAA
TTTGTAAATTCGCACCGCGTAATAAGTTGACAGTGATCACCCGGTTCGCGGTTATTTGATCAAGAAGAGT
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ cat D_R1_sorted_peaks.fa |wc -l
1824
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$
```

There are 1824 sequences

Motif analysis using Hommer

findMotifs.pl D_R1_sorted_peaks.fa fasta homer_motif_output/ -len 8,10,12 -norevopp
D_R1_sorted_peaks.fa → input sequences

fasta → input type

homer_motif_output/ → output directory

```
ibab@IBAB-MSc14-Comp003:~/data_analysis/chip_Seq$ findMotifs.pl D_R1_sorted_peaks.fa fasta hom
er_motif_output/ -len 8,10,12 -norevopp

Selected Options:
        Input file = D_R1_sorted_peaks.fa
        Promoter Set = fasta
        Output Directory = homer_motif_output/
        Motif length set at 8, 10, 12,
        Will not search the reverse strand

        !Warning - no background FASTA file specified (Highly recommended)
        !Your input sequences will be randomized to serve as a background instead.

        Found 912 sequences
        Using custom gene IDs for GO analysis
        Parsing FASTA format files...
        Found 912 sequences
        !! 1 of 912 contained bad nucleotide characters [not ACGTN], replaced with N
        Found 4560 sequences
        !! 5 of 4560 contained bad nucleotide characters [not ACGTN], replaced with N
```

-len 8,10,12 → motif lengths to search

-norevopp → only forward strand (optional)

HOMER outputs:

Known motifs enriched in your peaks

De novo motifs discovered

Motif logos and statistics

Phylogenetic analysis using MAFFT and FastTree

mafft alanine.fasta>align_seq_.fasta

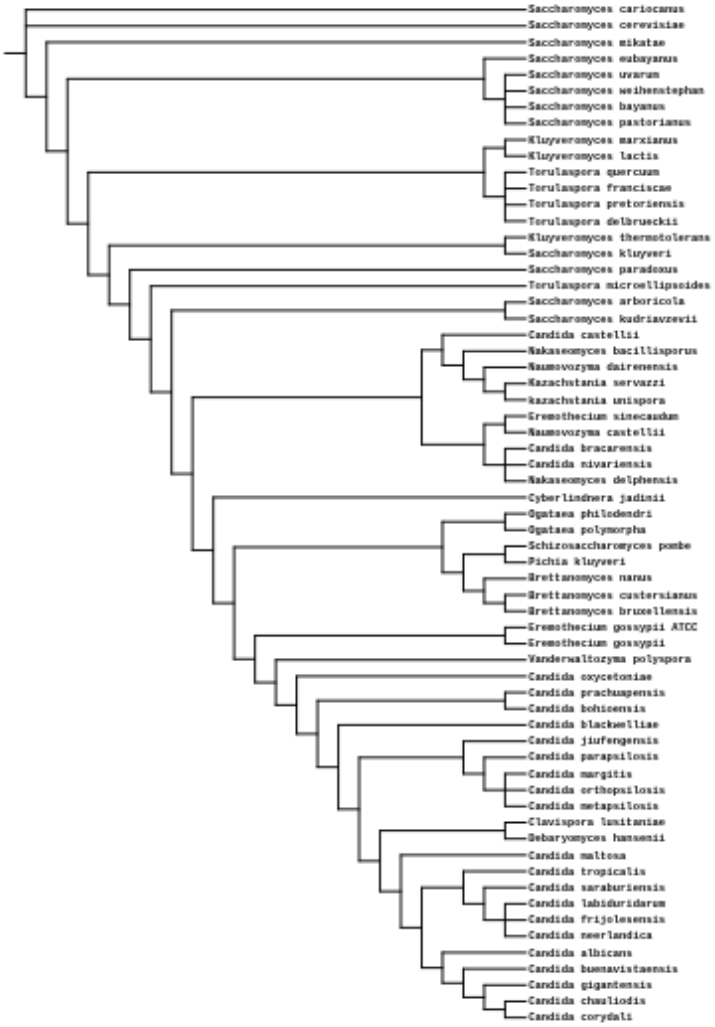
```
--dash :         Add structural information (Rozewicki et al, submitted)
ibab@IBAB-MSc14-Comp003:~/data_analysis/phylo$ mafft alanine.fasta>align_seq_.fa
sta
nthread = 0
nthreadpair = 0
nthreadtb = 0
ppenalty_ex = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00



Making a distance matrix ..
    1 / 63
done.
```

```
ibab@IBAB-MSc14-Comp003:~/data_analysis/phylo$ cat align_seq_.fasta
>Saccharomyces_cerevisiae
-ggggttatagttaaatttggtagaacgactgcgttgcatgcatttaatatgagttcaag
tctcattaact------
>Saccharomyces_weihenstephan
-ggggttatagttcaatttggtagaacgactgcgttgcatgcatttaatatgagttcaag
tctcattaactcca---
>Saccharomyces_paradoxus
-ggggttatagttaaatttggtagaacgattgcgttgcatgcatttaatatgagttcaag
tctcattaactccaata
>Saccharomyces_mikatae
-ggggttatagttaaatttggtagaacgactgcgttgcatgcatttaatatgagttcaag
tctcattaactccaata
>Saccharomyces_uvarum
-ggggttatagttcaatttggtagaacgactgcgttgcatgcatttaatatgagttcaag
tctcattaactcca---
>Saccharomyces_eubayanus
aggggttatagttcaatttggtagaacgactgcgttgcatgcatttaatatgagttcaag
tctcattaactccaata
>Saccharomyces_pastorianus
```

```
ibab@IBAB-MSc14-Comp003:~/data_analysis/phylo$ FastTree -nt align_seq_.fasta >ph
ytree.nwk
FastTree Version 2.1.11 Double precision (No SSE3)
Alignment: align_seq_.fasta
Nucleotide distances: Jukes-Cantor Joins: balanced Support: SH-like 1000
Search: Normal +NNI +SPR (2 rounds range 10) +ML-NNI opt-each=1
TopHits: 1.00*sqrtN close=default refresh=0.80
ML Model: Jukes-Cantor, CAT approximation with 20 rate categories
Initial topology in 0.00 seconds
Refining topology: 22 rounds ME-NNIs, 2 rounds ME-SPRs, 11 rounds ML-NNIs
Total branch-length 3.287 after 0.02 sec
ML-NNI round 1: LogLk = -1166.829 NNIs 9 max delta 3.02 Time 0.03
Switched to using 20 rate categories (CAT approximation)
Rate categories were divided by 0.840 so that average rate = 1.0
CAT-based log-likelihoods may not be comparable across runs
Use -gamma for approximate but comparable Gamma(20) log-likelihoods
ML-NNI round 2: LogLk = -1061.820 NNIs 3 max delta 0.42 Time 0.05
ML-NNI round 3: LogLk = -1061.739 NNIs 1 max delta 0.00 Time 0.05
Turning off heuristics for final round of ML NNIs (converged)
```



Synteny using mummer

nucmer --prefix=alignment alanine.fasta glycine.fasta

```
Try '/usr/bin/nucmer -h' for more information.
ibab@IBAB-MSc14-Comp003:~/data_analysis/phylo$ nucmer --prefix=alignment alanine
.fasta glycine.fasta
1: PREPARING DATA
2,3: RUNNING mummer AND CREATING CLUSTERS
# reading input file "alignment.ntref" of length 4621
# construct suffix tree for sequence of length 4621
# (maximum reference length is 536870908)
# (maximum query length is 4294967295)
# CONSTRUCTIONTIME /usr/bin/mummer alignment.ntref 0.00
# reading input file "/home/ibab/data_analysis/phylo/glycine.fasta" of length 45
25
# matching query-file "/home/ibab/data_analysis/phylo/glycine.fasta"
# against subject-file "alignment.ntref"
# COMPLETETIME /usr/bin/mummer alignment.ntref 0.00
# SPACE /usr/bin/mummer alignment.ntref 0.01
4: FINISHING DATA
```

Synten plot was not able to generate the time.