Анализ данных MNase seq

Кожевникова Дарья

Авг - Сен 2020

Цели и задачи

Цели:

- Научиться обрабатывать данные MNase seq

- Сопоставить результаты анализа с имеющимся теоретическим знанием

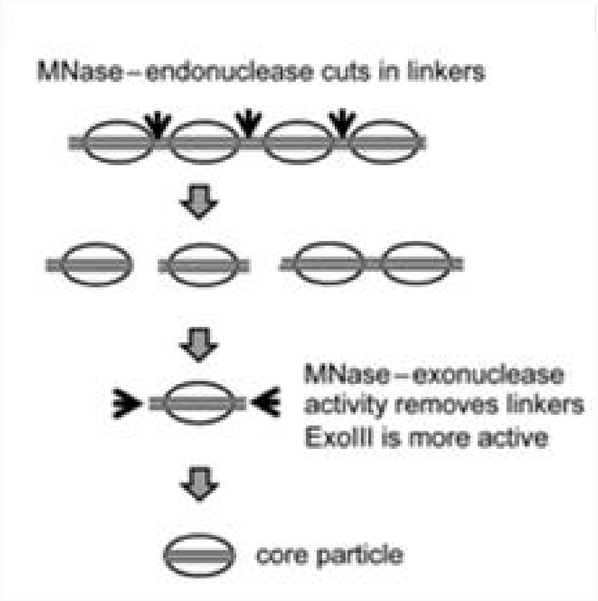
Задачи:

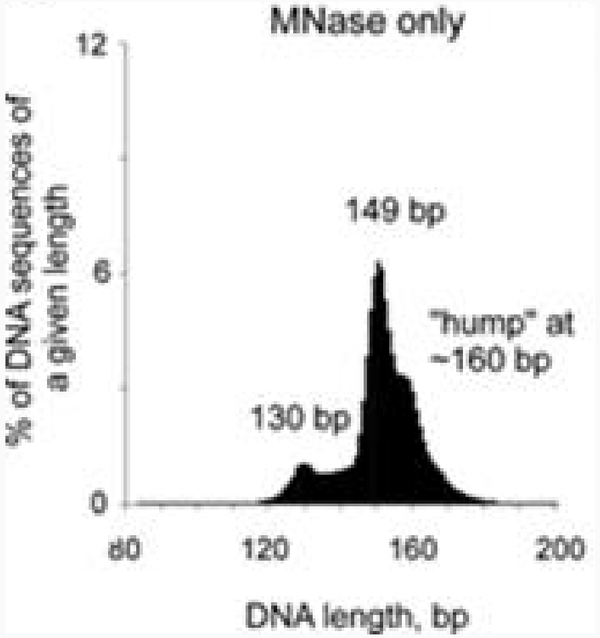
- Сделать обзор литературы
- Скачать данные из статьи Feng Cui, Hope A. Cole, David J. Clark, Victor B. Zhurkin, Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing, Nucleic Acids Research, Volume 40, Issue 21, 1 November 2012, Pages 10753–10764.
- Обработка данных с использованием инструментов биоинформатики
- Формулировка выводов

MNase seq data

The most frequently used method of mapping nucleosome positions and occupancy involves:

- Digestion of chromatin with micrococcal nuclease (MNase), an endo- and exo-nuclease that preferentially digests the naked DNA between nucleosomes, releases the nucleosomes from chromatin, and enriches the nucleosome-protected DNA fragments.
- To determine nucleosome positions and occupancy, the resulting undigested DNA is subjected to high throughput sequencing (MNase-seq).
- Mapping reads to the reference genome.





by Cole HA, Cui F, Ocampo J, et al

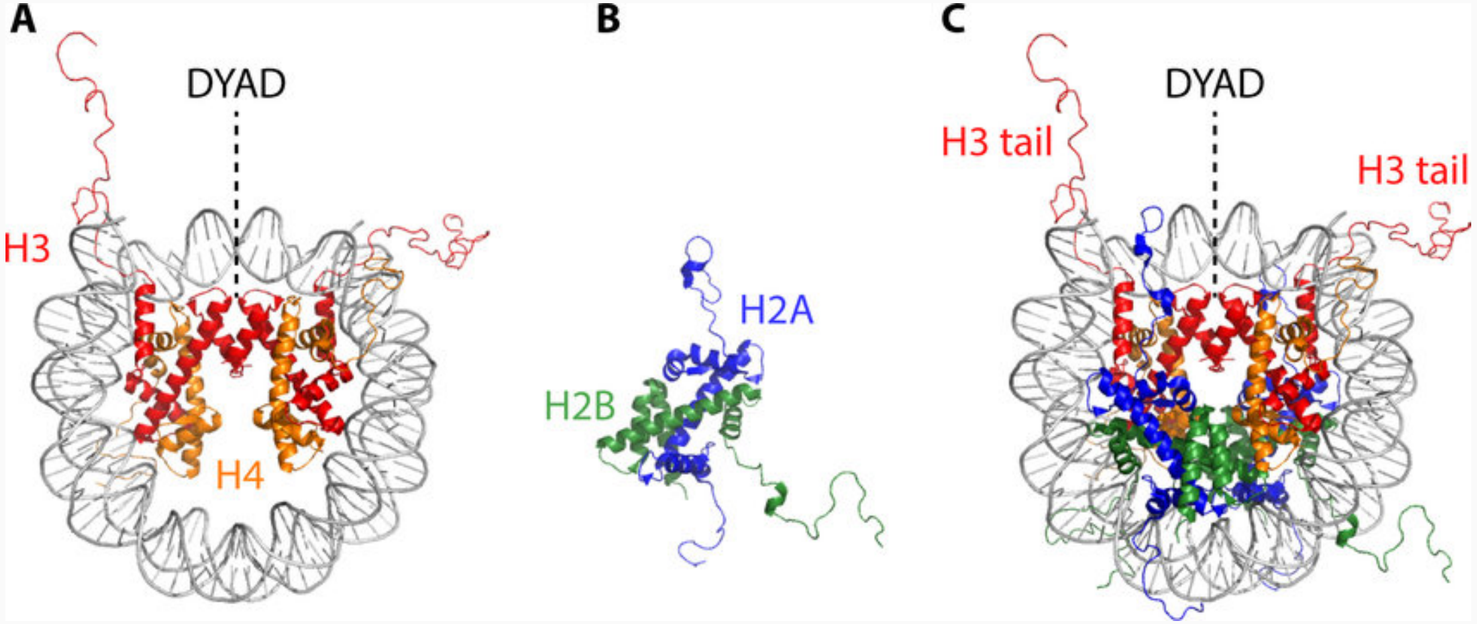
Feng Cui, Hope A. Cole, David J. Clark, Victor B. Zhurkin,

Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing

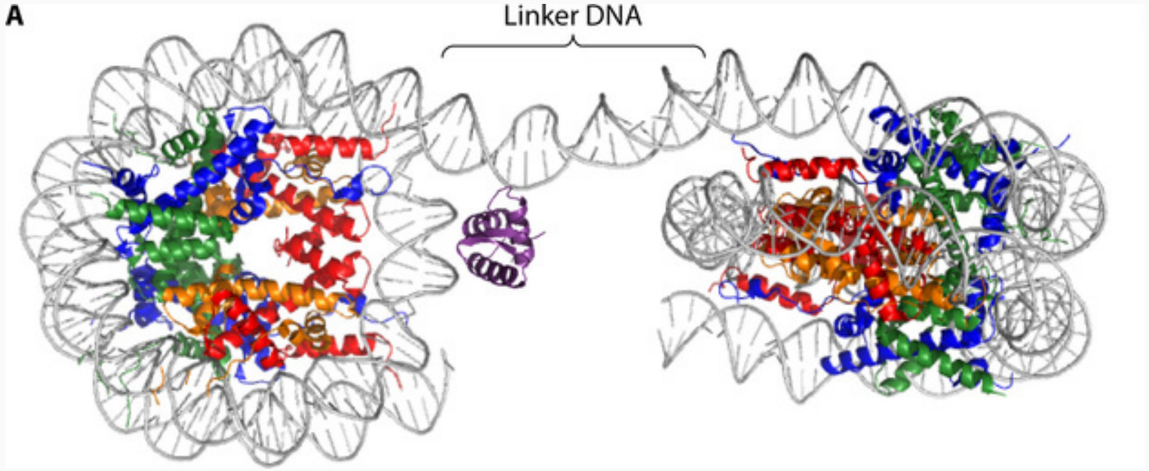
Nucleic Acids Research, Volume 40, Issue 21, 1 November 2012, Pages 10753–10764,
https://doi.org/10.1093/nar/gks870

Nucleosomes often undergo extensive rearrangement when genes are activated for transcription.
It was shown previously, using paired-end sequencing of yeast nucleosomes, that major changes in chromatin structure occur when genes are activated by **3-aminotriazole (3AT),** an inducer of the **transcriptional activator Gcn4.**
At the genomic level, **nucleosomes are regularly phased** relative to the transcription start site.
However, for a subset of 234 strongly induced genes, this **phasing is much more irregular after induction,** consistent with the loss of some nucleosomes and the re-positioning of the remaining nucleosomes.

DAC analysis of the 3AT-induced genes suggests that transcription activation coincides with rearrangement of nucleosomes into **irregular arrays with longer spacing.**
Sequence analysis of the +1 nucleosomes belonging to the 45 most strongly activated genes reveals **a distinctive periodic oscillation in the A/T-dinucleotide occurrence** that is present throughout the nucleosome and extends into the linker. This unusual pattern suggests that the **+1 nucleosomes might be prone to sliding,** thereby facilitating transcription.



Assembly of the nucleosome core particle. (A) Association of the (H3/H4)2 tetramer with DNA nucleates nucleosome ssembly and defines the dyad axis. (B) H2A/H2B dimer. (C) Two H2A/H2B dimers are deposited to generate the nucleosome ore particle. H3 N-terminal tails emerge near the DNA entry/exit points (based on data reported under PDB accession number 1KX5).



Histone H1 associates with linker DNA. (A) The globular domain of histone H1 (purple) binds the nucleosome at the dyad.

Nucleosome positioning

Translational positioning

- defined by the nucleosome midpoint (or dyad) with regard to the DNA sequence
- the DNA sequence patterns specifying translational nucleosome positioning are far from clear. The only well-established feature is the tendency of long A/T-rich fragments, and the A-tracts in particular, to be excluded from nucleosomes
- nucleosome positioning can be affected by DNA-binding transcription factors and chromatin remodeling enzymes

Rotational positioning

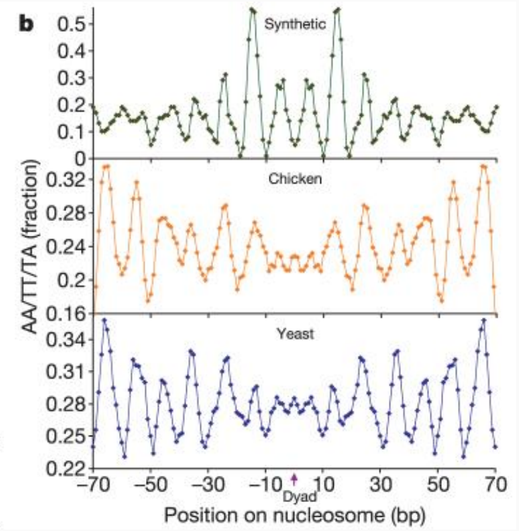
- defined by the side of the DNA helix that faces the histones
- related to the sequence-dependent preferences for DNA deformation, e.g. bending:

In particular, the **A/T**-containing dimeric steps AA:TT, AT and TA preferentially occur where the DNA is bent into the **minor groove,** while **G/C**-containing dimers GG:CC, GC and CG are frequently situated at the sites where DNA is bent toward the **major groove.** The occurrences of AT and GC dimers in nucleosome core DNA both display **sinusoidal patterns** with **~10-bp periodicity**, but they are **~5 bp out of phase** with one another. These sequence patterns are observed in nucleosomal DNA from chicken, yeast, fruit fly, nematode and human, indicating that the sequence rules for rotational positioning are essentially **the same across species.**

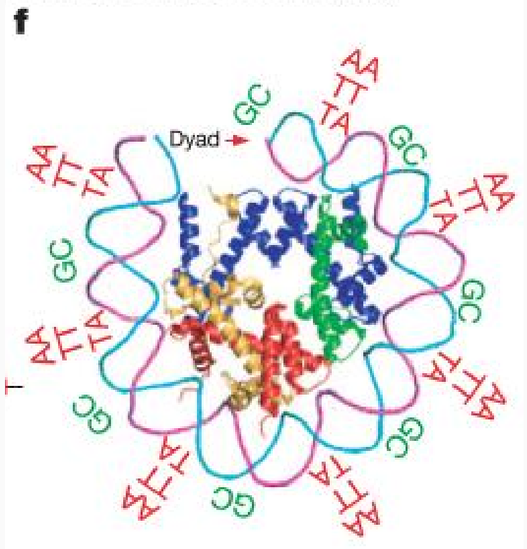
Nucleosome positioning code

Segal E, Fondufe-Mittendorf Y, Chen L, et al.: "Our findings demonstrate that **eukaryotic genomes use a nucleosome positioning code,** and link the resulting nucleosome positions to specific chromosome functions."

As expected for a nucleosome–DNA interaction model, the resulting model exhibits distinctive sequence motifs that recurperiodically at the DNA helical repeat and are known to facilitatethe sharp bending of DNA around the nucleosome3. These include **10-bp periodic AA/TT/TA dinucleotides** that oscillate in phasewith each other and out of phase with, **10-bp periodic GC dinucleotides.**
Improvingthe agreement of **a sequence with these motifs increased its bindingaffinity to the nucleosome**, whereas changing the periodicity ordeleting the key motifs decreased that affinity.

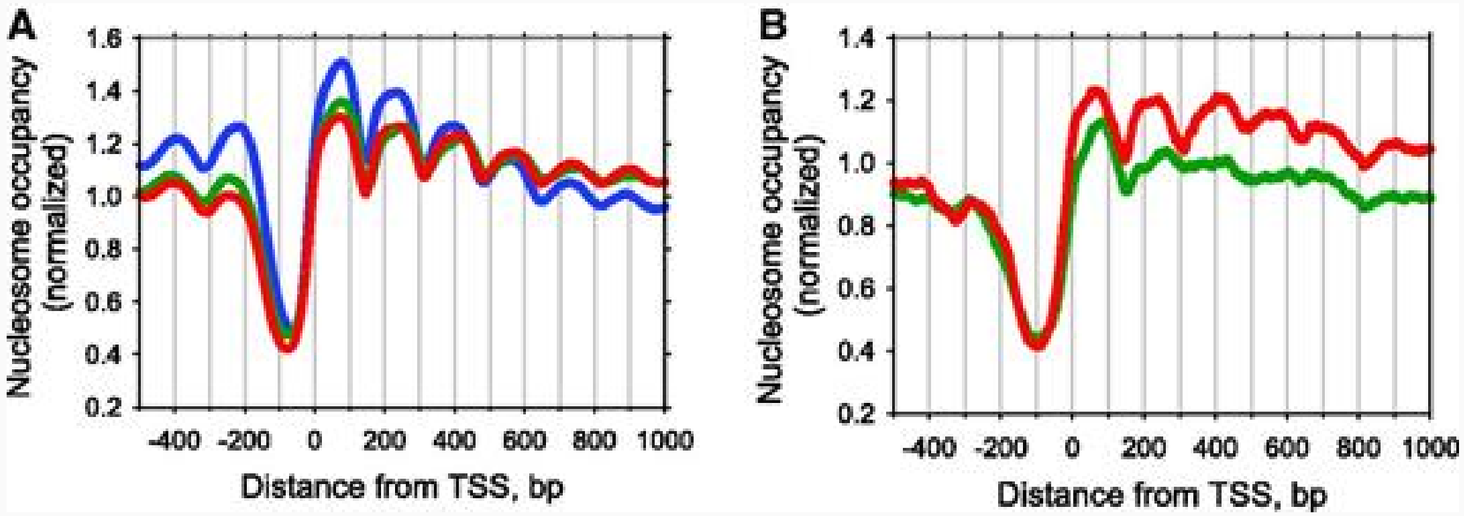


b, Fraction (3-bp moving average) of AA/TT/TAdinucleotides at each position of centre-aligned yeast, chicken or randomchemically synthesized nucleosome-bound DNA sequences, showing,10-bp periodicity of these dinucleotides.



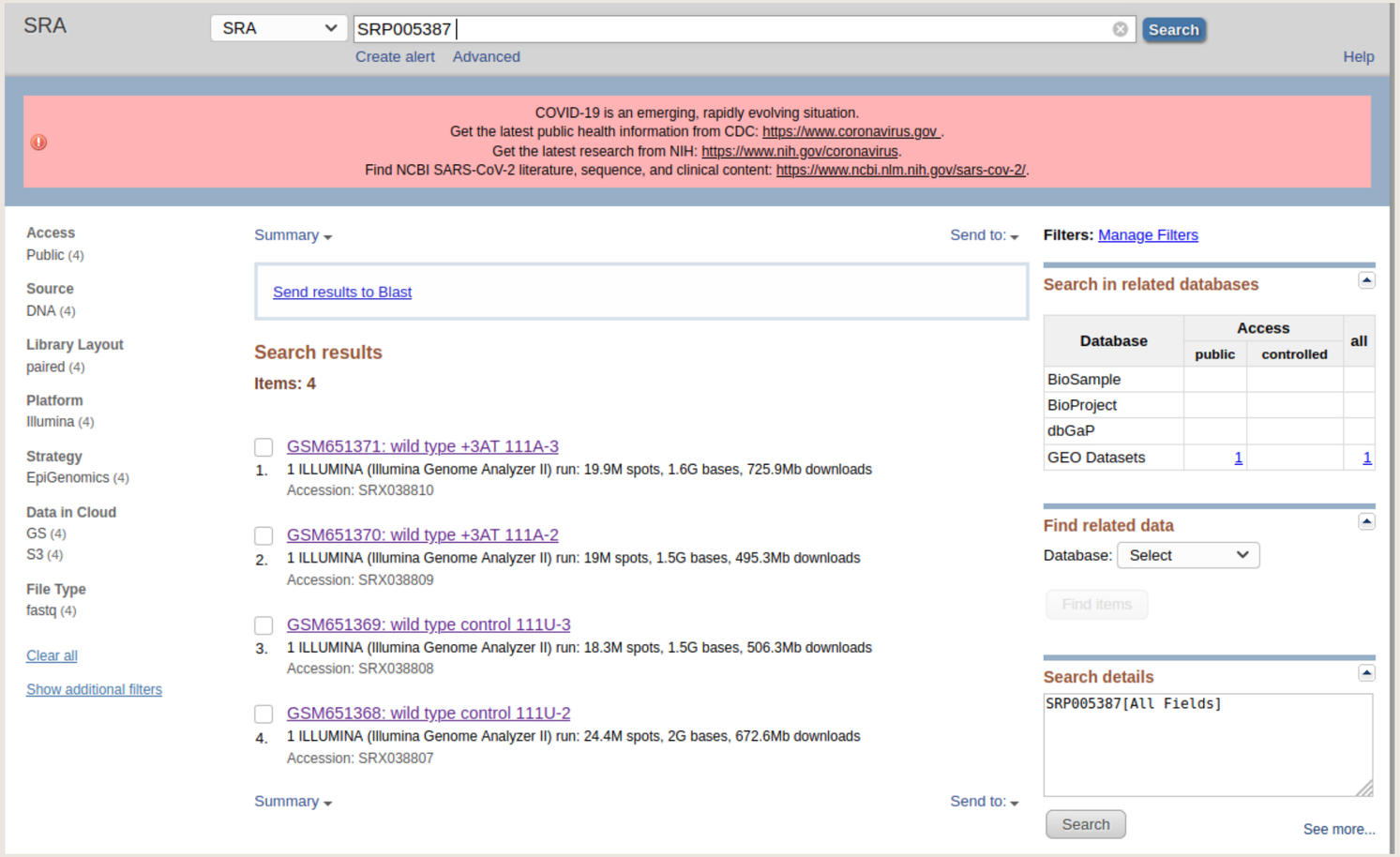
f, Key dinucleotidesinferred from the alignments are shown relative to the three-dimensionalstructure of one-half of the symmetric nucleosome.

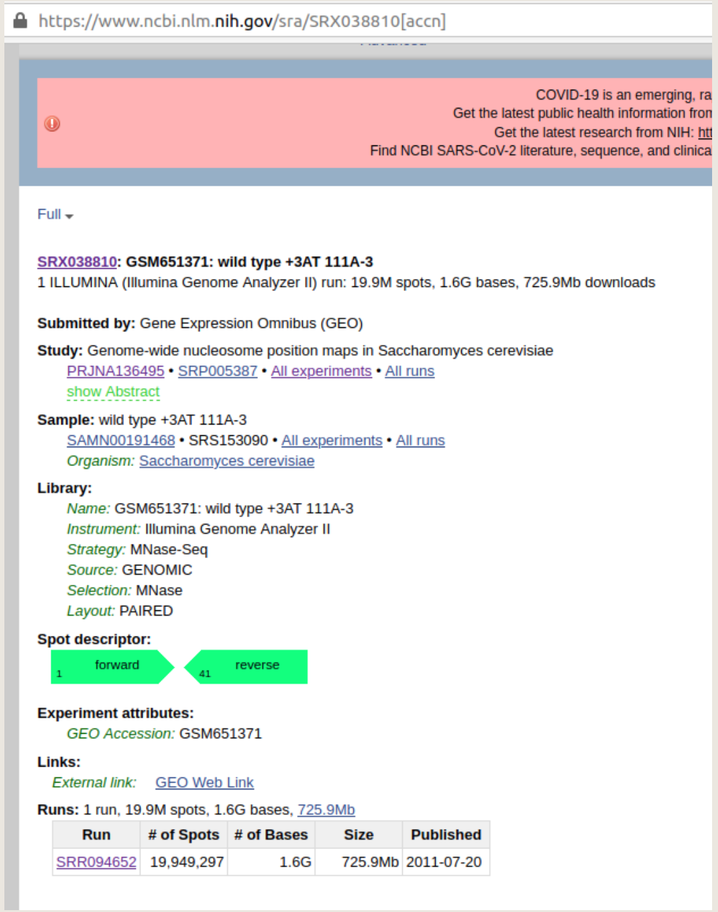
Local changes in nucleosome organization upon 3AT induction



Nucleosome organization around the 5′-end of yeast genes. (A) Overlays of nucleosome occupancy profiles of 4792 S. cerevisiae genes (30) relative to TSS (position 0). Nucleosome occupancy values are either taken directly from Kaplan et al. (29) (blue) or recalculated from Cole et al. (16 , 27): 3AT set (green) and CC set (red), respectively. In the latter two cases, the NCP fragments 147–152 bp in length were selected to calculate occupancy profiles.
(B) Nucleosome occupancy map for 234 genes (out of 4792 genes) that are induced by 3AT by more than 2-fold (25). Note that the occupancy value at each nucleotide is normalized by summing all the nucleosome sequences covering this nucleotide and dividing that number by the average number of nucleosome sequences per base pair across the genome.

SRR data from NCBI





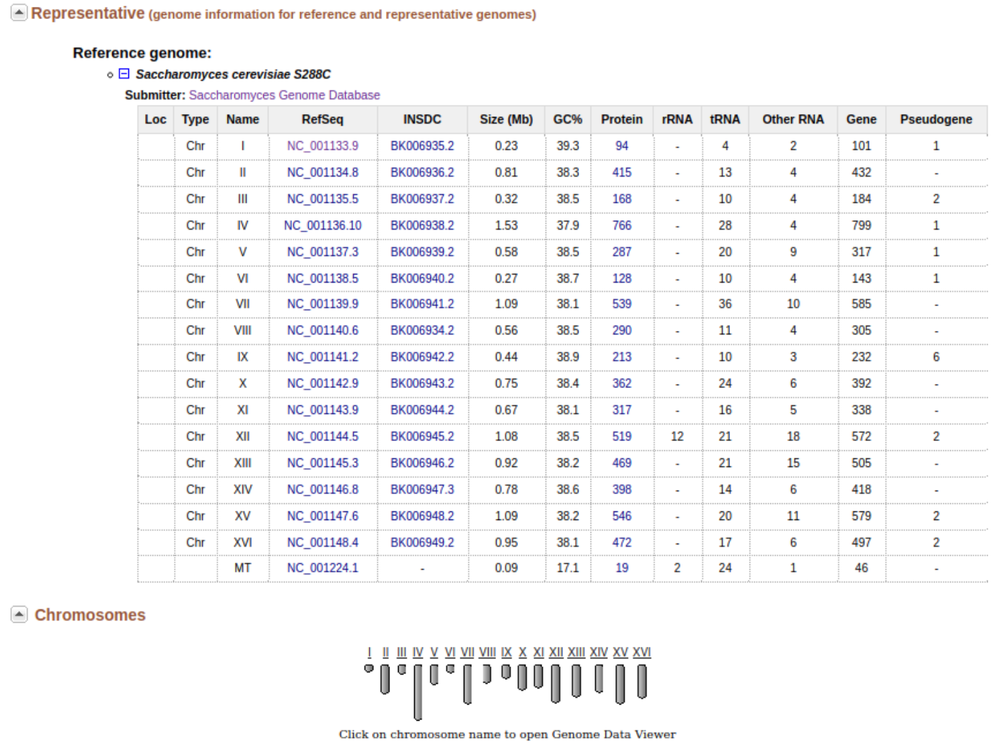
CC: SRR094649, SRR094650

3AT: SRR094651, SRR094652

SRR data processing

- Alignment:

Reference genome: Saccharomyces cerevisiae S288C (assembly R64)
Aligner: BWA (Burrows-Wheeler Aligner)
Tool: Jupiter Notebook: Enviroment:Genomics



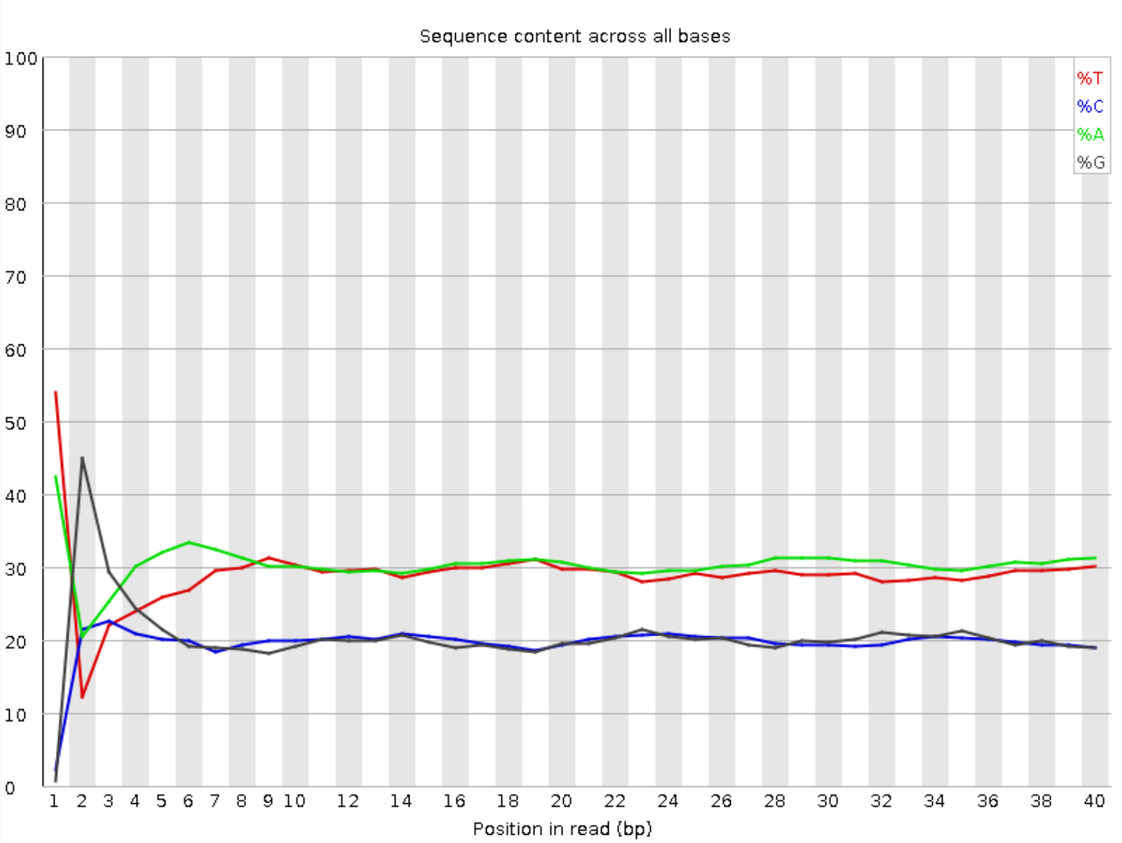
Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	I	NC_001133.9	BK006935.2	0.23	39.3	94	-	4	2	101	1
	Chr	II	NC_001134.8	BK006936.2	0.81	38.3	415	-	13	4	432	-
	Chr	III	NC_001135.5	BK006937.2	0.32	38.5	168	-	10	4	184	2
	Chr	IV	NC_001136.10	BK006938.2	1.53	37.9	766	-	28	4	799	1
	Chr	V	NC_001137.3	BK006939.2	0.58	38.5	287	-	20	9	317	1
	Chr	VI	NC_001138.5	BK006940.2	0.27	38.7	128	-	10	4	143	1
	Chr	VII	NC_001139.9	BK006941.2	1.09	38.1	539	-	36	10	585	-
	Chr	VIII	NC_001140.6	BK006934.2	0.56	38.5	290	-	11	4	305	-
	Chr	IX	NC_001141.2	BK006942.2	0.44	38.9	213	-	10	3	232	6
	Chr	X	NC_001142.9	BK006943.2	0.75	38.4	362	-	24	6	392	-
	Chr	XI	NC_001143.9	BK006944.2	0.67	38.1	317	-	16	5	338	-
	Chr	XII	NC_001144.5	BK006945.2	1.08	38.5	519	12	21	18	572	2
	Chr	XIII	NC_001145.3	BK006946.2	0.92	38.2	469	-	21	15	505	-
	Chr	XIV	NC_001146.8	BK006947.3	0.78	38.6	398	-	14	6	418	-
	Chr	XV	NC_001147.6	BK006948.2	1.09	38.2	546	-	20	11	579	2
	Chr	XVI	NC_001148.4	BK006949.2	0.95	38.1	472	-	17	6	497	2
		MT	NC_001224.1	-	0.09	17.1	19	2	24	1	46	-

SAM Format (Sequence Alignment Map):

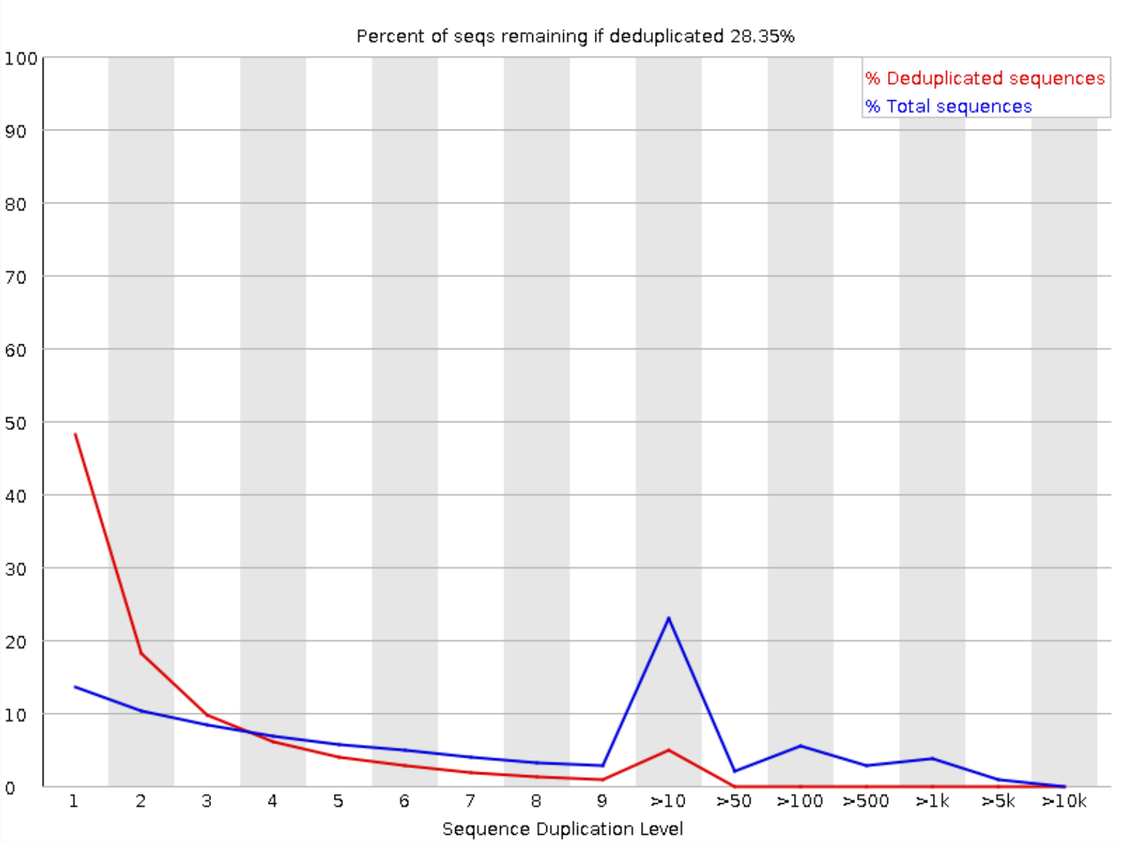
```
SRR094649.1     137     NC_001148.4     367277  37      40M     =       367277  0       ACNAGATTTTTTCAAATGGTGCAATAAACCAGATTG
TAAT    00!003<<;<CCCCCCCACCCCC@CCCCCBCCCCC@@C?C        XT:A:U  NM:i:1  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:1  XO:i:0
XG:i:0  MD:Z:2A37
SRR094649.2     69      NC_001140.6     247350  0       *       =       247350  0       TGGAAAGGTCCTTTGCTCTGTTNCTNNTNTCCTNNN
NNNN    CCCCCCCCCCCCCBCACCCC##!##!!#!####!!!!!!
SRR094649.2     137     NC_001140.6     247350  37      40M     =       247350  0       TGNAAATGAGCAAGTTATCAAAGACGTTCGTAAGAT
AANA    00!0097876CABCCACBCBBCCCCCCCCC5BBB####!#        XT:A:U  NM:i:2  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:2  XO:i:0
XG:i:0  MD:Z:2G35C1
SRR094649.3     117     NC_001144.5     638620  0       *       =       638620  0       NNNNNNNGATACCNNAGNTGTTCAAGTAGTTAACTA
ACCA    !!!!!!!######!!##!##C?CCCCC??@@CCCCC?CCC
SRR094649.3     153     NC_001144.5     638620  37      40M     =       638620  0       AAGGTTTGAAGTTGAGAAAGTACAAGGGTAACTTTA
CNGA    9<>49<AA:A?<9??0<<*9B)<BB9;>8>788;800!00        XT:A:U  NM:i:1  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:1  XO:i:0
XG:i:0  MD:Z:37C2
SRR094649.4     69      NC_001144.5     459272  0       *       =       459272  0       AAGGTGACAGGTTATGAAGATANGGNNCAAAAGNNA
NNNG    >?>>8?=>>/3:2?>=<3?###!##!!#######!!#!!!#
SRR094649.4     137     NC_001144.5     459272  0       40M     =       459272  0       TCNTCCCGTCATTATCGCCCGCATCCGGTGCCGTAA
AAGC    00!00:<:7:AAAA@@A:@=AA:=:A?:AA@':?>@@+?#        XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:2  X1:i:0  XM:i:2  XO:i:0
XG:i:0  MD:Z:2T34T2     XA:Z:NC_001144.5,+468409,40M,2;
SRR094649.5     83      NC_001145.3     567542  29      7S33M   =       567436  -139    ANNNTNNGAGCCANNCGNGGAATATGTTCCTTTAGA
AACA    #!!!#!!######!!##!##BBBAB?CCCCCCCCC:A?A=        XT:A:M  NM:i:3  SM:i:29 AM:i:29 XM:i:0  XO:i:0  XG:i:0  MD:Z:6T0
C2T22
SRR094649.5     163     NC_001145.3     567436  29      40M     =       567542  139     TANATTAGATTTGTGAAACTTCCATAATTTTTCGAA
AGGA    00!//88799BBABBA@BBB?A)AAB8??BB@BAB96:>9        XT:A:U  NM:i:1  SM:i:29 AM:i:29 X0:i:1  X1:i:0  XM:i:1  XO:i:0
XG:i:0  MD:Z:2A37
SRR094649.6     117     NC_001136.10    1210226 0       *       =       1210226 0       CNNNGTNTCTGAANTCTNTAGATTTACTGGCGTGCG
ACGA    #!!!##!######!###!###?6:?CC;CCA@?7@>B=<<
SRR094649.6     153     NC_001136.10    1210226 0       40M     =       1210226 0       GTTCCGCAGATAAGTGACCAAGAGACTGAGAAAAGG
ANTA    ABBA:3?9;?AA8BAA>?<??*<0?BB??<//00000!00        XT:A:R  NM:i:1  SM:i:0  AM:i:0  X0:i:21 X1:i:0  XM:i:1  XO:i:0
XG:i:0  MD:Z:37T2
SRR094649.7     99      NC_001144.5     280226  60      40M     =       280343  157     AGATGGACAACATTACAGAGAATTCCNATACACNAA
ACNG    ????87BB:ABAAAA6AA?A@;@@###!######!####!#        XT:A:U  NM:i:3  SM:i:25 AM:i:25 X0:i:1  X1:i:0  XM:i:3  XO:i:0
XG:i:0  MD:Z:26A6C4A1
SRR094649.7     147     NC_001144.5     280343  60      40M     =       280226  -157    CATAAATGTTTTAAGTATAACAAGAACACCAAATGT
TCCT    5>4@=AAA?A>;5>>36:/27:<66>>)7<@@@A@5A@@@        XT:A:U  NM:i:0  SM:i:37 AM:i:25 X0:i:1  X1:i:0  XM:i:0  XO:i:0
```

- # Quality check with FastQC

Per base sequence content:

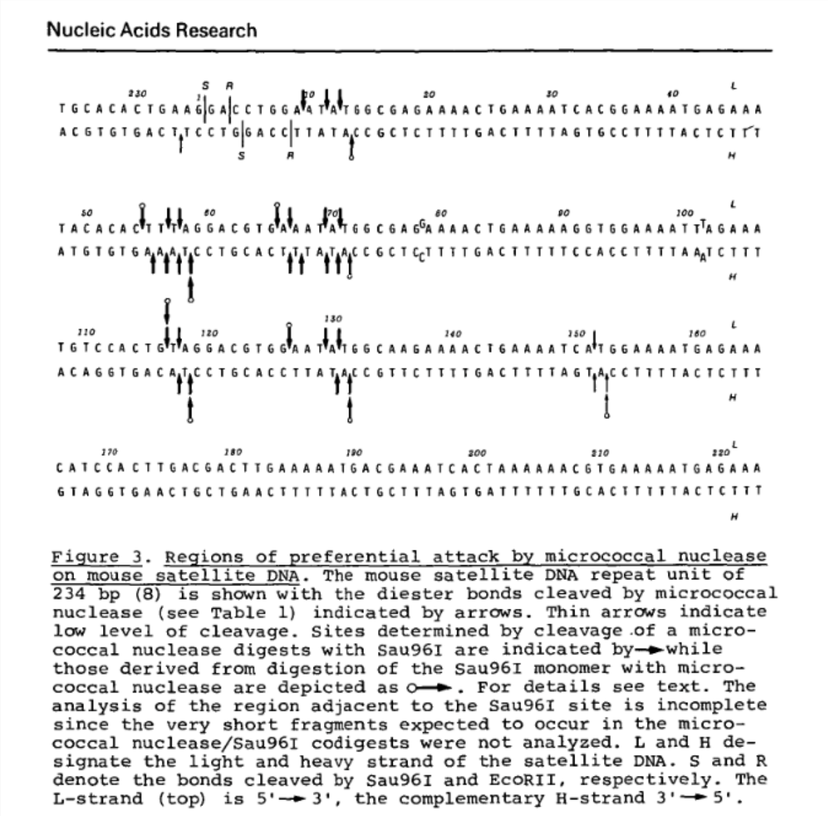


Sequence Duplication Levels:



Sequence specificity in the digestion of double-stranded DNA by micrococcal nuclease

'...it was concluded that thenuclease cleaves initially in AT-rich regions and it was sug-gested that this specificity depended on the greater conforma-tional motility of such regions in the DNA...



Nucleic Acids Research

Figure 3. Regions of preferential attack by micrococcal nuclease on mouse satellite DNA. The mouse satellite DNA repeat unit of 234 bp (8) is shown with the diester bonds cleaved by micrococcal nuclease (see Table 1) indicated by arrows. Thin arrows indicate low level of cleavage. Sites determined by cleavage of a micrococcal nuclease digests with Sau96I are indicated by ➞ while those derived from digestion of the Sau96I monomer with micrococcal nuclease are depicted as ○➞. For details see text. The analysis of the region adjacent to the Sau96I site is incomplete since the very short fragments expected to occur in the micrococcal nuclease/Sau96I codigests were not analyzed. L and H designate the light and heavy strand of the satellite DNA. S and R denote the bonds cleaved by Sau96I and EcoRII, respectively. The L-strand (top) is 5'➞3', the complementary H-strand 3'➞5'.

Wolfram Hörz, Werner Altenburger (1981):
"Sequences of the type **5′CATA and 5′CTA are attacked preferentially,** followed by exonucleotic degradation at the newly generated DNA termini. **GC-rich flanking sequences** further **increase** the probability of initial attack. Unexpectedly, **long stretches containing only A and T are spared** by the nuclease. These results, which were obtained with mouse satellite DNA and two fragments from the plasmid pBR322, do not support the previous contention that it is the regions of high AT-content which are initially cleaved by micrococcal nuclease. This specificity of micrococcal nuclease complicates its use in experiments intended to monitor the nucleoprotein structure of a DNA sequence in chromatin."

HTSeq: Analysing high-throughput sequencing data with Python

The various classes of HTSeq:

–Sequences and FASTA/FASTQ files

In order to represent sequences and reads (i.e., sequences with base-call quality information), the classes *Sequence* and *SequenceWithQualities* are used. The classes *FastaReader* and *FastqReader* allow to parse FASTA and FASTQ files.

–Genomic intervals and genomic arrays

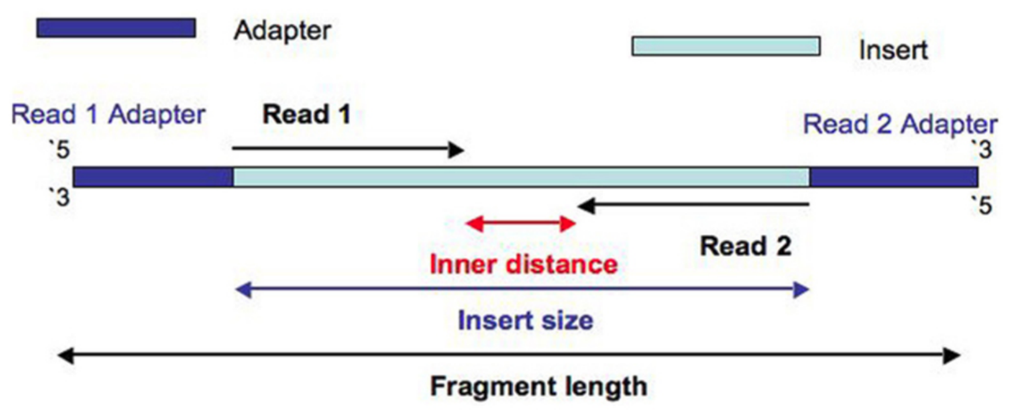
The classes *GenomicInterval* and *GenomicPosition* represent intervals and positions in a genome. The class *GenomicArray* is an all-purpose container with easy access via a genomic inter-val or position, and *GenomicArrayOfSets i*s a special case useful to deal with genomic features (suchas genes, exons, etc.)

–Read alignments

To process the output from short read aligners in various formats (e.g., SAM), the classes described hereare used, to represent output files and alignments, i.e., reads with their alignment information.

–Features

The classes *GenomicFeature* and *GFF_Reader* help to deal with genomic annotation data.

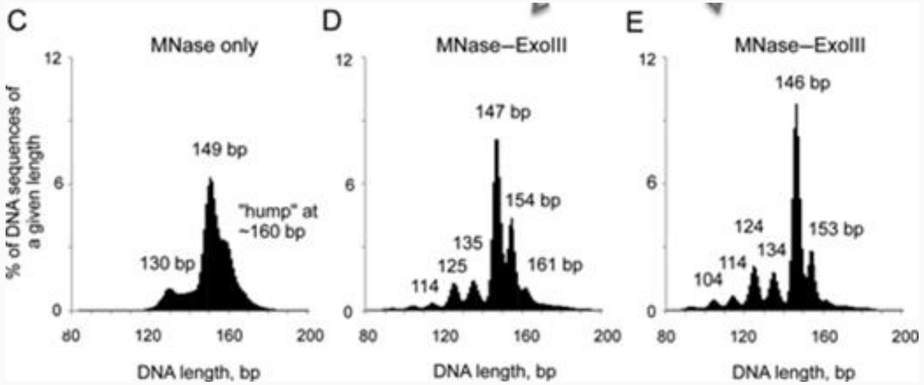


- Length distribution

Tool: Jupiter Notebook

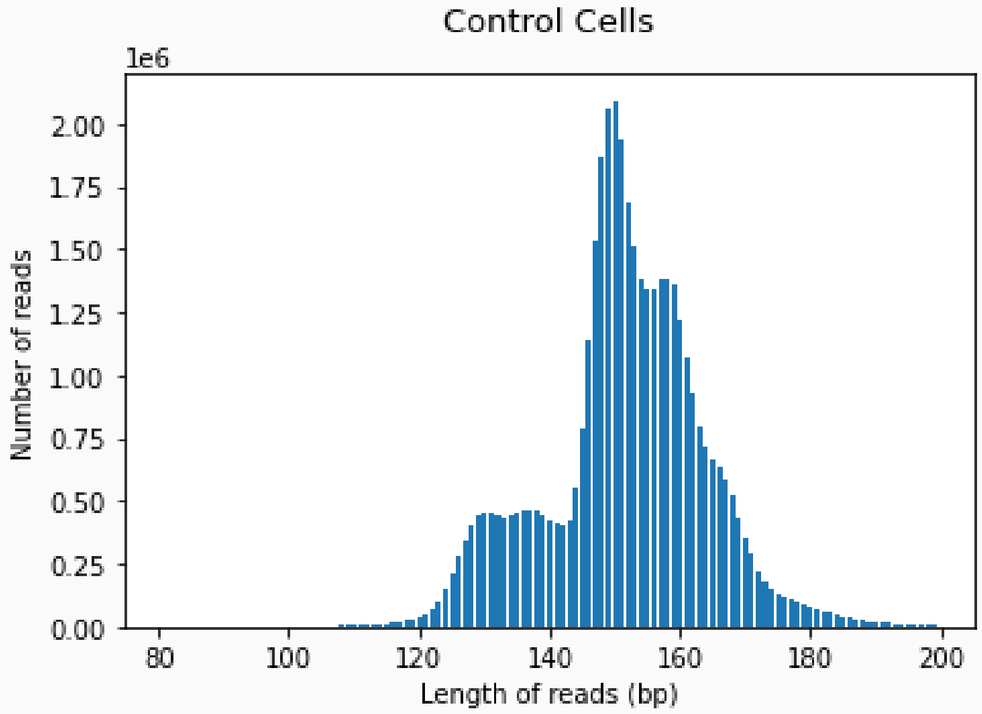
Python packages: HTSeq, pysam, matplotlib

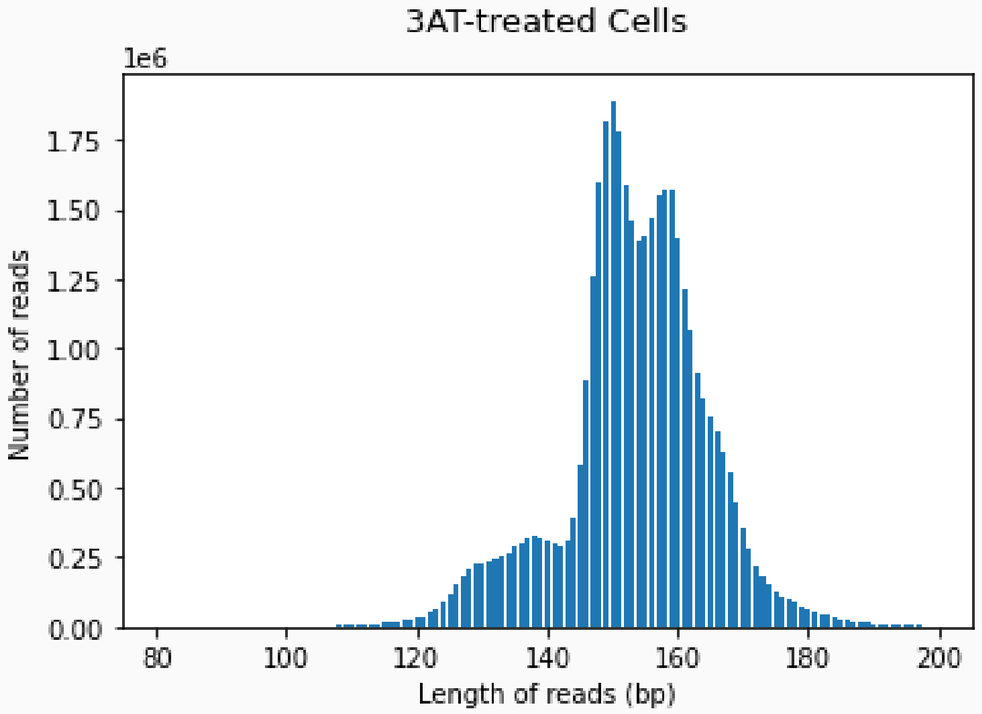
by Cole HA, Cui F, Ocampo J, et al



(C) Length distribution histograms for nucleosome sequences obtained by paired-end sequencing of the mono-nucleosome band from wild type strain JRY4012: MNase-only. (D, E) Length histograms for MNase-ExoIII at two different MNase concentrations

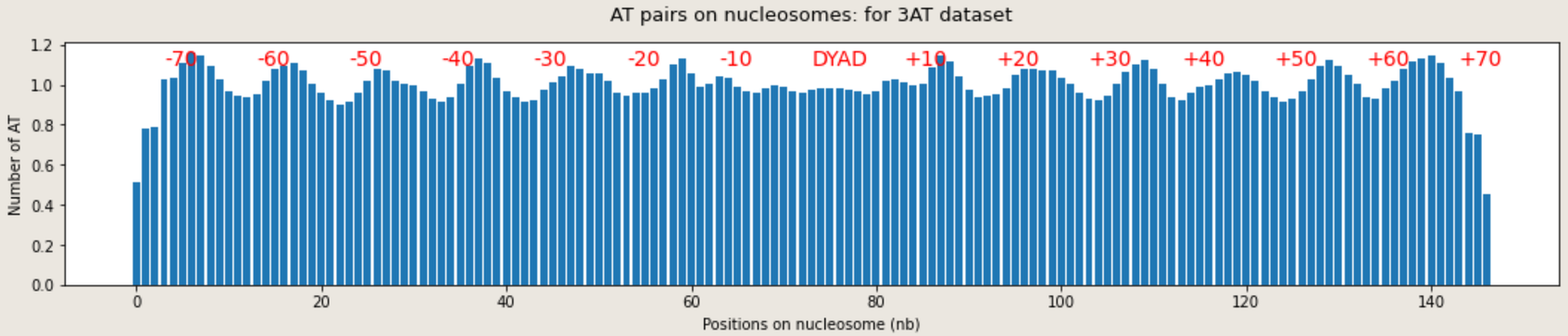
Peak on 150 length both in CC and 3AT sets:

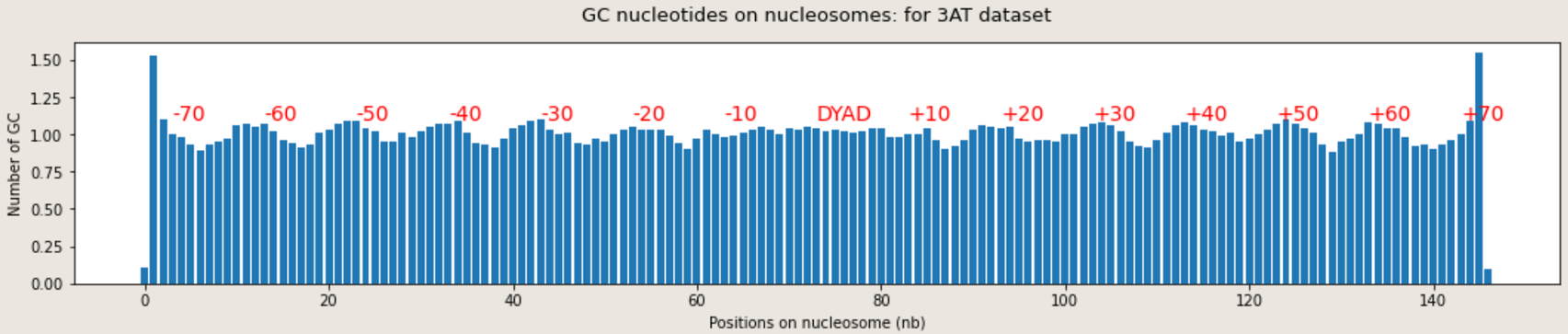




- A/T and G/C patterns

Done on dataset of reads with length 147





Conclusions

- Weak exonuclease activity of MNase
- AT and GC sinusoidal patterns

References

1. Cole HA, Cui F, Ocampo J, et al. Novel nucleosomal particles containing core histones and linker DNA but no histone H1. Nucleic Acids Res. 2016;44(2):573-581.
2. Feng Cui, Hope A. Cole, David J. Clark, Victor B. Zhurkin, Transcriptional activation of yeast genes disrupts intragenic nucleosome phasing, Nucleic Acids Research, Volume 40, Issue 21, 1 November 2012, Pages 10753–10764.
3. Segal E, Fondufe-Mittendorf Y, Chen L, et al. A genomic code for nucleosome positioning. Nature. 2006;442(7104):772-778. doi:10.1038/nature04979
4. Vladimir B. Teif, Nucleosome positioning: resources and tools online, Briefings in Bioinformatics, Volume 17, Issue 5, September 2016, Pages 745–757
5. Hope A. Cole, Bruce H. Howard, David J. Clark. The centromeric nucleosome of budding yeast is perfectly positioned and covers the entire centromere. Proceedings of the National Academy of Sciences Aug 2011, 108 (31) 12687-12692
6. Colin Dingwall, George P. Lomonossoff, Ronald A. Laskey, High sequence specificity of micrococcal nuclease, Nucleic Acids Research, Volume 9, Issue 12, 25 June 1981, Pages 2659–2674
7. Wolfram Hörz, Werner Altenburger, Sequence specific cleavage of DNA by micrococcal nuclease, Nucleic Acids Research, Volume 9, Issue 12, 25 June 1981, Pages 2643–2658