ChIP-Seq Analysis

07.02.19 | Shane Crinion |13326096

Chromatin immunoprecipitation (ChIP-Seq) assays are used to analyse DNA-protein interactions. ChIP-Seq combines next generation sequencing (NGS) with chromatin immunoprecipitation to identify genome-wide DNA binding sites for transcription factors (TFs). TFs control the level of transcription of its target genes by binding to their matching motif. In cancer research, ChIP can generate a library of target DNA sites that bound to transcription factors that may be deregulated in cancer cells.

Breast cancer cells lines can be used to indicate potential targets for malignancy treatment and identifying potential drug targets in oncology. ChIP-Seq analysis and biocomputational methods are used below by the Irish Cancer Society BreastPredict project to extract a human chromosome 21 TF and obtain information including genome-wide binding locations, regulatory role information and binding motif identification. The analysis uses two ChIP-Seq files named *chip.fa* and *input.fa* which contain test and control data respectively and will be referred to accordingly. The test data contains data which has been chromatin immunoprecipitated prior to sequencing while the control has not.

Data Preparation

Sequencing data is contained on the university cluster and can be accessed using `ssh`.

1. Log onto your cluster

Command:

```
# use ssh <username>@<hostname>
ssh scrinion@smgate.nuigalway.ie
ssh syd.nuigalway.ie
```

Log onto the cluster using your `ssh` to access FASTA (.fa) files.

2. Create a new directory

Command:

```
mkdir /data4/scrinion/chip-assignment
```

Create a new directory using `mkdir` for files generated during analysis.

3. Copy FASTA files to your new directory

Command:

```
cp chip.fa input.fa /data4/scrinion/chip-assignment
```

4. Transfer your data using `sftp`.

```
# transfer from cluster to TCP port
sftp scrinion@smgate.nuigalway.ie
put chip.fa
put input.fa

# transfer to TCP port to local machine
sftp <username>@smgate.nuigalway.ie
get chip.fasta
get input.fasta
```

`sftp` is used to transfer files securely across machines. This is used to transfer files that are generated from the analysis to the local machine to inspect results in the HTML files.

Initial QC

Quality control was performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), a bioinformatics tools used for quality control on raw sequence data. FASTQC is used to obtain an overview of sequence data using summary graphs and tables. If any area may be problematic, it is highlighted for attention in the report.

1. Load module

Command:

```
module load FASTQC
```

2. Run QC report

Command:

```
fastqc chip.fa
fastqc input.fa
```

QC is performed on ChIP-Seq data and the control (input data). The control is analysed to ensure high quality reference data and subsequently high experiment quality. The QC report can be inspected using the generated HTML file.

Test Data

The following areas are analysed from the ChIP-Seq test data:

Basic statistics (table 1), per base sequence quality (figure 1), per tile sequence quality, per sequence quality score, per base sequence content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, adaptor content and kmer content.

Of these 12 areas analysed, 3 were highlighted for attention: per tile sequence quality, per sequence quality score and kmer content.

Measure	Value
Filename	chip.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	295896
Sequences flagged as poor quality	0
Sequence length	37
%GC	48

Table 1: Summary statistics. This generates basic statistical data of the file. More information can be found at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/1%20Basic%20Statistics.html

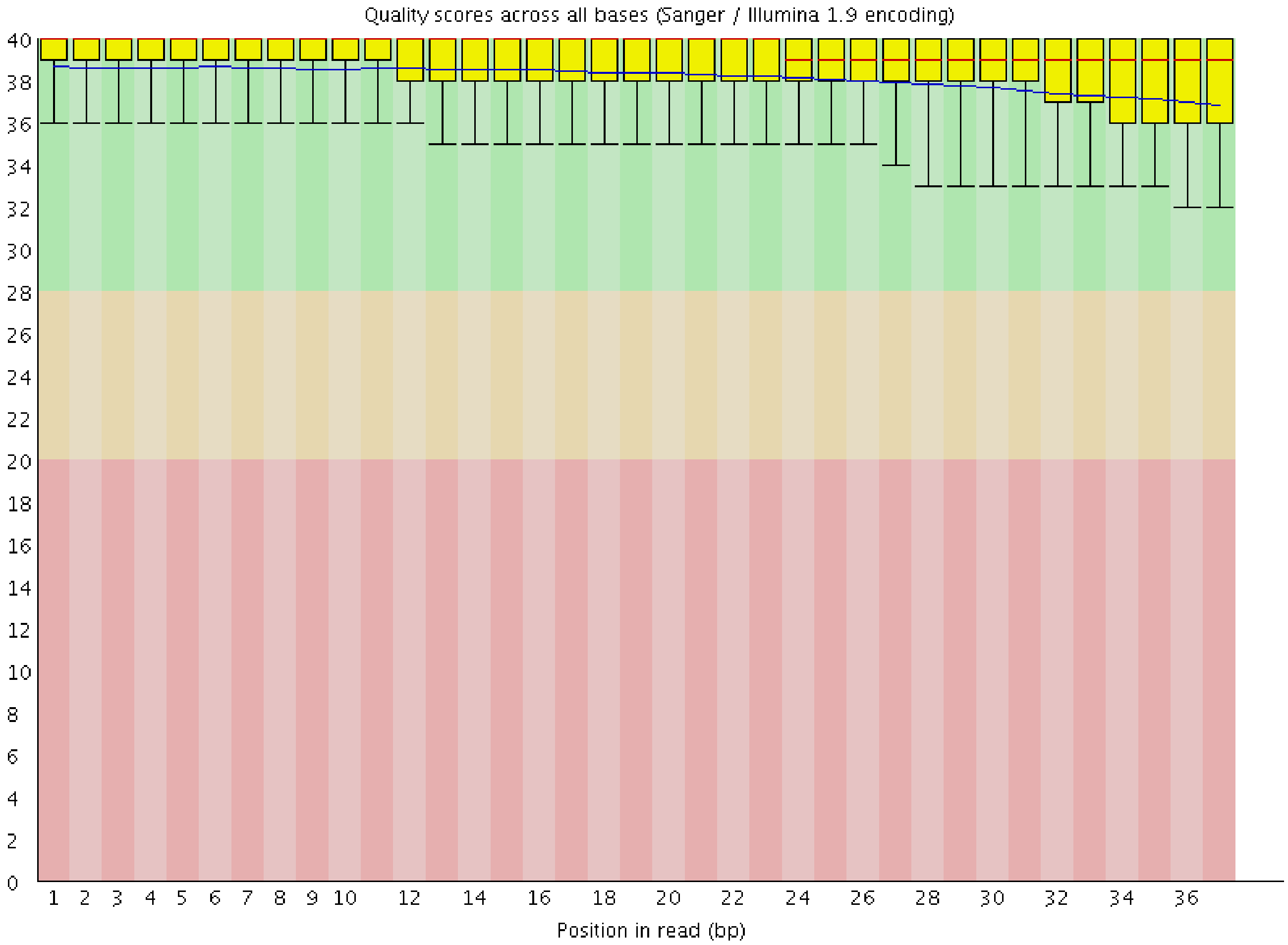


Figure 1: Per base sequencing quality. The BoxWhisker plot above indicates that high quality sequencing occurs throughout position in read base pair. The red line indicates median. The yellow box indicates the interquartile range. The upper and lower whiskers represent the 10% and 90% points. The blue line represents the mean quality. Quality can decrease towards orange at read ends however all reads remain of very high quality (indicated as green region). Phred quality score is above 35 (y axis) in all cases indicating 99.9% base accuracy in all instances. The heading indicates that Sanger / Illumina 1.9 encoding was used for quality score. No warnings/failure is raised. More information can be found at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

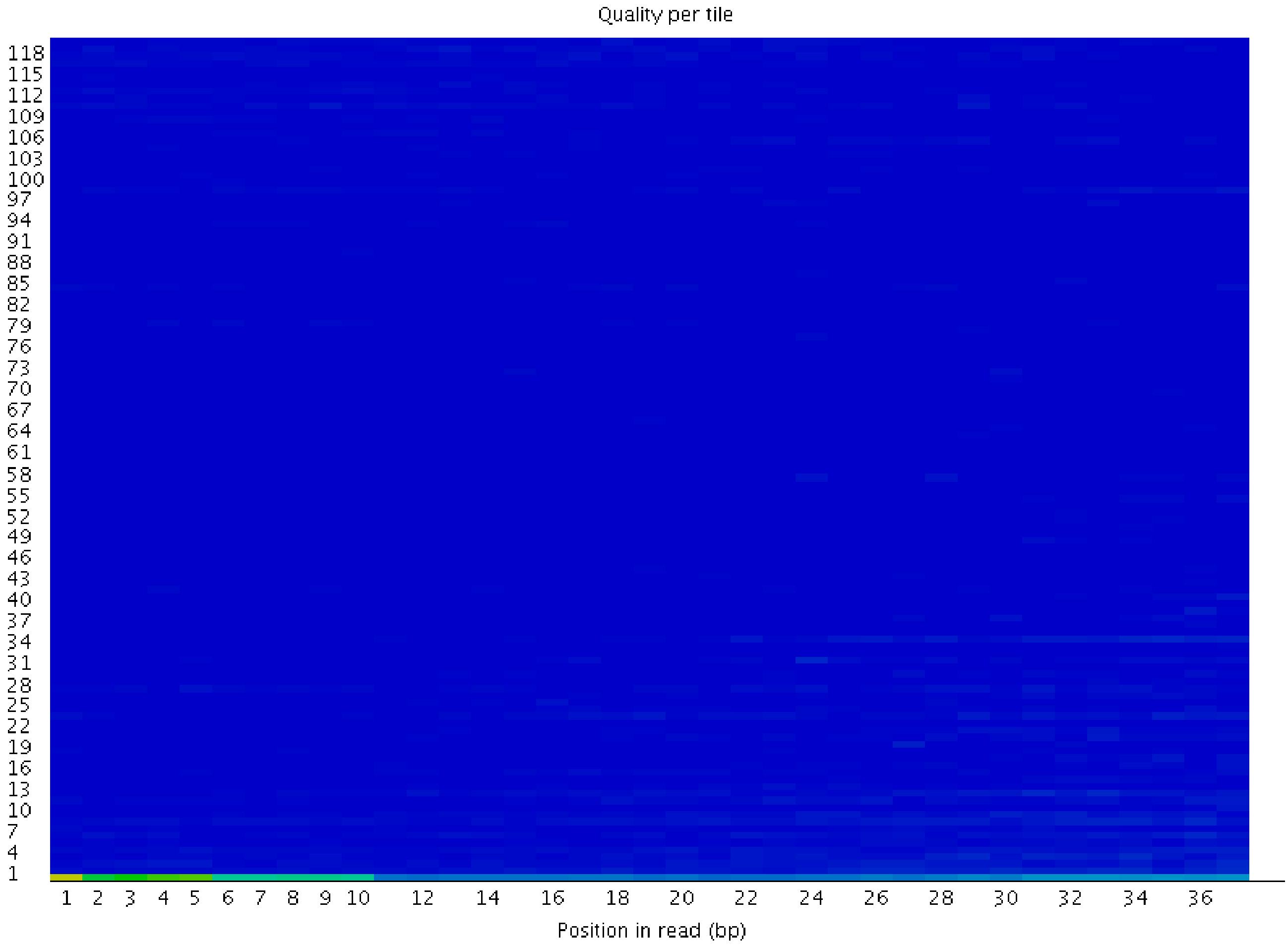


Figure 2: Per tile sequencing quality. This is generated if using Illumina libraries with the sequence identifiers. This can be generated used sequence identifiers. This indicates the loss of quality across each part of the flowcell and the deviation of quality for each score. A good analysis has consistent blue all over. This example returned a warning which indicates Phred score < 2 and less than mean for the bases accross the tile. Considering all tiles have the same characteristics with no variation, this warning can be ignored.

More information can be found at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html

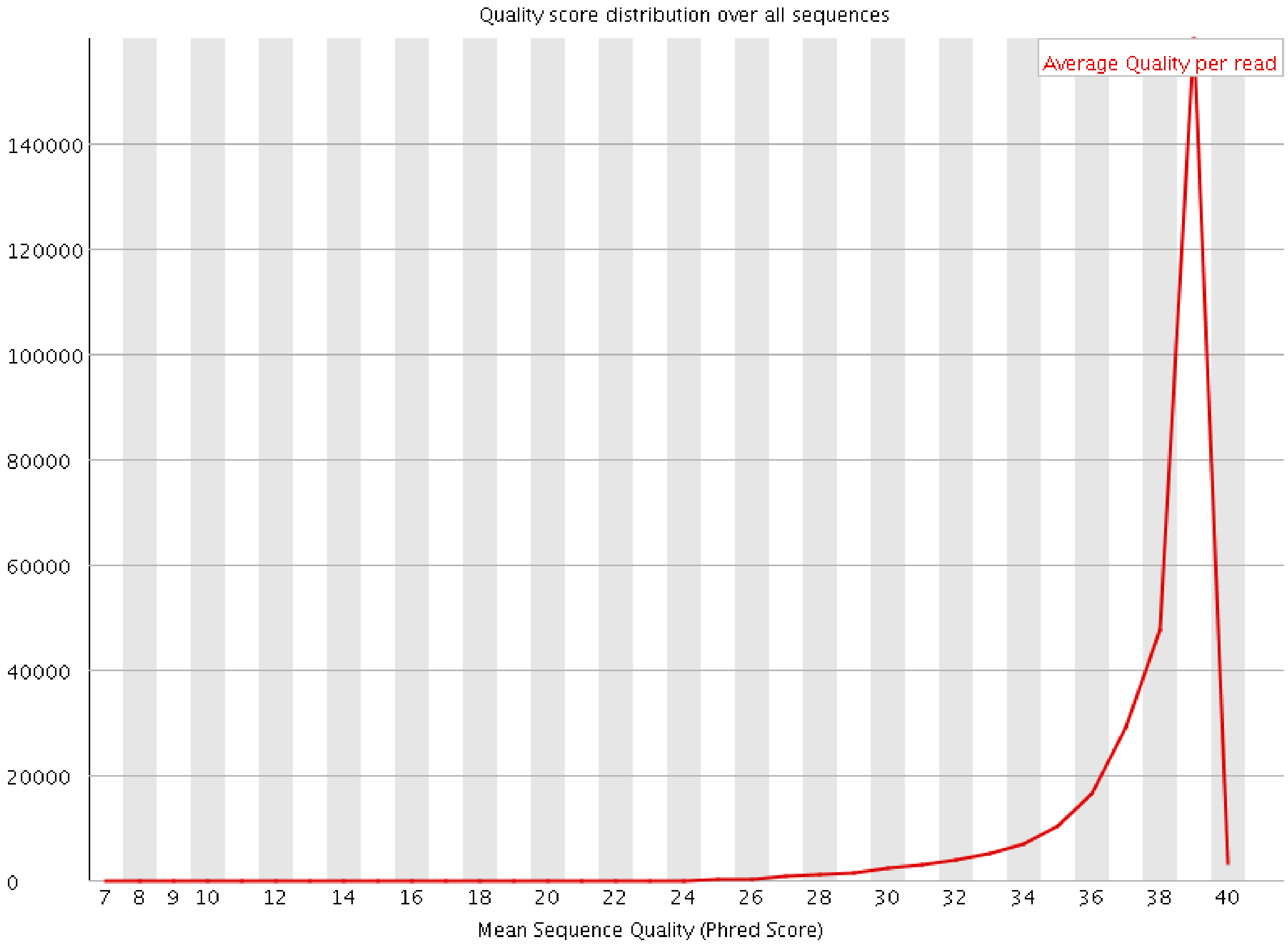


Figure 3: Per sequence quality scores. This indicates if a portion of the sample will have universally low quality values. These can indicate if there is a significant error rate in the sample. The mean quality is very high for the data and indicates no part of the sequence has universal poor quality. No warnings/failure occur which indicates error rate < 0.2%.

More information can be found at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html

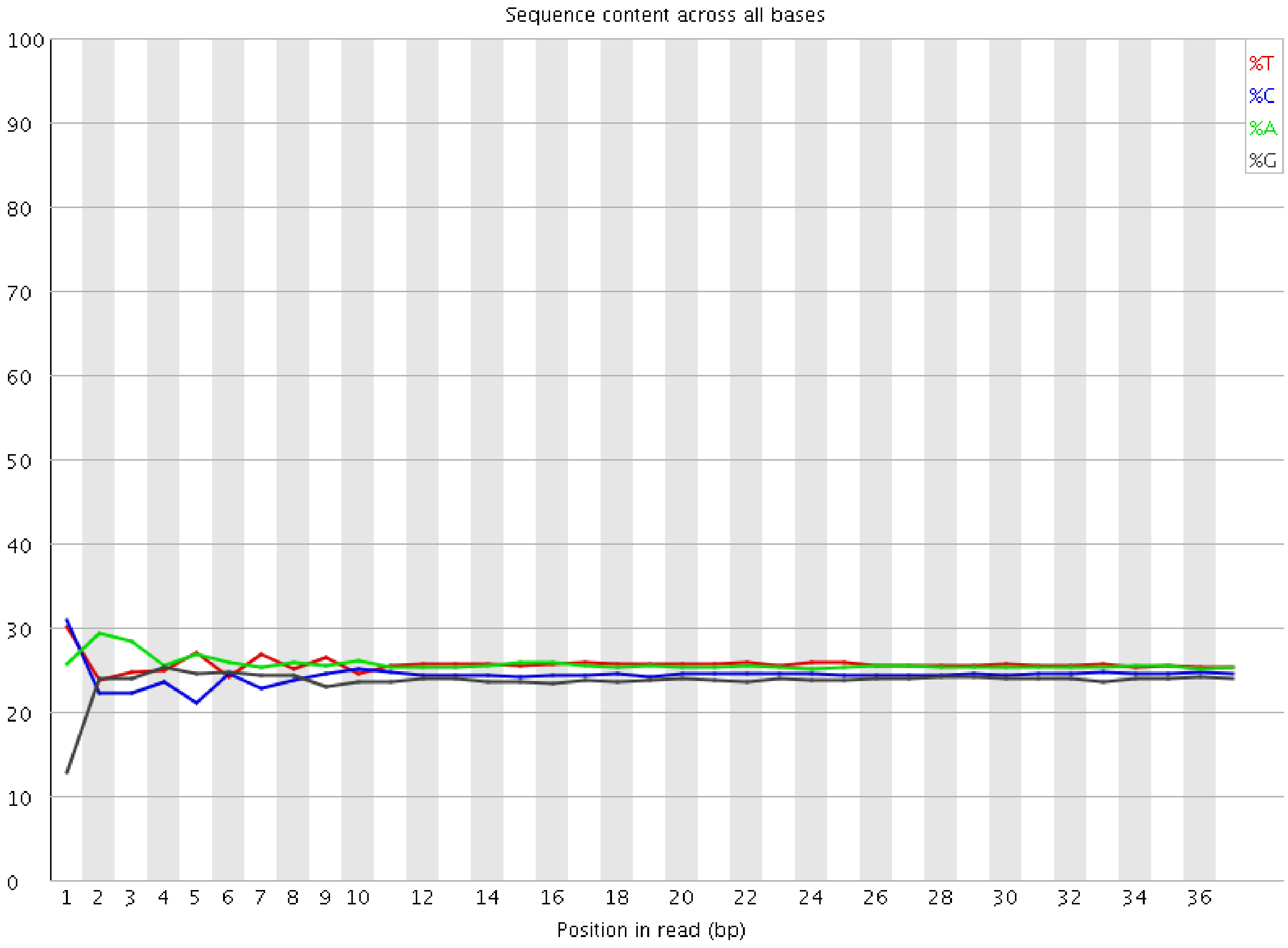


Figure 4: Per base sequence content. The library of reads should have a minimum difference in content between base sequences. The bases in the graph should represent the amount of bases in the genome and should a similar amount of each base. Warning occurred which indicates < 10% in base difference at some point of the genome. This is likely due to intrinsic variation during early position in read. This can be ignored as the long end indicates no difference in sequence content across all bases. More information can be found at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html.

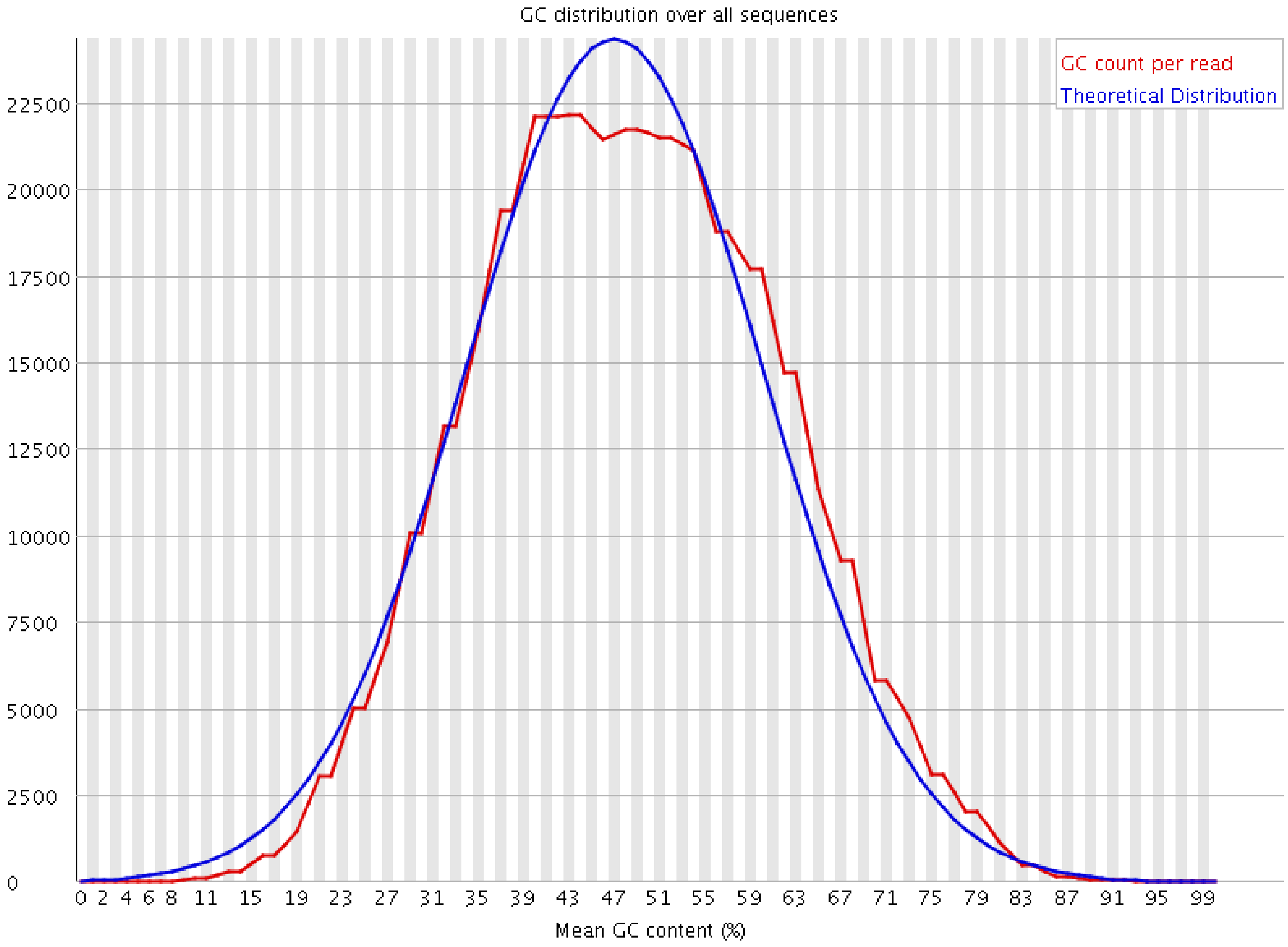


Figure 5: Per sequence GC content. This models a reference GC count across the genome against the actual GC count. The central point corresponds to the overall GC content. No warnings/failure occurred which indicates a deviation from the normal distribution. No warning/failure occured which indicates the GC count deviated < 15% from the model throughout the genome. More information can be found at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html

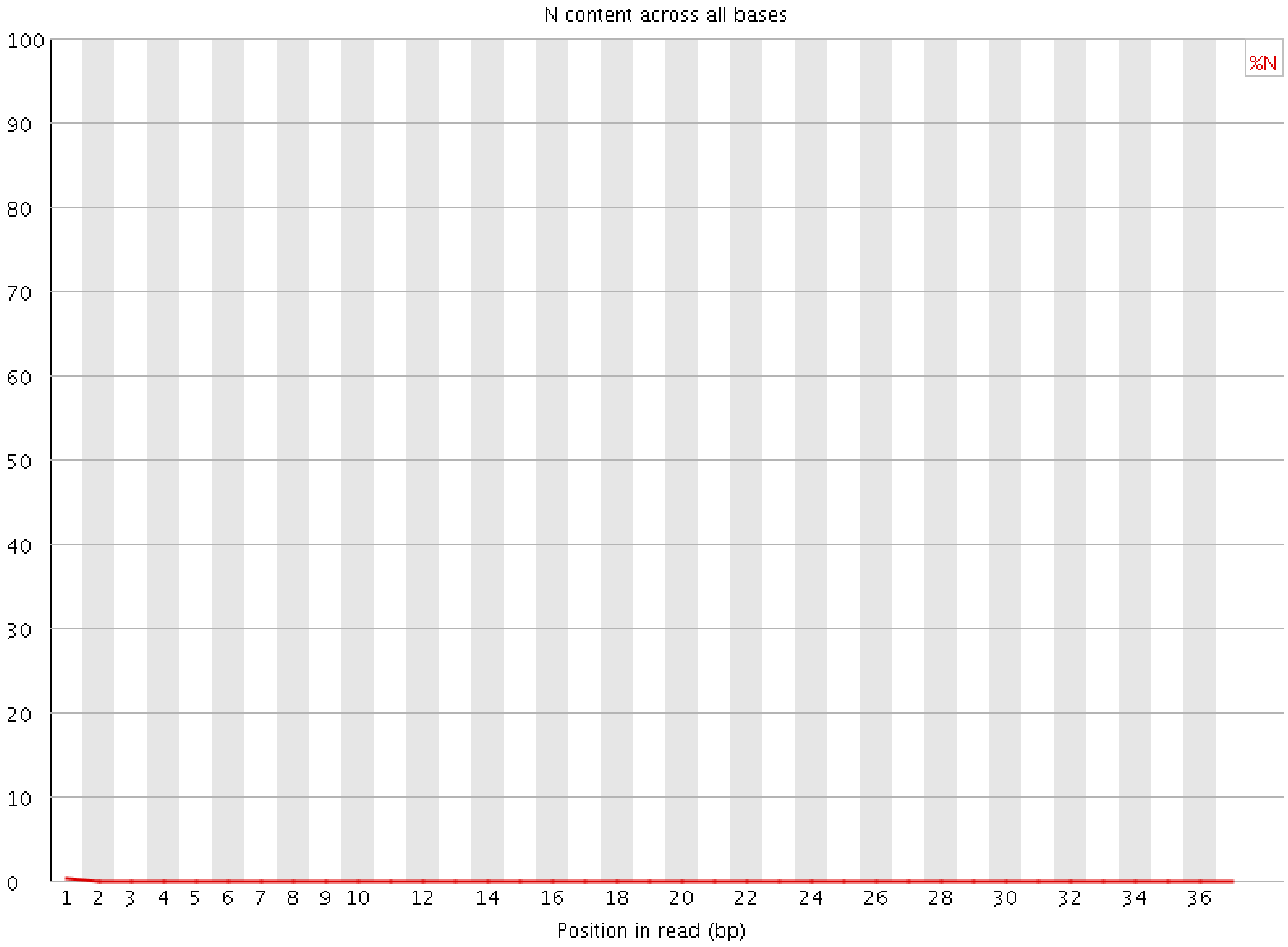


Figure 6: Per base N content. This indicates if a base call cannot be performed. The unknown base is then represented by an N. No warning/failure occurred which indicates no position shows an N content of >5%.

More information can be found at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/6%20Per%20Base%20N%20Content.html

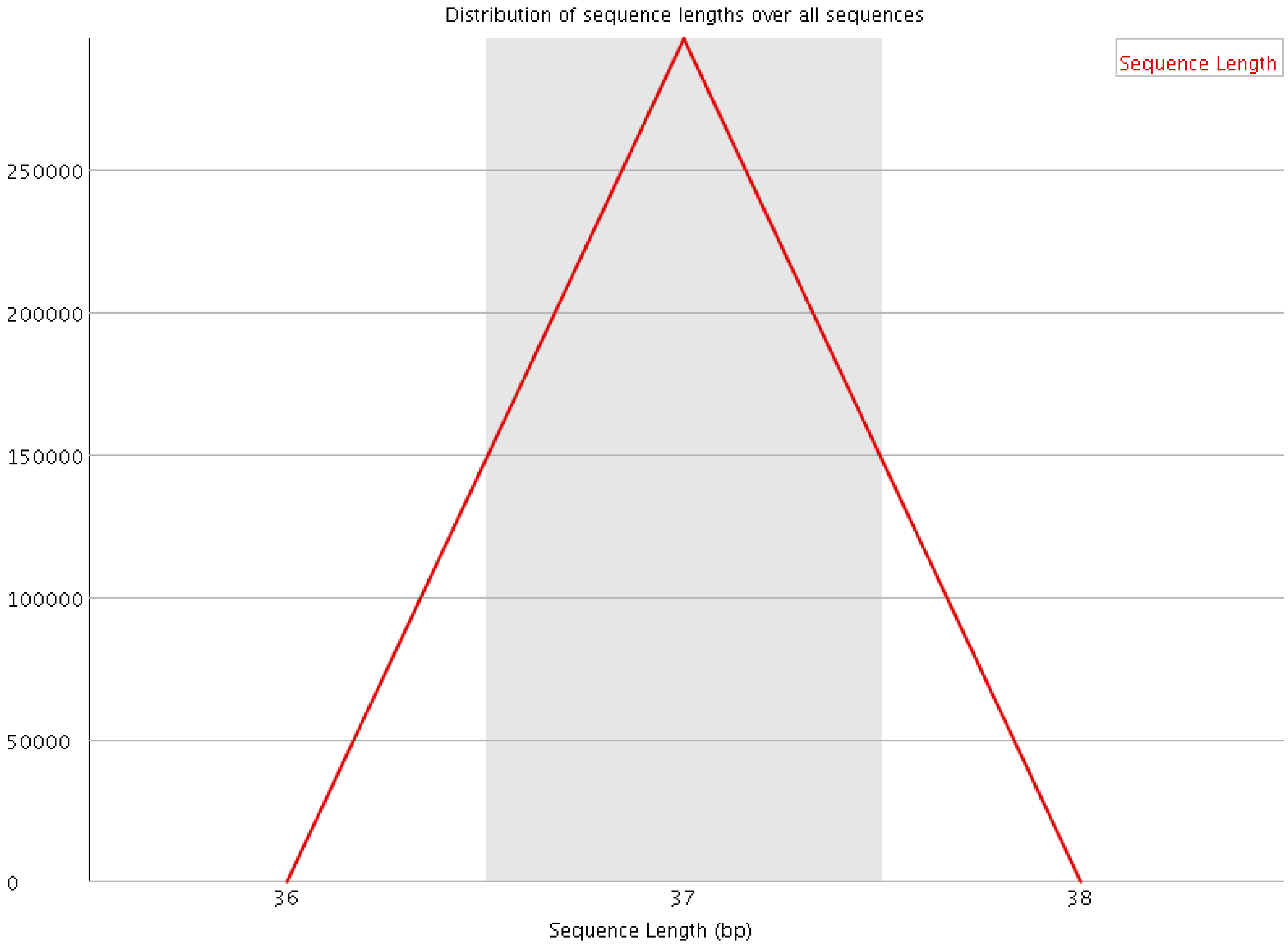


Figure 7: Sequence length distribution. This indicates the distribution of sequence fragment length to ensure that read length is typical and no wild variation occurs. The sequence length is within the range of 36-38 bp which indicates high quality, reliable reads. Trimming may still be used to remove poor quality base calls at the ends. More information can be found at:

More information can be found at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html

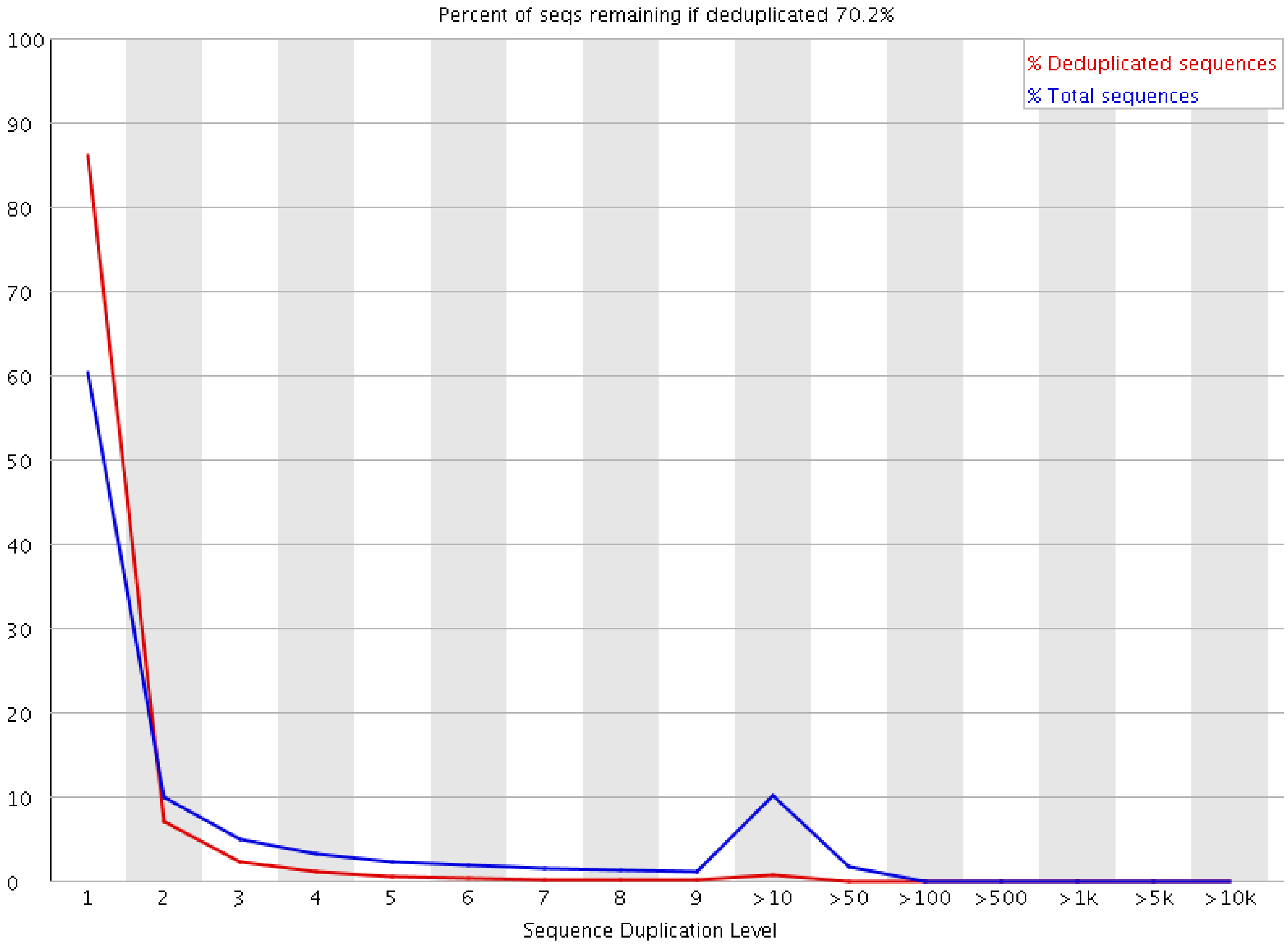


Figure 8: Sequence duplication levels. This indicates the level of duplication that occurs throughout the genome as most sequences occur only once in the sequence. Low duplication level indicates high coverage while high duplication may indicate some enrichment. A specific enrichment can be seen by the spike in the right end of the graph. No warning/failure occurs which indicates that non-unique sequences make up < 20%.

More information can be found at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html

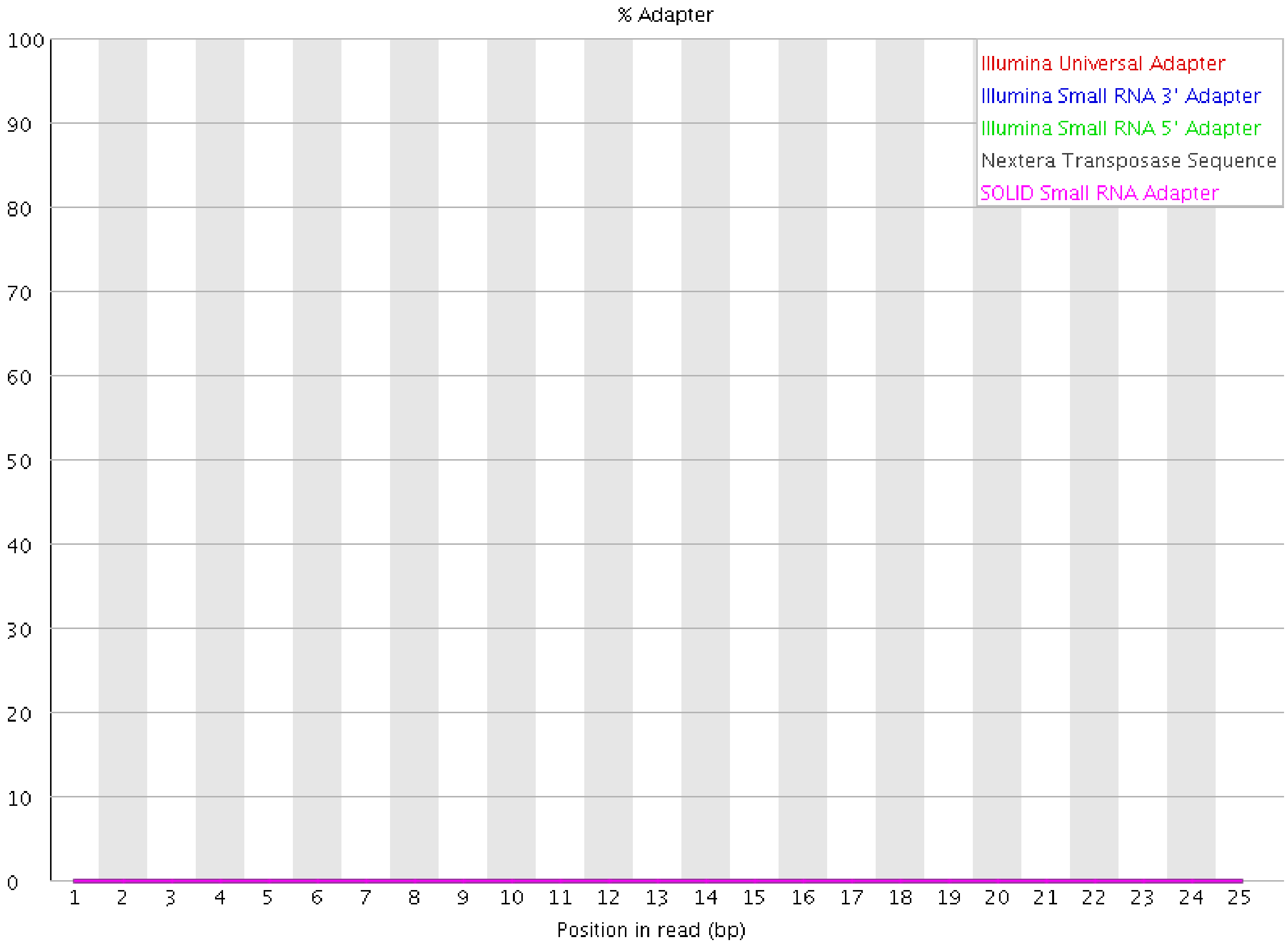
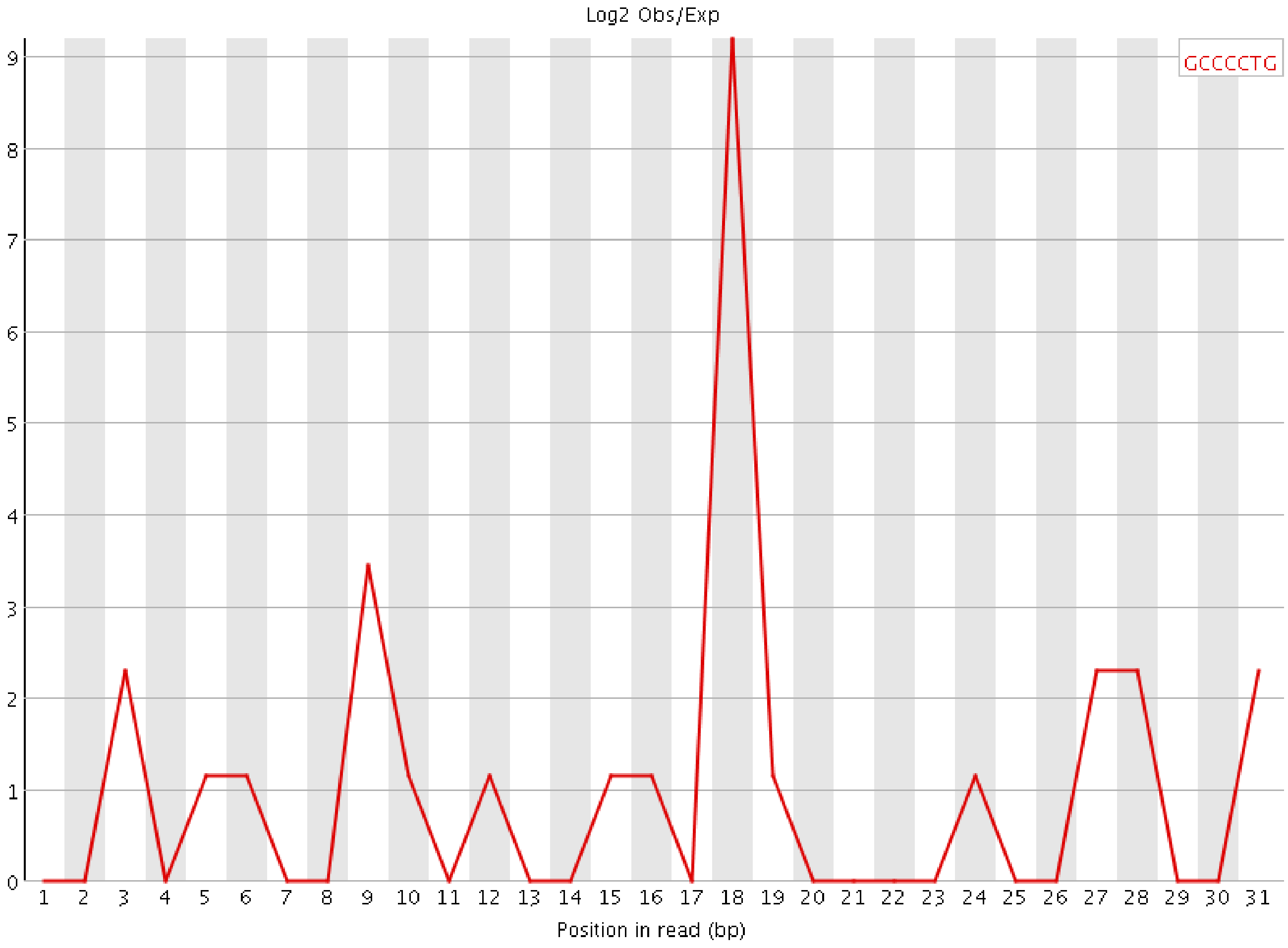


Figure 9: Adaptor content. The adaptor sequence content is analysed to determine if a significant amount of adaptors are in the sequence. This can be done to determine if adaptor trimming is required. The graph indicates that no adaptors are identified in the sequence. No warning/failure occurred which indicates adaptor content of < 5 % in all reads. More information can be found at:

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/10%20Adapter%20Content.html



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GCCCCTG	135	0.005943457	9.183783	18

Figure 10: K-mer content. This will perform an analysis to indicate the coverage per each length of read. This can indicate the presence of read-through adaptor sequences at the ends of the sequence. The presence of sequences that are overrepresented can be identified using this method. The graph indicates the proportion of each read length. A warning occurred which indicates that one kmer is in more than 5 % of reads. The warning can occur where a proportion of the read lengths are shorter than the sequence length. This may indicate adaptor trimming is required, however no adaptors were detected in the previous module.

More information can be found at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/11%20Kmer%20Content.html

Control Data

The '*input.fastq*' control file was also analysed using FASTQC. Warnings occurred for per tile sequencing quality and per base sequence count. Per tile sequencing quality can be ignored as the quality is ambiguous across all tiles and the quality is proven in other modules. The per base content warning is due to higher base levels for A and T. This can be ignored as the A-T and G-C levels are concordant which indicates no sequencing faults.

Sequence Alignment

Sequence alignment is performed using Bowtie (http://bowtie-bio.sourceforge.net/index.shtml), a short-read aligner which reads to the human genome at ultrafast rate (25 million 35-bp reads per hour). Bowtie is used for aligning to create the genome index. UCSC Genome Browser is also used to download chromosome 21 of the GRCh37/hg19 reference genome. The reference genome can be downloaded from: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/chr21.fa.gz.

Sequence Alignment is used in this analysis to annotate the peaks in test data against control data. This is can be used to determine expression level variation of TF targets that may be associated with breast cancer.

1. Load Bowtie

Command:

```
module load bowtie
```

Following login to the cluster, `bowtie` must be loaded to use its sequence alignment functions on the ChIP-Seq and control files.

2. Build the genome index

2.1 Create chromosome 21 index

Command:

```
bowtie2-build chr21.fa chr21
```

The genome index is created using `bowtie2-build` and chromosome 21 FASTA (.fa) file is downloaded from UCSC Genome Browser. Chromosome 21 was uploaded from the link above using `sftp` to the cluster. bowtie2-build creates a 'small' index for the sequence using 32-bit numbers. The result is 6 .bt2 files which contain the small indexes for the sequence.

2.2 Create SAM files

Command:

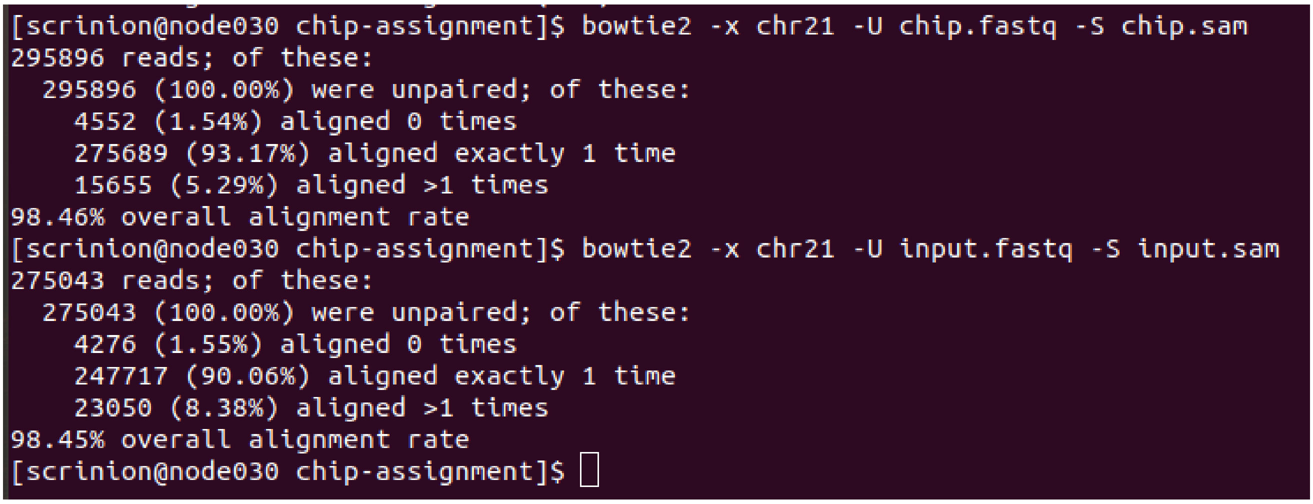
```
bowtie2 -x chr21 -U chip.fastq -S chip.sam
bowtie2 -x chr21 -U input.fastq -S input.sam
```

> `-x` Indicates the index to use

> `-U` Unpaired reads to be aligned

> `-S` File to write SAM alignments to

Output:



```
[scrinion@node030 chip-assignment]$ bowtie2 -x chr21 -U chip.fastq -S chip.sam
295896 reads; of these:
  295896 (100.00%) were unpaired; of these:
    4552 (1.54%) aligned 0 times
    275689 (93.17%) aligned exactly 1 time
    15655 (5.29%) aligned >1 times
98.46% overall alignment rate
[scrinion@node030 chip-assignment]$ bowtie2 -x chr21 -U input.fastq -S input.sam
275043 reads; of these:
  275043 (100.00%) were unpaired; of these:
    4276 (1.55%) aligned 0 times
    247717 (90.06%) aligned exactly 1 time
    23050 (8.38%) aligned >1 times
98.45% overall alignment rate
[scrinion@node030 chip-assignment]$ 
```

The unaligned chip and input FASTQ files were aligned to 98.45 % & 98.46 % respectively. SAM (Sequence Alignment/Map, .sam) alignment files contain the chip and input sequences aligned to the reference genome.

Alignment Postprocessing

`samtools` is used for alignment postprocessing and to generate BAM (Binary Alignment Map, .bam) files. BAM files contain the binary, compressed version of the genomic sequencing data in SAM files.

Samtools can be used for the manipulation of BAM files. The functions from `samtools` include sorting, duplicated removal, indexing and obtaining mapping statistics. The file name should indicate the change made at each step - files are binary and changes can be non-determinable by inspection.

Alignment postprocessing is performed in this analysis to clean the data and to create index required to inspect peaks associated with BC.

1. Load `samtools` module

Command:

```
module load samtools
```

2. Create BAM files

Command:

```
samtools view -Sb chipseq.sam > chipseq.bam
```

> `-Sb` specifies to ignore previous samtools versions and generate the BAM file `chipseq.bam`. This indicates that 1 sequence is present.

A BAM file (.bam) is the binary version of a SAM file. The BAM and SAM files are both required for the Integrative Genomics Viewer (IGV).

3. Remove duplicates

Command:

```
samtools rmdup chip.bam chip.rmdup.bam
samtools rmdup input.bam input.rmdup.bam
```

> `rmdup` removes potential PCR duplicates.

4. Sort the files

Command:

```
samtools sort chip.rmdup.bam chip.rmdup.sorted
samtools sort input.rmdup.bam input.rmdup.sorted
```

> `sort` sort by alignment to leftmost coordinates and writes to new file `sorted.bam`.

5. Index the BAM files

Command:

```
samtools index chip.rmdup.sorted.bam
samtools index input.rmdup.sorted.bam
```

> `index` is used to create the the BAI (Binary Alignment Index, .bai) file.

The BAI (Binary Alignment Index, .bai) file is generated using `index`. Indexing is required for correct alignment of the genome when using IGV.

6. Generate mapping statistics

Command:

```
samtools flagstat chip.rmdup.sorted.bam > chip_mappingstats.txt
samtools flagstat input.rmdup.sorted.bam > input_mappingstats.txt
```

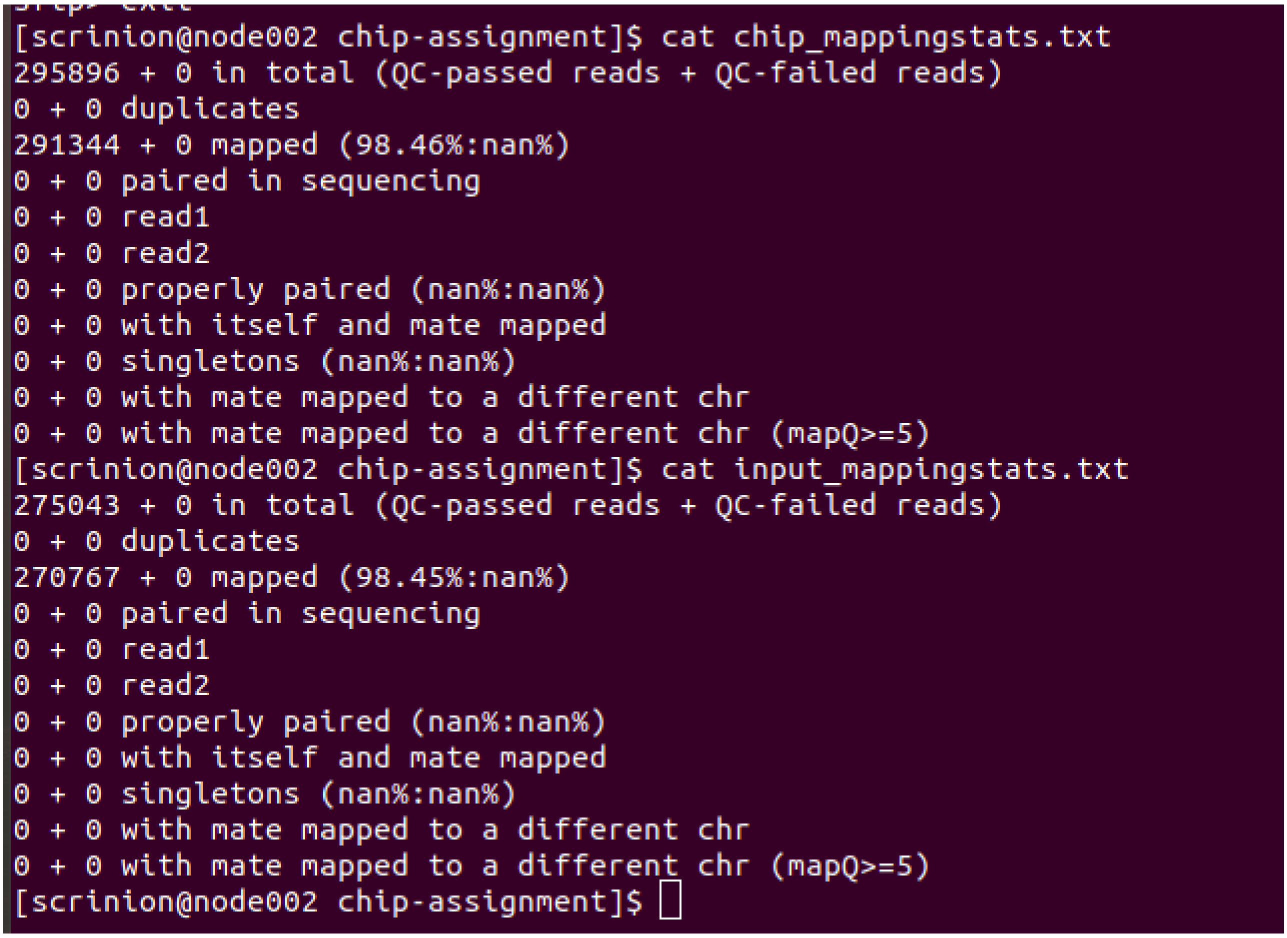
> `flagstat` this does complete run of the bam file to print read counts and summary statistics.

7. Inspect the mapping statistics

Command:

```
cat chip_mappingstats.txt
cat input_mappingstats.txt
```

Output:



```
[scrinion@node002 chip-assignment]$ cat chip_mappingstats.txt
295896 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
291344 + 0 mapped (98.46%:nan%)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (nan%:nan%)
0 + 0 with itself and mate mapped
0 + 0 singletons (nan%:nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
[scrinion@node002 chip-assignment]$ cat input_mappingstats.txt
275043 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
270767 + 0 mapped (98.45%:nan%)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (nan%:nan%)
0 + 0 with itself and mate mapped
0 + 0 singletons (nan%:nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
[scrinion@node002 chip-assignment]$ 
```

The `flagstat` flag generates statistics about the read QC and mapping. Counts are taken for 13 categories of potential issues, of which the ChIP-Seq and input files passed QC and experienced no failures. The first row indicates the total number of reads for ChIP-Seq and control files as 295,896 and 275,043 reads respectively. The statistics above indicate high quality reads for both files.

Alignment Visualisation

Alignment visualisation is performed using the Integrative Genomics Viewer (IGV), a high performance visualisation tool provided by the Broad Institute. This used to explore integrated genomic data. For this experiment, IGV is used to compare peaks in the ChIP-Seq and control. The IGV user manual can be found at: https://software.broadinstitute.org/software/igv/UserGuide.

Alignment visualisation can identify coverage changes in the test vs. control data. This can be used to identify novel peaks but can also be used to inspect expression levels of known BC genes.

1. Load IGV

Command:

```
module load java
module load IGV/2.1.17
```

Java is required to use IGV so load Java in preparation. Alternative to using the cluster modules, IGV can be downloaded (https://software.broadinstitute.org/software/igv/download) and files transferred using `sftp` to perform the alignment visualisation on your local machine.

2. Load a genome

Use the drop-down list from the IGV drop down. The 'Human hg19' genome is selected. Select the appropriate chromosomal region of the analysis, chromosome 21. Select the region of interest by entering the gene or locus. The sequencing data is visualised using the BAM or SAM file. The BAI files are loaded automatically to index the alignments. The coverage track is also automatically generated for visualisation.

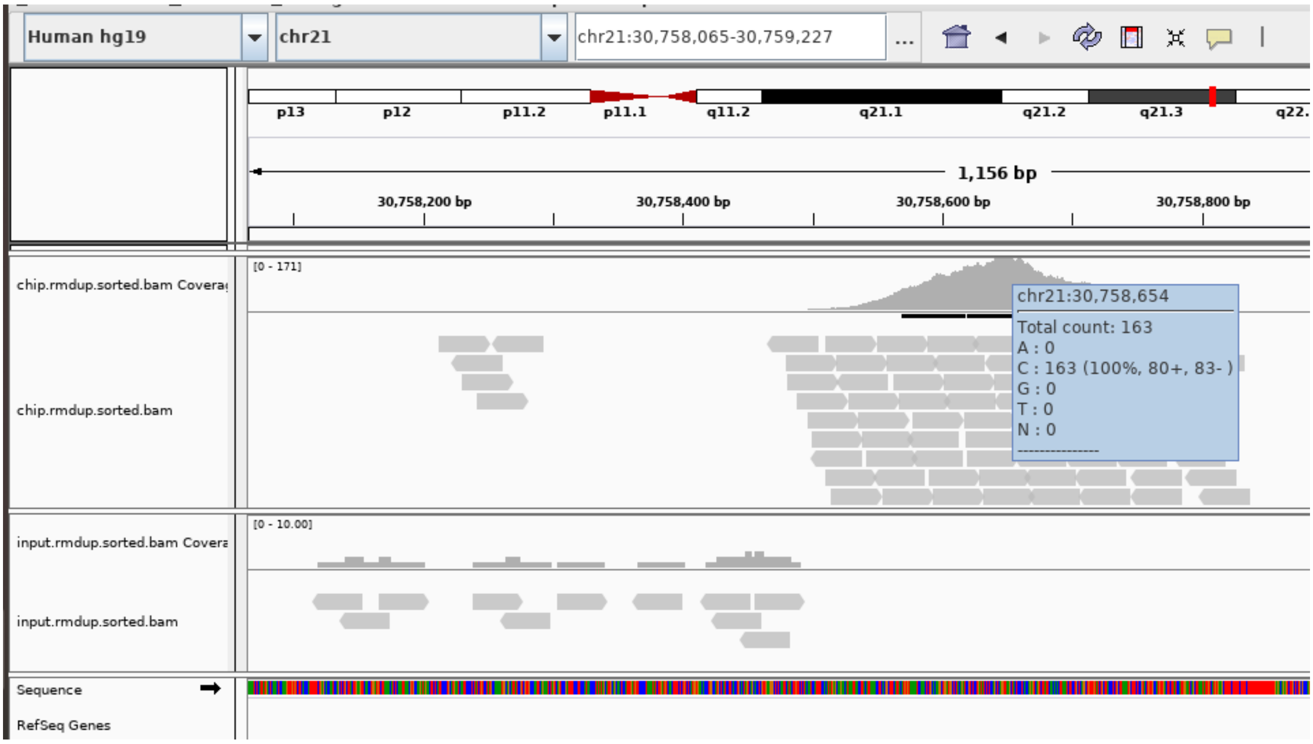
Navigate to:

> file > load from file > chip.sorted.bam
>
> file > load from file > input.sorted.bam

3. Explore the genome

The IGV tracks are used to search the genome for any peak data of interest. 2 genes from chromosome 21 associated with breast cancer (http://www.cancerindex.org/geneweb/clinkc21.htm) were selected to inspect.

3.1 Peak Comparison


Human hg19

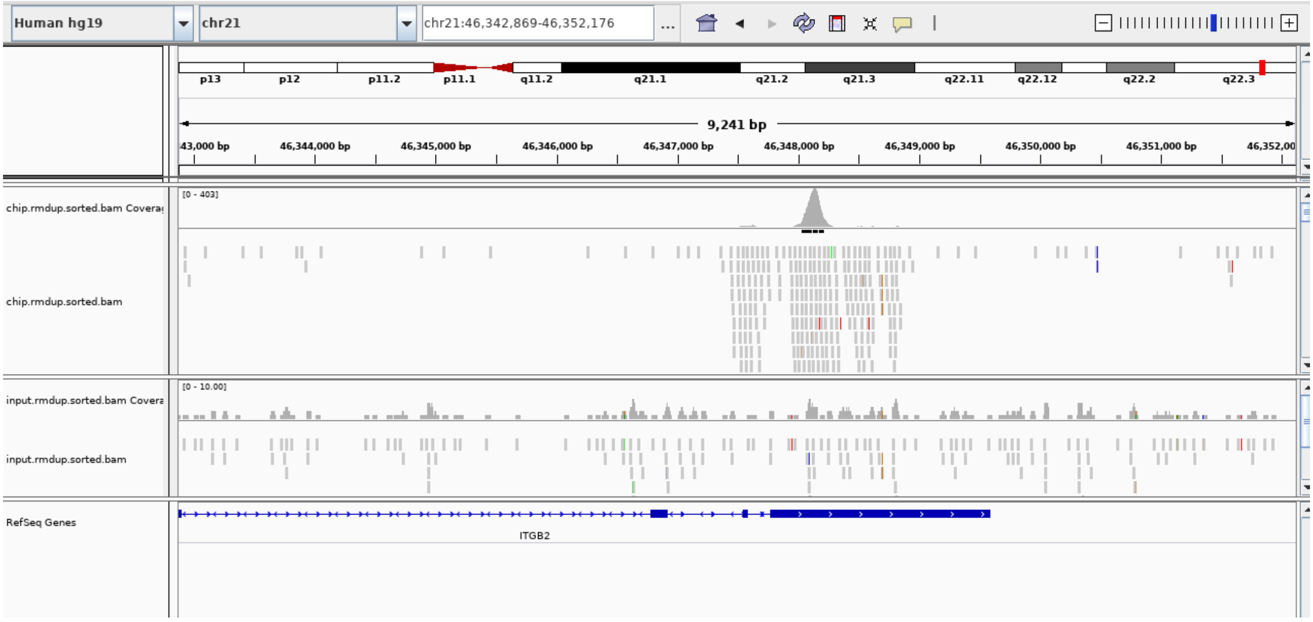
chr21

chr21:30,758,065-30,759,227

chr21:30,758,654
Total count: 163
A : 0
C : 163 (100%, 80+, 83-)
G : 0
T : 0
N : 0

Figure 1: Example of read variation in test vs. control. The loaded data can be used to identify regions or genes of interest. The top track indicates the location across the genome. The second track indicates the test data peaks range and coverage. The third track indicates the peaks and coverage of the control data. The final track indicates the sequence nucleotide, each corresponding with a colour (A = green, C = blue, G = orange and T = red) and the RefSeq Genes are listed corresponding with the sequence. Hovering over a region indicates the total count for the peak. The above indicates a peak of the test data with 163 counts (chr21:30,758,654) that is not found found in the control sequence. The region above is an example of peak differences in the test vs. control data. A high peak can indicate that there an association with the region and the test genomic data.

3.2 Integrin Subunit Beta 2 (ITGB2)


Human hg19

chr21

chr21:46,342,869-46,352,176

Figure 2: Peak variation identified in ITGB2. This is a gene associated with breast cancer which shows different levels of reads in the test vs the control. Read levels increase is synonymous with increased expression. Here is indication of higher expression level of a BC-linked gene as seen using IGV. The increased expression level appears to be in the promoter region of the gene which indicates influence to transcription.

3.3 Trefoil Factor 3 (TFF3)

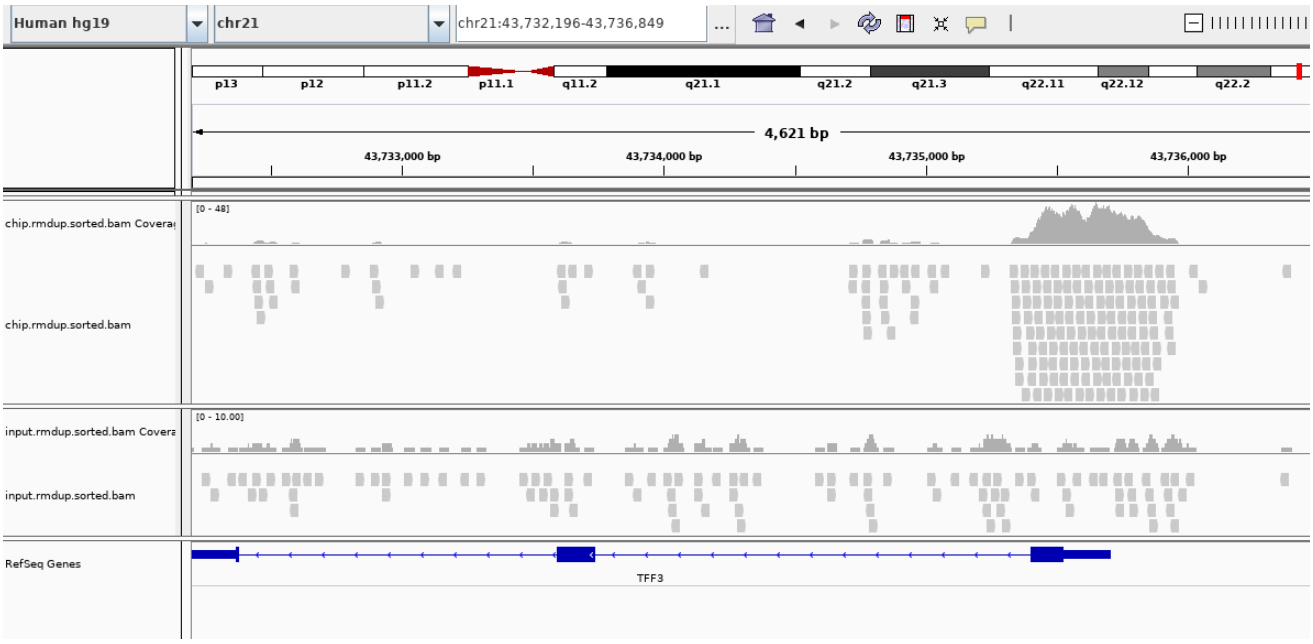


Figure 3: Peak variation identified in TFF3. TFF3 is a transcription factor on chromosome 21 that is associated with breast cancer (https://www.ncbi.nlm.nih.gov/gene/7031). The test data indicates that a high number of reads occurs at the promoter region. Due to the previous association of TFF3 with breast cancer and the DNA interaction pattern identifiable with IGV, this may have an important regulatory role in the cancer cell line.

Peak Calling

The peak calling is performed using `macs2`. Model-based Analysis of ChIP-Seq (MACS) is a novel algorithm used to identify transcription factor binding sites. MACS indicates the importance of chromatin complexity and its relationship to gene expression. More information regarding MACS usage can be found from the manual: https://github.com/taoliu/MACS/blob/macs_v1/README.rst.

Peak calling is used to limit the analysis to regions of interest in the sequence. This is required for computational purposes and increased speed in TF identification.

1. Generate XLS files

Command:

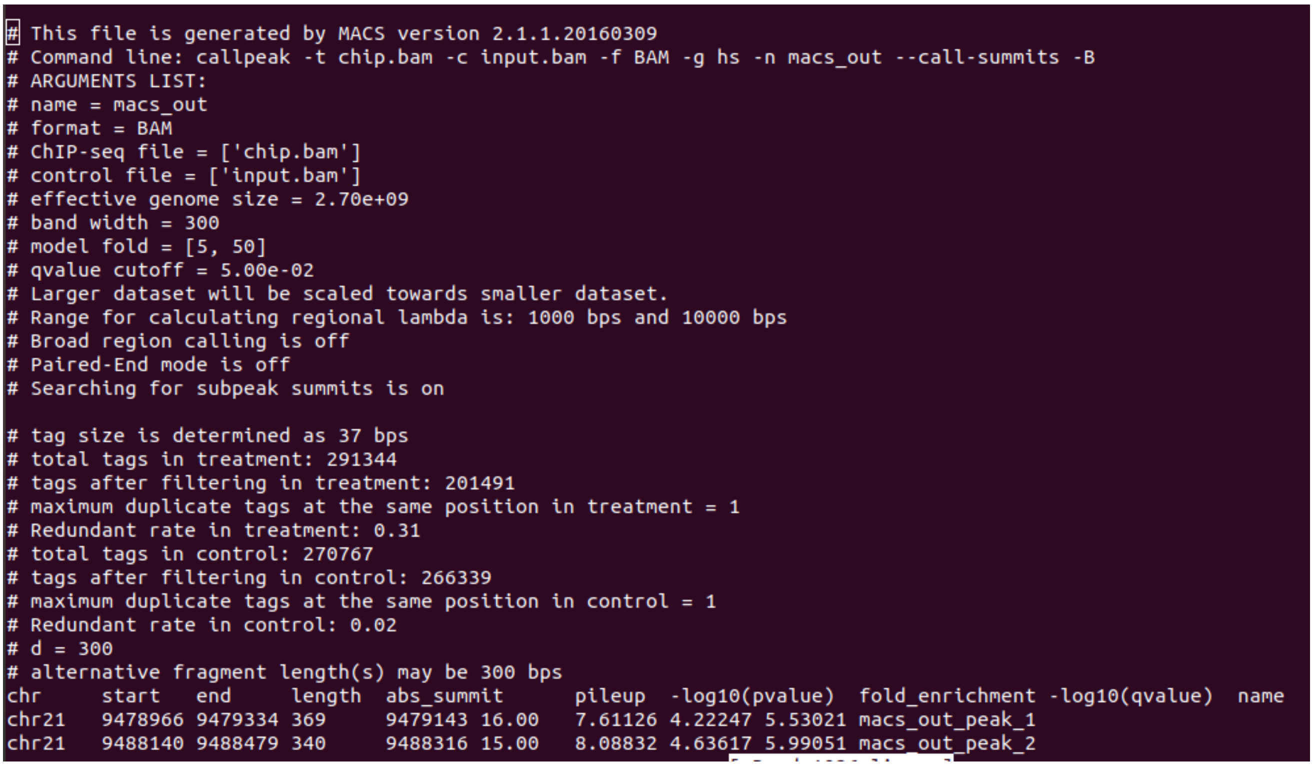
```
macs2 callpeak -t chip.bam -c input.bam -f BAM -g hs -n macs_out --call-summits -B
```

> `callpeak` call peaks from the alignment results.

> `-t` indicates the treatment file
>
> `-c` indicates the control file
>
> `-f` indicates the input is in BAM format
>
> `-g` indicates the size which is 'human genome'
>
> `-n` indicates the output file name
>
> `--call-summits` can be used to deconvolve the subpeaks within each peak by reanalysing the signal shape.

The XLS file that is created is a tabular file that contains information about the peaks called. The file contains information that can be used for targeted analysis of the peaks.

2. Extract the peaks



```
# This file is generated by MACS version 2.1.1.20160309
# Command line: callpeak -t chip.bam -c input.bam -f BAM -g hs -n macs_out --call-summits -B
# ARGUMENTS LIST:
# name = macs_out
# format = BAM
# ChIP-seq file = ['chip.bam']
# control file = ['input.bam']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off
# Paired-End mode is off
# Searching for subpeak summits is on

# tag size is determined as 37 bps
# total tags in treatment: 291344
# tags after filtering in treatment: 201491
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.31
# total tags in control: 270767
# tags after filtering in control: 266339
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.02
# d = 300
# alternative fragment length(s) may be 300 bps
chr     start   end     length  abs_summit      pileup  -log10(pvalue)  fold_enrichment -log10(qvalue)  name
chr21   9478966 9479334 369     9479143 16.00   7.61126 4.22247 5.53021 macs_out_peak_1
chr21   9488140 9488479 340     9488316 15.00   8.08832 4.63617 5.99051 macs_out_peak_2
```

The information provided from the .xls contains the following:

- chromosome name
- start position of peak
- end position of peak
- length of peak region
- absolute peak summit position
- pileup height at peak summit, -log10(pvalue) for the peak summit (e.g. pvalue =1e-10, then this value should be 10)
- fold enrichment for this peak summit against random Poisson distribution with local lambda, -log10(qvalue) at peak summit

The XLS file is used to extract the chromosomal coordinates of the peaks however formatting needs to be universal and is edited using `nano`.

3. Annotate the peaks

Command:

```
nano macs_out_peaks.xls
```

`nano` is used for editing basic text in a Linux environment. The data contained in the above output is removed from the .xls file so only chromosomal peak data is contained.

Command:

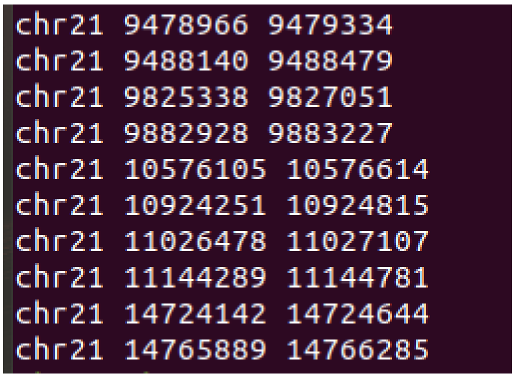
```
awk '{print $1,$2,$3}' macs_out_peaks.xls > peaks.bed
```

Following deletion of the top lines, the peak coordinates only are extracted in the correct format (chr start stop) by using `awk` to create a new file of only the peak data.

Command:

```
head peaks.bed
```

Output:



```
chr21 9478966 9479334
chr21 9488140 9488479
chr21 9825338 9827051
chr21 9882928 9883227
chr21 10576105 10576614
chr21 10924251 10924815
chr21 11026478 11027107
chr21 11144289 11144781
chr21 14724142 14724644
chr21 14765889 14766285
```

The `head` command is used to inspect and ensure that the .bed file contains a list of the genome-wide binding locations in the correct format.

Peak Annotations

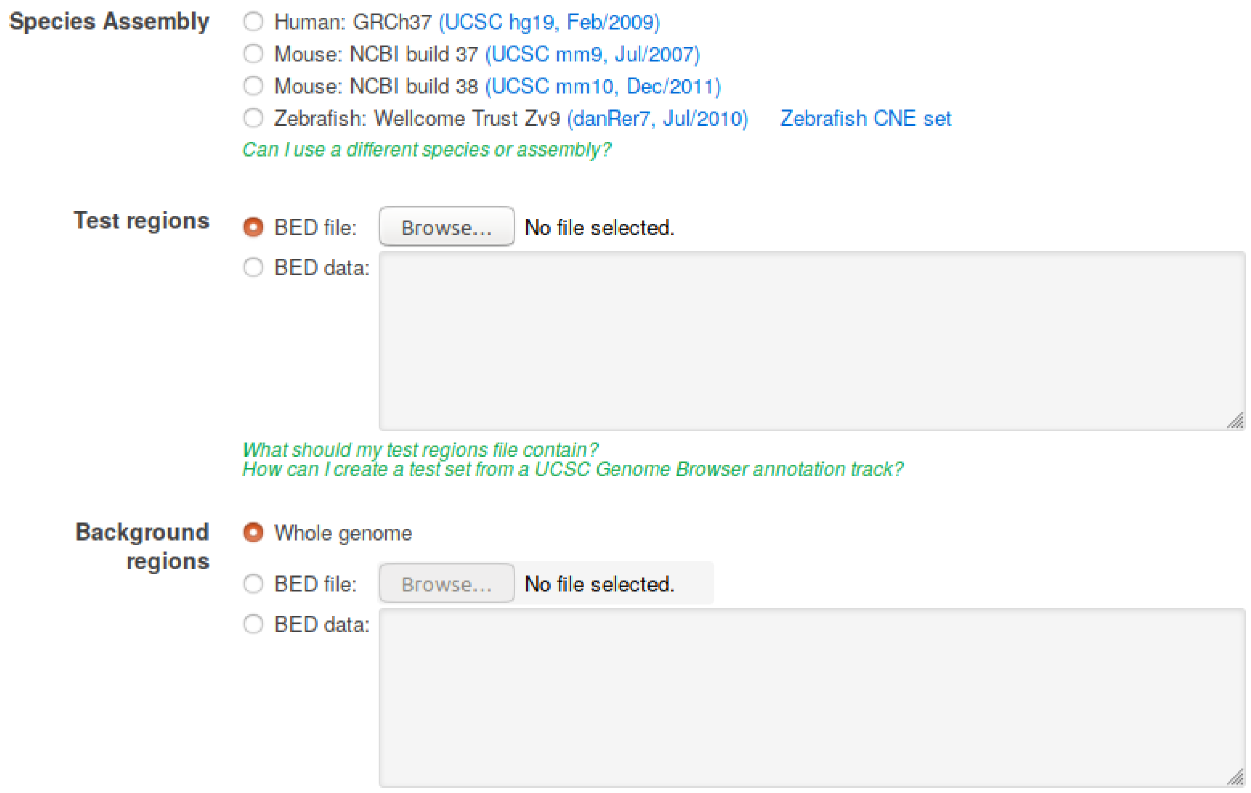
Peak annotation is performed using GREAT (Genomic Regions Enrichment of Annotations Tool), a software which is used to predict functions of cis-regulatory regions using binding events across the genome. GREAT associates function with non-coding genomic regions by analysing the annotation of nearby genes. Annotation of distal binding sites and false positive control is possible using a binomial test over the input region. The GREAT process performs peak annotation using the following steps:

1. Input the genomic regions

2. GREAT indicates the proximal and distal input regions with their assocaited target genes.
3. Gene annotations from ontologies are outline to associate the regions with disease or gene set.
4. Statistical enrichment is calculated for association of the genomic regions with annotations.
5. The output contains annotations significantly associated with a set the input regions.

Further information regarding annotation mechanisms using GREAT can be found at: http://bejerano.stanford.edu/papers/GREAT.pdf.

1. Upload peaks to GREAT



Species Assembly
○ Human: GRCh37 (UCSC hg19, Feb/2009)
○ Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
○ Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
○ Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) Zebrafish CNE set
Can I use a different species or assembly?

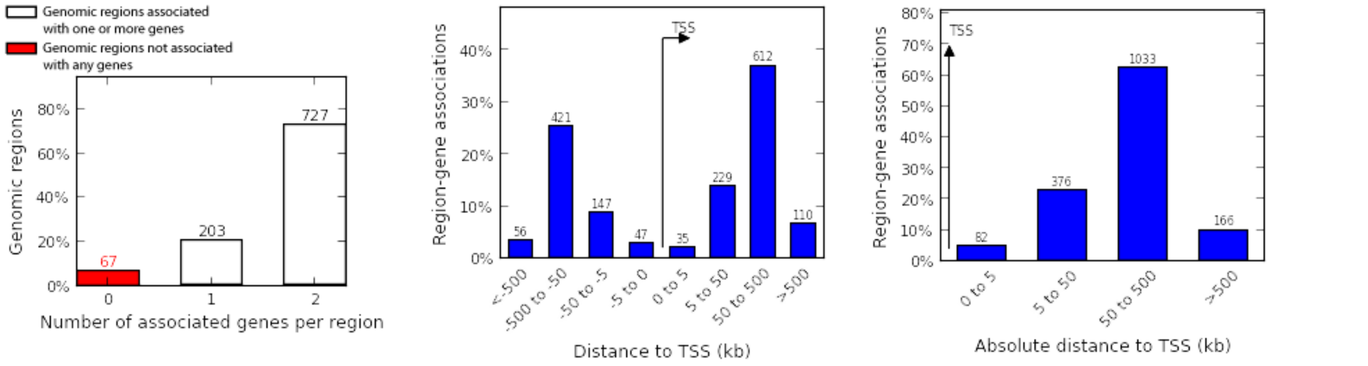
Test regions
● BED file: Browse... No file selected.
○ BED data:
What should my test regions file contain?
How can I create a test set from a UCSC Genome Browser annotation track?

Background regions
● Whole genome
○ BED file: Browse... No file selected.
○ BED data:

Peaks called from test data on breast cancer cells are uploaded to http://great.stanford.edu/public/html/. The correct build of human genome is selected. The peak annotations in uniform format (and stored as BED file) are uploaded and compared to the whole genome.

2. Explore Region-Gene Association Graphs

2.1 Gene-Region Association Graphs



Of the 997 genomic regions contained in the BED file, 157 (1%) of 18,041 genes were contained. The imputed data is the

Graph 1: Number of associated genes per region. Most regions are associated with < 2 genes per region. This indicates that for each peak which was imputed, there are 727 cases in which genes are found.

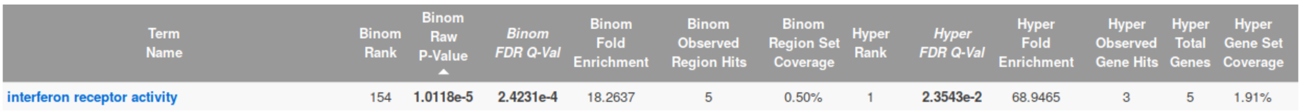
Graph 2: Distance and orientation to transcription start site (TSS). This indicates that most genes are about 50 to 500 kb away from a TSS, a point at which the transcription begins on the 5' end. This information can be used to identify if a TF is in within region of potential targets and influence expression levels. The orientation is important as TF regulation towards the 3' end of the strand.

Graph 3: Absolute distance to TSS: Indicates the total gene from the potential target site of the transcription factor. This reveals whether identified regions are within range to either regulate or be regulated by a transcription factor.

2.2 Gene Ontology (GO) Molecular Function

The GO indicates the ontology of which the molecular function is most related to. This can be used to identify if the sequence region is enriched for a particular function. This may reveal particular biological process, cellular component and molecular function that may be dyregulated during oncogenesis.

Interferon Receptor Activity

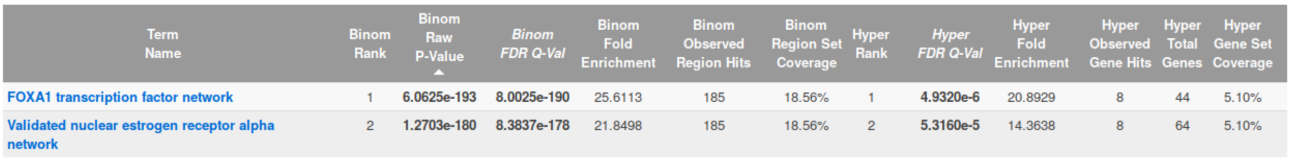


Term Name	Binom Rank	Binom Raw P-Value ▲	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
interferon receptor activity	154	**1.0118e-5**	**2.4231e-4**	18.2637	5	0.50%	1	**2.3543e-2**	68.9465	3	5	1.91%

The GREAT analysis indicates that there is a significant enrichment of genes associated with interferon receptor activity. This can be considered in motif analysis if any results are from this ontology.

2.3 Molecular Signature Database (MSigDB) Pathway

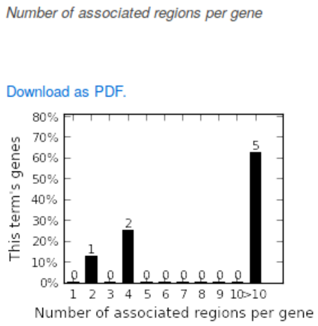
The molecular signature database (MSigDB) contains annotated gene sets that represent biological processes in large-scale genomic data. MSigDB contains the largest number of gene sets overall and combines manual curation and automatic computational means.



Term Name	Binom Rank	Binom Raw P-Value ▲	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
FOXA1 transcription factor network	1	6.0625e-193	8.0025e-190	25.6113	185	18.56%	1	4.9320e-6	20.8929	8	44	5.10%
Validated nuclear estrogen receptor alpha network	2	1.2703e-180	8.3837e-178	21.8498	185	18.56%	2	5.3160e-5	14.3638	8	64	5.10%

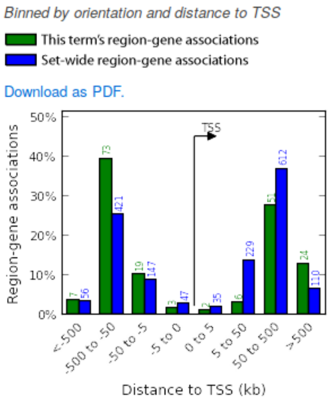
The results indicated an enrichment for FOXA1 and oestrogen receptors. It is noteworthy however that the MSigBD has 1320 terms covering 8117 (45%) of all 18041 genes which indicates that the FDR is high.

FOXA1 transcription factor network



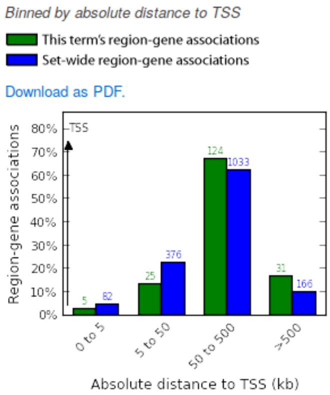
Number of associated regions per gene

Download as PDF.



Binned by orientation and distance to TSS

Download as PDF.

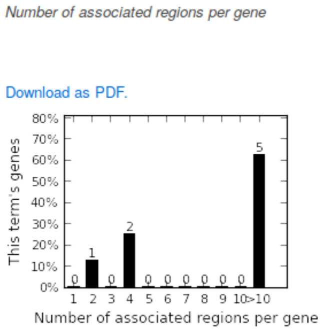


Binned by absolute distance to TSS

Download as PDF.

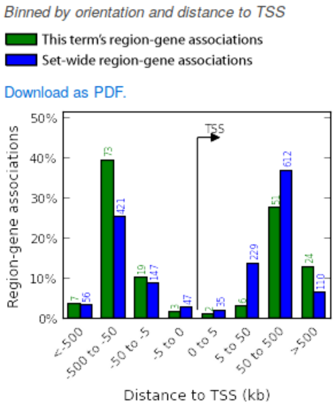
The most common number of associated regions per gene is >10 which indicates a high level of enrichment in some regions for the FOXA1 genes. Most are within 50-500 kb of a TSS which indicates the potential for TF influencing expression. Again, the FOXA1 network can be considered during motif analysis.

Validated nuclear oestrogen receptor alpha network



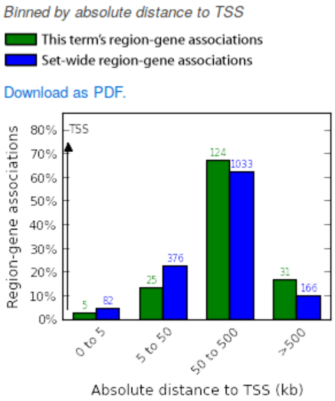
Number of associated regions per gene

Download as PDF.



Binned by orientation and distance to TSS

Download as PDF.



Binned by absolute distance to TSS

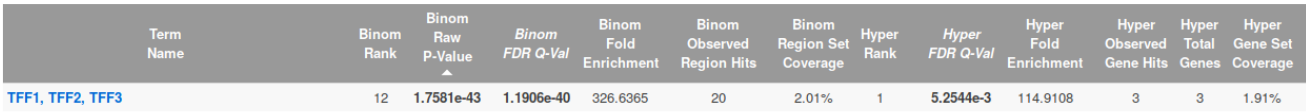
Download as PDF.

Again, the association regions is >10 per gene. The genes are most often within region of a TSS. However, the significance of this enrichment is questionable: this analysis is being performed on a BC cell line so some oestrogen association should be expected.

2.4 TreeFam

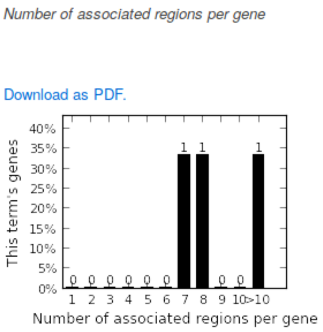
TreeFam (Tree families database) is a database of phylogenetic trees of animal genes that provides homology predictions. TreeFam has 8,126 terms covering 13,550 (75%) of all 18,041 genes, indicating again that FDR could be high. This is considerably higher coverage to the MSigDB previously outlined.

TFF1, TFF2, TFF3

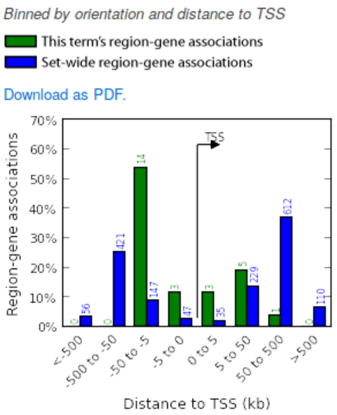


Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
TFF1, TFF2, TFF3	12	1.7581e-43	1.1906e-40	326.6365	20	2.01%	1	5.2544e-3	114.9108	3	3	1.91%

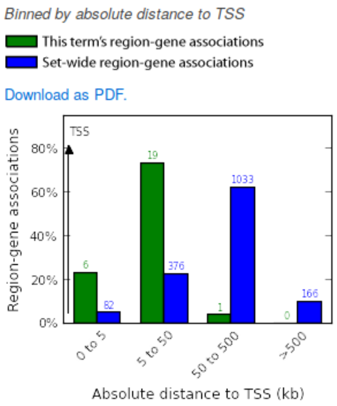
The results indicate association of the coordinate with the TFF family. This is an interesting result considering the association of this family with breast cancer (outlined above) and the high level of peak reads visualised around the promoter region using IGV.


Number of associated regions per gene

Download as PDF.


Binned by orientation and distance to TSS

Download as PDF.


Binned by absolute distance to TSS

Download as PDF.

As can be seen in the graphs, the level of association is not as high as seen in the MSigDB pathways. However, the rate of FDR is much lower for these samples given the higher power. Higher statistical power and the known association between the TFF genes and breast cancer suggest that these may be identified in the motif analysis.

Motif Analysis

Motif analysis is used to determine defining nucleotide sequences that can identify a TF. For the purpose of this study, motif analysis can identify the TF in the data with the regulatory role.

Numerous methods are used for motif detection. Bedtools is first used for compatibility. MEME is first used, followed by DREME which is an associated software with a different motif detection algorithm. MEME-ChIP is a specialised version that performs MEME/DREME which is also used. Each of these motif detection methods are paired with TOMTOM, a motif comparison software used to annotate the motifs to TFs. RSAT is then used to complement motif annotation results.

1. Bedtools

Bedtools provides a wide range of genomic analysis tools that can be used on BAM, BED and other file formats. More information can be found from the user manual at: https://media.readthedocs.org/pdf/bedtools/latest/bedtools.pdf

1.1 Ensure compatibility with *bedtools*

Command:

```
tr " " "\t" < peaks.bed > newpeaks.bed
```

`bedtools` has the additional requirement that all files are tab delimited. This requires using `awk` to replace all space characters with tab. Another requirement is that the chromosome naming follows an identical pattern in all files eg. 'chr21' and '1' do not work together.

1.2 Get sequences from FASTA file

Command:

```
module load bedtools
bedtools getfasta -fi chr21.fa -bed newpeaks.bed -fo peaks.fasta
```

> `-fi` indicates the FASTA input file
>
> `-bed` indicates the peak coordinates
>
> `-fo` indicates the file name of the new file

Use intervals to extract sequences defined by the BED file and create a new FASTA file following the same format.

2. MEME

MEME (Multiple Expected Maximisation for Motif Elicitation) and DREME (Discriminative Regular Expression for Motif Elicitation) are used for motif-based sequence analysis. MEME discovered novel, ungapped motifs in sequences and patterns identified into separate motifs.

2.1 Load MEME

Command:

```
module load python
module load meme
```

2.2 Run MEME analysis

```
meme peaks.fasta -dna -mod zoops -minw 6 -maxw 10000 -nmotifs 5 -o meme_out
```

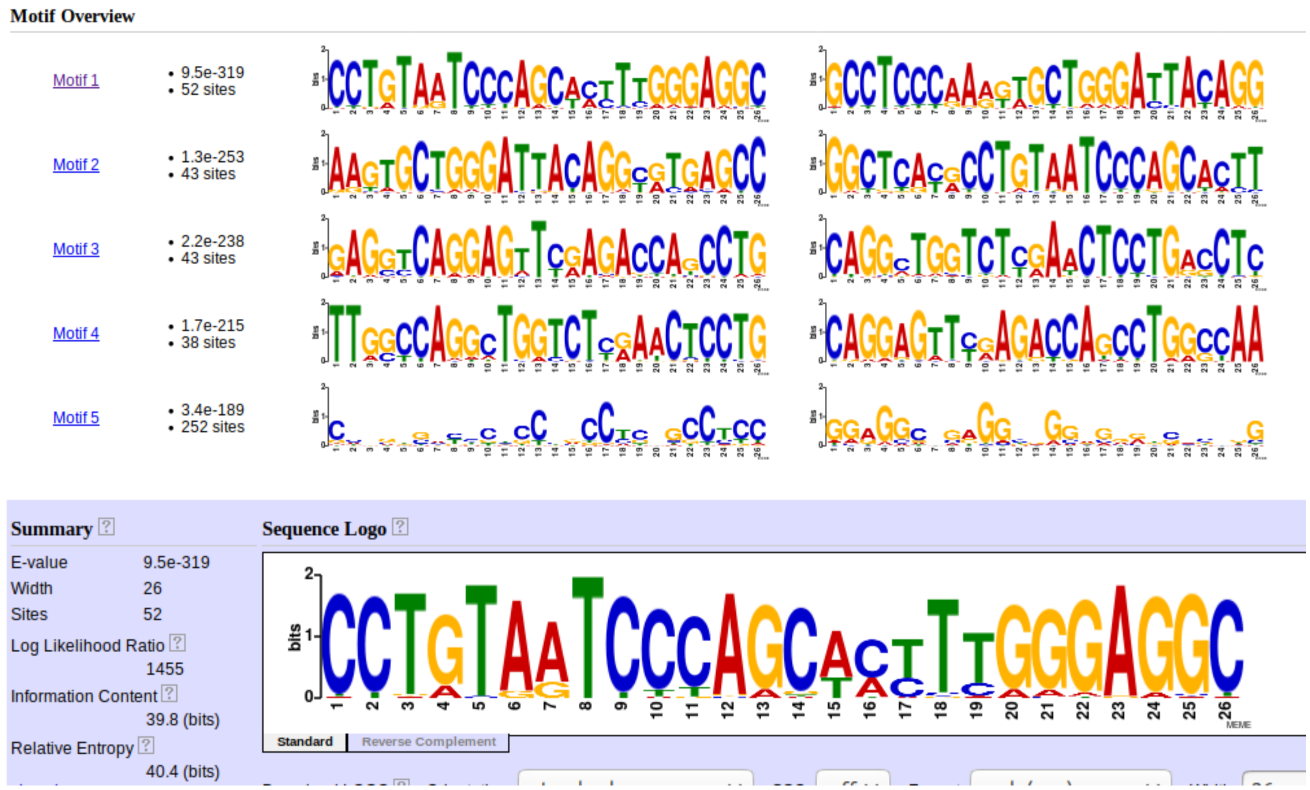
> `-dna` indicates that the FASTA file contains DNA sequences.
>
> `-mod` indicates to count either 0 or 1 occurrence of a motif using `zoops`
>
> `-minw` indicates the minimum motif width
>
> `-maxw` indicates the maximum motif width

> `-nmotifs` limits the number of motifs detected for computational purposes
>
> `-o` indicates the output file name

The FASTA file is selected for analysis and contains the genomic data contained within the chromosomal peak regions. This data will be used to identify any motifs that may be used to find the TF.

The command specifies that DNA data is contained. The `zoops` is used to specify that some motifs may be missing from the primary sequence and identify more accurate results than the standard setting.

2.3 Output: Top Motifs



Motif Overview

Motif 1	• 9.5e-319 • 52 sites	
Motif 2	• 1.3e-253 • 43 sites	
Motif 3	• 2.2e-238 • 43 sites	
Motif 4	• 1.7e-215 • 38 sites	
Motif 5	• 3.4e-189 • 252 sites	

Summary

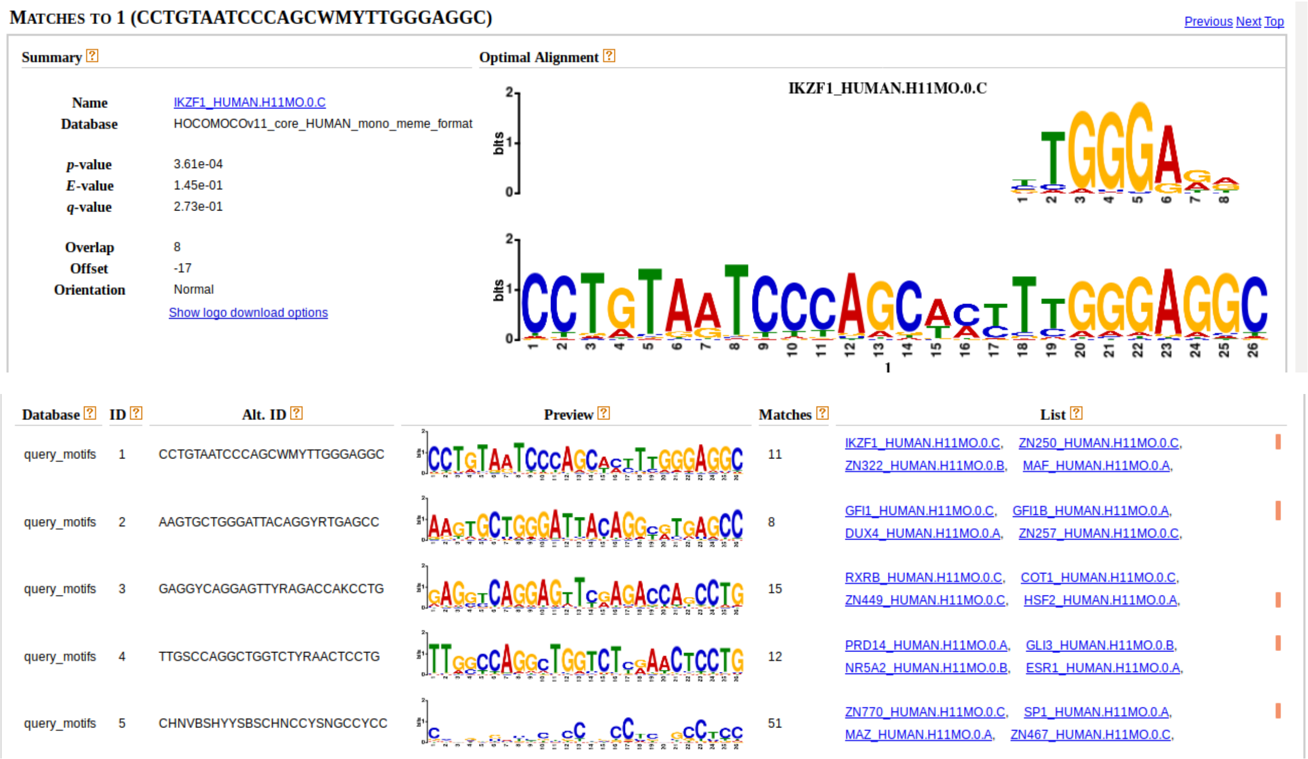
E-value	9.5e-319
Width	26
Sites	52
Log Likelihood Ratio	1455
Information Content	39.8 (bits)
Relative Entropy	40.4 (bits)

The top motifs identified range from 38-252 base pairs in length. The motif most commonly occurring is predominantly Cs and Gs so can be discounted. This PSPM coordinates are obtained to use for TOMTOM analysis.

2.4 Output: TOMTOM

TOMTOM is used to compare one or more motifs against a database of known motifs to rank the motifs in the database and provide alignment for each significant match. A job can be submitted for analysis at: http://tools.genouest.org/tools/meme/cgi-bin/tomtom.cgi. TOMTOM searches for the query motif against JASPER or the selected database. This can identify if the motif is specifically associated with any transcription factor that may be implemented in BC.

The TOMTOM generates reports on the motif-based sequence alignment. Each motif as a p-value which indicates the probability of a random motif of the same width having the same alignment. The e-value is a parameter to describe the number of expected hits, similarly to the p-value. The q-value is the p-value adjusted for FDR.



MATCHES TO 1 (CCTGTAATCCCAGCWMYTTGGGAGGC)

Summary ?

Name	IKZF1_HUMAN.H11MO.0.C
Database	HOCOMOCOv11_core_HUMAN_mono_meme_format
p-value	3.61e-04
E-value	1.45e-01
q-value	2.73e-01
Overlap	8
Offset	-17
Orientation	Normal
	Show logo download options

Optimal Alignment ?

Database ?	ID ?	Alt. ID ?	Preview ?	Matches ?	List ?
query_motifs	1	CCTGTAATCCCAGCWMYTTGGGAGGC		11	IKZF1_HUMAN.H11MO.0.C, ZN250_HUMAN.H11MO.0.C, ZN322_HUMAN.H11MO.0.B, MAF_HUMAN.H11MO.0.A,
query_motifs	2	AAGTGCTGGGATTACAGGYRTGAGCC		8	GFI1_HUMAN.H11MO.0.C, GFI1B_HUMAN.H11MO.0.A, DUX4_HUMAN.H11MO.0.A, ZN257_HUMAN.H11MO.0.C,
query_motifs	3	GAGGYCAGGAGTTYRAGACCAKCCTG		15	RXRB_HUMAN.H11MO.0.C, COT1_HUMAN.H11MO.0.C, ZN449_HUMAN.H11MO.0.C, HSF2_HUMAN.H11MO.0.A,
query_motifs	4	TTGSCCAGGCTGGTCTYRAACTCCTG		12	PRD14_HUMAN.H11MO.0.A, GLI3_HUMAN.H11MO.0.B, NR5A2_HUMAN.H11MO.0.B, ESR1_HUMAN.H11MO.0.A,
query_motifs	5	CHNVBSHYYSBSCHNCCYSNGCCYCC		51	ZN770_HUMAN.H11MO.0.C, SP1_HUMAN.H11MO.0.A, MAZ_HUMAN.H11MO.0.A, ZN467_HUMAN.H11MO.0.C,

List of genes associated with the top motif: FOXJ3_HUMAN.H11MO.0.A, PRDM6_HUMAN.H11MO.0.C, MEF2B_HUMAN.H11MO.0.A, SRY_HUMAN.H11MO.0.B, GATA3_HUMAN.H11MO.0.A, MEF2D_HUMAN.H11MO.0.A, FOXQ1_HUMAN.H11MO.0.C, ANDR_HUMAN.H11MO.0.A, MEF2A_HUMAN.H11MO.0.A, ZFP28_HUMAN.H11MO.0.C, SOX5_HUMAN.H11MO.0.C

TOMTOM results indicate that FOXJ3 genes are asssociatred with the highest matching motif. This is interesting considering the GREAT results which indicated an increased expression level of the FOXA.

3. DREME

DREME discovers short, ungapped motifs that are relatively enriched in the sequence to perform motif discovery for short regularly expressed motifs. DREME uses regular expression restriction which allows for high speed detection of motifs. DREME is available from the same site (http://meme-suite.org/doc/dreme.html) and can be used with larger dataset such as ChIP-seq data. DREME can scale to a better quality than MEME. DREME is loaded as part of MEME.

3.1 Run MEME Analysis

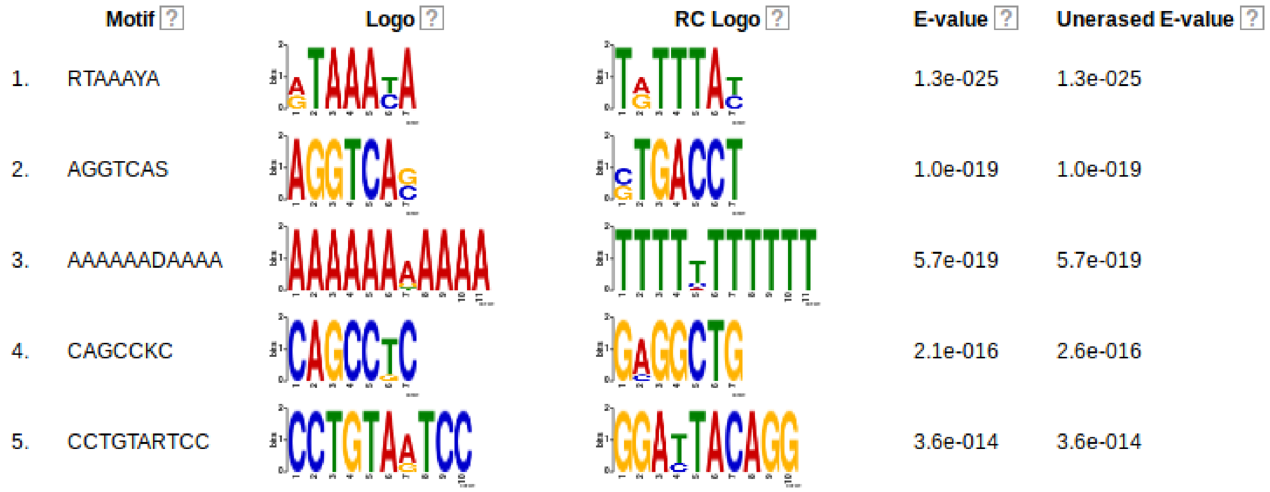
Command:

```
dreme -m 5 -mink 6 -maxk 26 -p peaks.fasta -o dreme_out
```

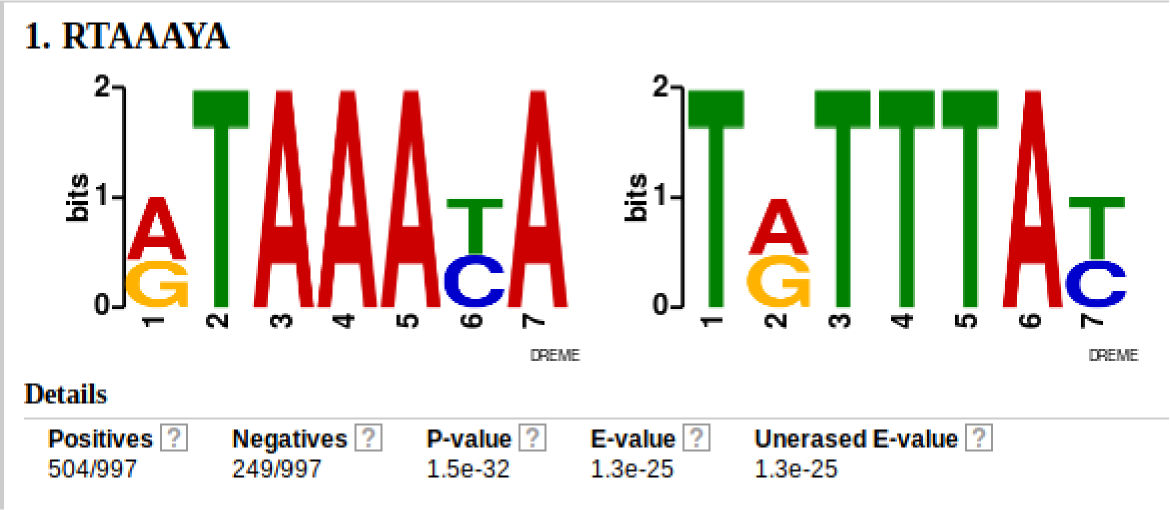
> `-m` indicates the number of motifs to test
>
> `-mink` indicates the minimum width of the motif score
>
> `-maxk` indicates the maximum width of the motif core
>
> `-p` indicates the primary FASTA sequence file to use
>
> `-o` indicates the output file to use

The number of motifs is limited to 5. The minimum and maximum motif width is set as 5 and 26 as these are the initial motifs used for detection.

3.2 Output: Top motifs



	Motif	Logo	RC Logo	E-value	Unerased E-value
1.	RTAAAYA			1.3e-025	1.3e-025
2.	AGGTCAS			1.0e-019	1.0e-019
3.	AAAAAADAAAA			5.7e-019	5.7e-019
4.	CAGCCKC			2.1e-016	2.6e-016
5.	CCTGTARTCC			3.6e-014	3.6e-014



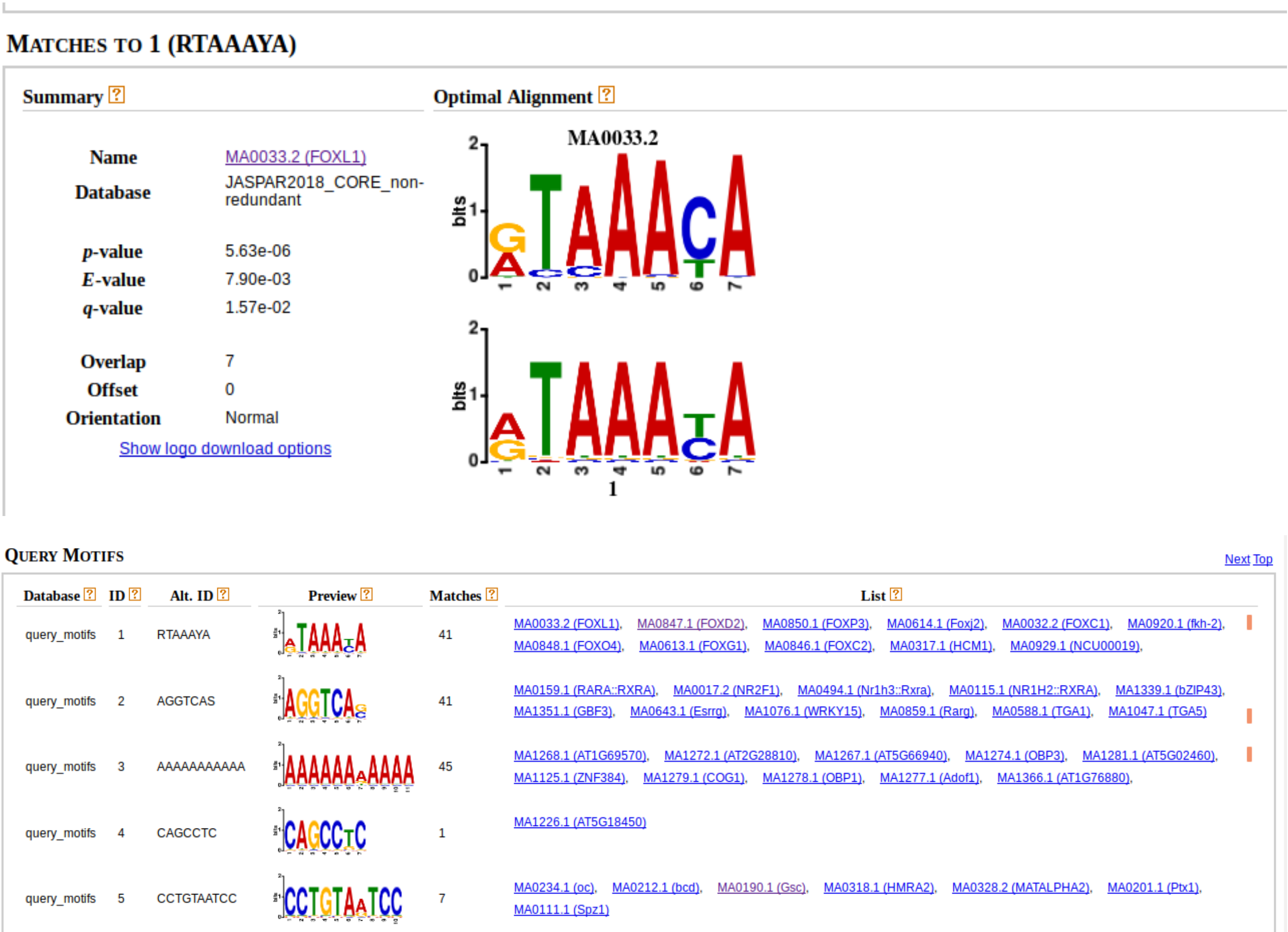
1. RTAAAYA

Details

Positives	Negatives	P-value	E-value	Unerased E-value
504/997	249/997	1.5e-32	1.3e-25	1.3e-25

The top motif RTAAYA was identified in 504/997 peaks. Although this is a high number of matches, the negatives value indicates the number of times the sequence is found in a randomly shuffled sequence. This indicates that this motif is truly less significant than it may seem.

3.3 Output: TOMTOM



Matches to 1 (RTAAAYA)

Summary ?

Name	MA0033.2 (FOXL1)
Database	JASPAR2018_CORE_non-redundant
***p*-value**	5.63e-06
***E*-value**	7.90e-03
***q*-value**	1.57e-02
Overlap	7
Offset	0
Orientation	Normal

Show logo download options

Optimal Alignment ?

Query Motifs

Next Top

Database ?	ID ?	Alt. ID ?	Preview ?	Matches ?	List ?
query_motifs	1	RTAAAYA		41	MA0033.2 (FOXL1), MA0847.1 (FOXD2), MA0850.1 (FOXP3), MA0614.1 (Foxj2), MA0032.2 (FOXC1), MA0920.1 (fkh-2), MA0848.1 (FOXO4), MA0613.1 (FOXG1), MA0846.1 (FOXC2), MA0317.1 (HCM1), MA0929.1 (NCU00019),
query_motifs	2	AGGTCAS		41	MA0159.1 (RARA::RXRA), MA0017.2 (NR2F1), MA0494.1 (Nr1h3::Rxra), MA0115.1 (NR1H2::RXRA), MA1339.1 (bZIP43), MA1351.1 (GBF3), MA0643.1 (Esrrg), MA1076.1 (WRKY15), MA0859.1 (Rarg), MA0588.1 (TGA1), MA1047.1 (TGA5)
query_motifs	3	AAAAAAAAAAA		45	MA1268.1 (AT1G69570), MA1272.1 (AT2G28810), MA1267.1 (AT5G66940), MA1274.1 (OBP3), MA1281.1 (AT5G02460), MA1125.1 (ZNF384), MA1279.1 (COG1), MA1278.1 (OBP1), MA1277.1 (Adof1), MA1366.1 (AT1G76880),
query_motifs	4	CAGCCTC		1	MA1226.1 (AT5G18450)
query_motifs	5	CCTGTAATCC		7	MA0234.1 (oc), MA0212.1 (bcd), MA0190.1 (Gsc), MA0318.1 (HMRA2), MA0328.2 (MATALPHA2), MA0201.1 (Ptx1), MA0111.1 (Spz1)

List of genes associated with top motif:

MA0033.2 (FOXL1), MA0847.1 (FOXD2), MA0850.1 (FOXP3), MA0614.1 (Foxj2), MA0032.2 (FOXC1), MA0920.1 (fkh-2), MA0848.1 (FOXO4), MA0613.1 (FOXG1), MA0846.1 (FOXC2), MA0317.1 (HCM1), MA0929.1 (NCU00019), MA0849.1 (FOXO6), MA0042.2 (FOXI1), MA0297.1 (FKH2), MA0845.1 (FOXB1), MA0851.1 (Foxj3), MA1103.1 (FOXK2), MA0458.1 (slp1), MA0047.2 (Foxa2), MA0031.1 (FOXD1), MA0157.2 (FOXO3), MA0030.1 (FOXF2), MA0546.1 (pha-4), MA0593.1 (FOXP2), MA0040.1 (Foxq1), MA0480.1 (Foxo1), MA0148.3 (FOXA1), MA0446.1 (fkh), MA0013.1 (br(var.4)), MA0852.2 (FOXK1), MA0481.2 (FOXP1), MA0773.1 (MEF2D), MA0660.1 (MEF2B), MA0296.1 (FKH1),

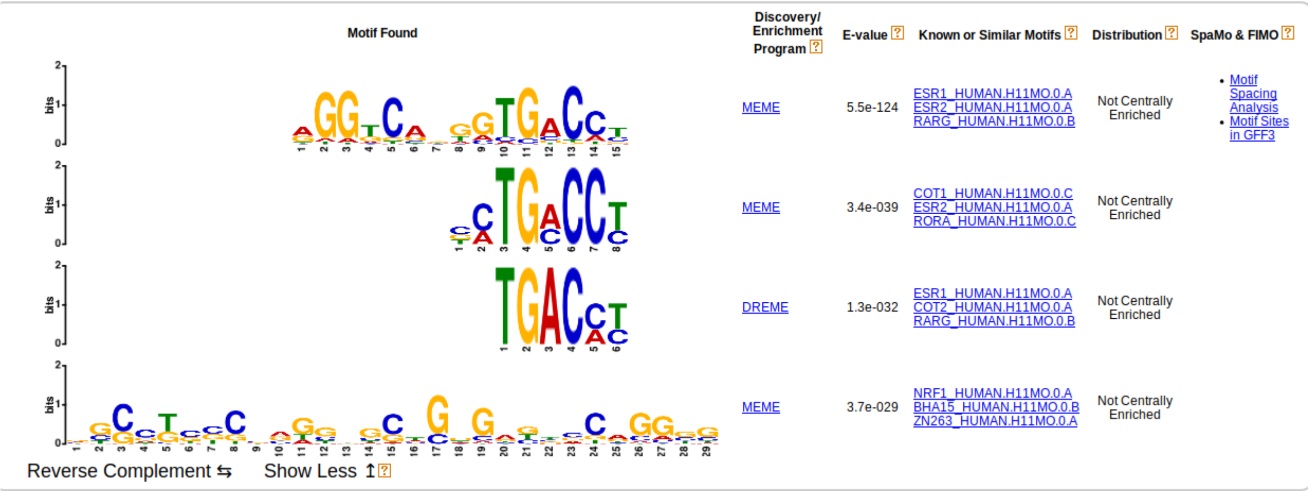
The TOMTOM analysis indicates that there are many associations of this motif with the FOXA1 transcription factor network. Many of these genes are from the FOX family which suggests that one may be the transcription factor of interest. Although promising, more analyses are performed due to the short motif length (7 bp) in an attempt to resolve ambiguity of TF identity.

4. MEME-ChIP

4.1 Submit Sequences

MEME-ChIP performs motif analysis on large sets of sequences such as those from ChIP-seq. The data can be submitted at: http://meme-suite.org/tools/meme-chip. Alternatively, the analysis command for upload can be found using the provided link.

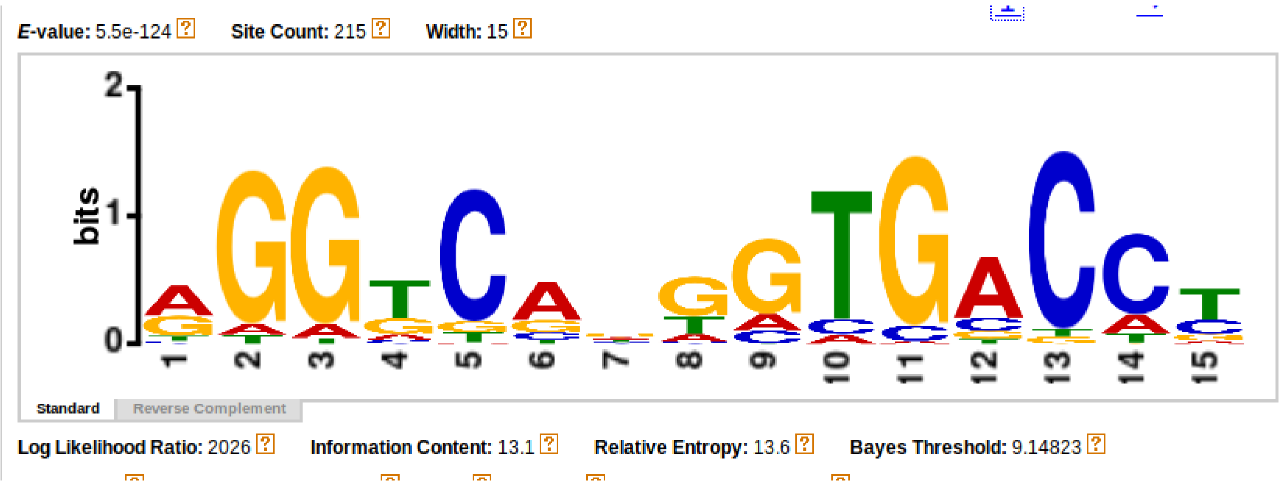
4.2 Output: Top Matching Motifs



Motif Found	Discovery/ Enrichment Program ?	E-value ?	Known or Similar Motifs ?	Distribution ?	SpaMo & FIMO ?
	MEME	5.5e-124	ESR1_HUMAN.H11MO.0.A ESR2_HUMAN.H11MO.0.A RARG_HUMAN.H11MO.0.B	Not Centrally Enriched	• Motif Spacing Analysis • Motif Sites in GFF3
	MEME	3.4e-039	COT1_HUMAN.H11MO.0.C ESR2_HUMAN.H11MO.0.A RORA_HUMAN.H11MO.0.C	Not Centrally Enriched	
	DREME	1.3e-032	ESR1_HUMAN.H11MO.0.A COT2_HUMAN.H11MO.0.A RARG_HUMAN.H11MO.0.B	Not Centrally Enriched	
	MEME	3.7e-029	NRF1_HUMAN.H11MO.0.A BHA15_HUMAN.H11MO.0.B ZN263_HUMAN.H11MO.0.A	Not Centrally Enriched	

The top matching motifs from MEME-ChIP indicate association with ESR1, ESR2 and RARG. It is possible to consider this as the expected binding motif considering its complex structure.

4.3 Output: Top Matching Motif

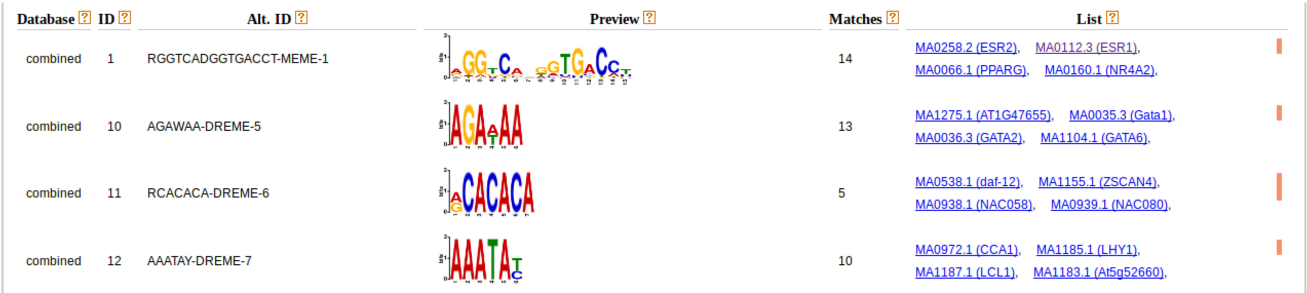


E-value: 5.5e-124 ?   Site Count: 215 ?   Width: 15 ?

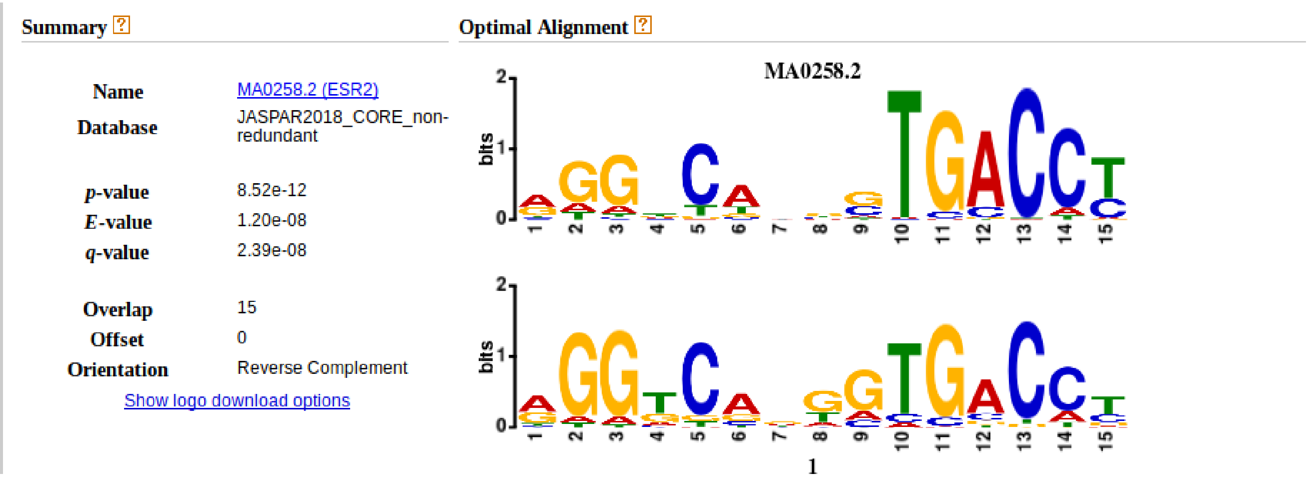
Log Likelihood Ratio: 2026 ?   Information Content: 13.1 ?   Relative Entropy: 13.6 ?   Bayes Threshold: 9.14823 ?

Upon further inspection, the motif is found to be within 215/997. Considering the complex motif structure, this could be considered a high match for the motif.

4.4 Output: TOMTOM



Database ?	ID ?	Alt. ID ?	Preview ?	Matches ?	List ?
combined	1	RGGTCADGGTGACCT-MEME-1		14	MA0258.2 (ESR2), MA0112.3 (ESR1), MA0066.1 (PPARG), MA0160.1 (NR4A2),
combined	10	AGAWAA-DREME-5		13	MA1275.1 (AT1G47655), MA0035.3 (Gata1), MA0036.3 (GATA2), MA1104.1 (GATA6),
combined	11	RCACACA-DREME-6		5	MA0538.1 (daf-12), MA1155.1 (ZSCAN4), MA0938.1 (NAC058), MA0939.1 (NAC080),
combined	12	AAATAY-DREME-7		10	MA0972.1 (CCA1), MA1185.1 (LHY1), MA1187.1 (LCL1), MA1183.1 (At5g52660),



Summary ?

Name	MA0258.2 (ESR2)
Database	JASPAR2018_CORE_non-redundant
***p*-value**	8.52e-12
***E*-value**	1.20e-08
***q*-value**	2.39e-08
Overlap	15
Offset	0
Orientation	Reverse Complement

Show logo download options

Optimal Alignment ?

List of genes associated with top motif:

MA0258.2 (ESR2), MA0112.3 (ESR1), MA0066.1 (PPARG), MA0160.1 (NR4A2), MA0505.1 (Nr5a2), MA0363.1 (REB1), MA1107.1 (KLF9), MA0858.1 (Rarb(var.2)), MA0159.1 (RARA::RXRA), MA0740.1 (KLF14), MA1147.1 (NR4A2::RXRA), MA0730.1 (RARA(var.2)), MA1112.1 (NR4A1), MA0413.1 (USV1)

The TOMTOM analysis indicates the ESR genes as top hits for the motif. This motif appears to be concordant with the ESR family of genes and is fthe likely TF with a reliably complex motif and high q-value. An area of concern however is the relatively low site count vs. total peaks.

---

5. RSAT

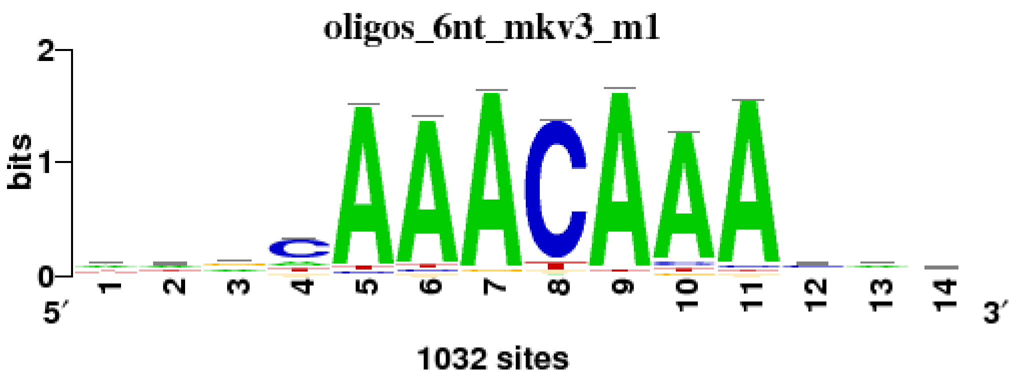
RSAT (Regulatory Sequence Analysis Tools) can also be used to analyse cis-regulatory functions as (http://rsat.sb-roscoff.fr/). The purpose of RSAT in this experiment was to confirm peak over-representation and synonymy with motif analysis obtained using MEME Suite tools.

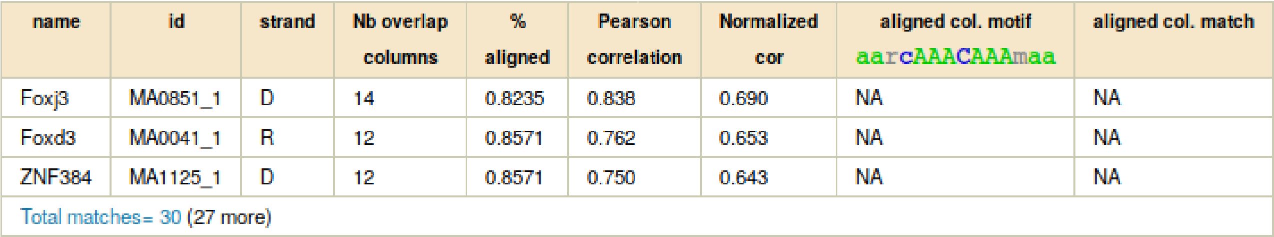
5.1 Submit Sequences

Sequences were submitted in FASTA format at www.rsat.sb-roscoff.fr/peak-motifs_form.cgi.

5.2 Output Analysis

**





name	id	strand	Nb overlap columns	% aligned	Pearson correlation	Normalized cor	aligned col. motif aarcAAACAAAmaa	aligned col. match
Foxj3	MA0851_1	D	14	0.8235	0.838	0.690	NA	NA
Foxd3	MA0041_1	R	12	0.8571	0.762	0.653	NA	NA
ZNF384	MA1125_1	D	12	0.8571	0.750	0.643	NA	NA
Total matches= 30 (27 more)								

The results of the analysis indicate the the Foxj3 gene is contains the most concordant motif. This result was also obtained using MEME analysis and GREAT analysis which indicated increased expression of the FOXA1 transcription factor network.

Conclusion:

Transcription factor mutations constitute a core component of the proliferative signalling that can occur in tumour growth. The aim of this study involved the detection of a TFs associated with BC using ChIP-Seq data. The analysis identified a potential relationship between the Oestrogen Receptor (ESR) 1 TF and oncogenesis using peak visualisation and motif analysis. IGV was used to study prior known BC genes such as TFF1-3. Following this, GREAT analysis could identify ontology associations such as those with FOX which was also identified the DREME analysis. RSAT and MEME TOMTOM results also correlated, indicating FOX gene association with the BC cell line. GREAT results indicated an association with the TFF genes which are known to be associated with BC development.

The most significant associations however were found using MEME-ChIP motif analysis to detect motifs associated with ESR1. The motif complexity and association indicates that ESR1 may be the TF with the regulatory role in the BC cell line. The transcriptionally active ESR1 can trigger endocrine therapy resistance and metastatic progression which is concordant with our peak analysis results (Lei et al., 2018). Although this appears to be the best match, the level of reads is still considerably low. This may be due to low antibody specificity but may also suggest the possibility of other transcription factors performing the regulation. The ESR1 appears the best candidate considering peak and motif analysis and due to its evident association with metastasis in breast cancer.