

对 NGS 数据的 ChIP-Seq 分析

生信 2001 张子栋

2023 年 6 月 15 日

摘要

目录

1	项目材料及目的	2
1.1	PubMed 数据库中搜索数据文献	2
1.1.1	选定研究物种	2
1.1.2	在 PubMed 中检索文献	2
2	数据与方法	2
2.1	NGS 数据获取	2
2.2	线虫参考基因组	2
2.2.1	WormBase	2
2.3	R 包及其他软件	3
2.3.1	FastQC	3
2.3.2	Trimmomatic	4
2.3.3	Bowtie2	5
2.3.4	Samtools	5
2.3.5	Bedtools	5
2.3.6	MACS2	6
2.3.7	MEME***	6
2.3.8	ChIPseeker	6
3	实验流程及结果	7
3.1	使用 FastQC 分析数据	7
3.2	使用 Trimmomatic 去除 adapter 序列	7
3.3	BWA 比对	8
3.4	Bowtie 比对	8
3.4.1	建立参考基因组索引	8
3.4.2	Bowtie 比对	8
3.5	Samtools 统计比对结果	9
3.5.1	Bowtie 比对结果	9
3.6	Peak Calling	9
3.7	Motif 分析	10
3.8	Peak 注释	10
4	结果与图表	10
4.1	FastQC Result	10
4.1.1	Original Data	10
4.1.2	Processed Data	10
4.2	Peak calling	10
4.3	Motif Analysis	10
4.4	Peak Annotation	10

目录	1
4.5 Gene Ontology	10

1 项目材料及目的

1.1 PubMed 数据库中搜索数据文献

1.1.1 选定研究物种

选定研究物种为线虫 (*nematode, Caenorhabditis Elegans*)。

1.1.2 在 PubMed 中检索文献

在 PubMed (nih.gov) 中搜索关键字 *Caenorhabditis Elegans ChIP-seq*. 并选定文献为 *The hypoxia response pathway promotes PEP carboxykinase and gluconeogenesis in C. elegans* (Nat Commun. 2022 Oct 18;13(1):6168. DOI: 10.1038/s41467-022-33849-x).

这篇文章研究的内容是在线虫中, 缺氧应答通路如何通过激活 HIF-1 转录因子来调节 PEP 羧激酶和糖异生的基因表达和代谢流动, 从而提高对氧化应激和缺氧应激的抵抗力。PEP 羧激酶是糖异生过程中的限速酶, 可以将草酰乙酸转化为磷酸烯醇式丙酮酸。

作者利用基因组编辑、转录组分析、代谢组分析和行为实验等方法来揭示 HIF-1 直接或间接调控的上百个基因的功能。这篇文章使用 ChIP-Seq 的目的是为了发现 HIF-1 直接调控的基因, 并分析它们在缺氧应答通路中的功能。

2 数据与方法

2.1 NGS 数据获取

文章中的 ChIP-seq 数据上传至 NIH/NCBI 数据库, 登录号为 GSE173333。SRA 数据库对应登录号为 SRP316378, 并最终选择 SRR14325856 和 SRR14325854 作为研究数据。

使用以下命令下载数据:

```
1 wget https://trace.ncbi.nlm.nih.gov/Traces/sra-reads-be/fastq?acc=SRR14325856
```

2.2 线虫参考基因组

2.2.1 WormBase

WormBase 是一个专门收录线虫的基因组信息的数据库, 它支持使用线虫作为模式生物的研究者, 提供了线虫的基因组序列、注释、变异、表达、互作等数据。WormBase 还包括一个子项目 WormBase ParaSite, 它收录了其他线虫和扁形动物寄生虫的基因组信息。参考基因组是一个完整的线虫基因序列, 基因组注释文件是对参考基因组中的基因、转录本、外显子等特征的描述。

通过 WormBase : Nematode Information Resource 网站获取线虫参考基因组。

2.3 R 包及其他软件

2.3.1 FastQC

FastQC 是一个质量控制分析工具，用于检测高通量测序数据中的潜在问题。它提供了一系列的分析模块，可以帮助快速了解数据是否有任何需要注意的问题，以便进行进一步的分析。FastQC 可以处理 fastq 或 bam 格式的原始序列文件，并生成一个报告，总结分析结果。

FastQC 的报告是一个 HTML 文件，包含了各种分析模块的结果和图表。FastQC 的报告中，每个分析模块都有一个结果图和一个状态图标。状态图标表示该模块的结果是否正常（绿色）、需要注意（黄色）或有问题（红色）。

FastQC 有以下几个分析模块：

- 基本统计：显示输入文件的名称、编码类型、总读数、读长和 GC 含量等信息。
- 每碱基质量分布：显示每个位置的平均质量得分，以及上下四分位数的范围。
- 每序列质量分布：显示每个序列的平均质量得分的频率分布，以及对应的合格率。
- 每碱基序列内容：显示每个位置的 A、T、G 和 C 的比例，以及与理论值的偏差。
- 每序列 GC 含量：显示每个序列的 GC 含量的频率分布，以及与整体 GC 含量的比较。
- 每碱基 N 含量：显示每个位置包含 N（未知）碱基的比例。
- 序列长度分布：显示不同长度序列出现的频率，以及平均长度和最大长度。
- 序列重复度：显示不同重复次数序列出现的频率，以及总重复度和最大重复度。
- 过表示 kmer 内容：显示在所有序列中过表达（出现次数超过预期）或者欠表达（出现次数低于预期）的 kmer（一般为 7 或 8 bp），以及它们在序列中出现的位置和比例。
- adapter 检测：检测输入文件中是否包含常见 adapter 序列，并显示它们在不同位置出现的比例。

FastQC 命令格式

```
1 fastqc [-o output dir] [--(no)extract] [-t thread num] [-f fastq|bam  
|sam] [-c contaminant file] seqfile1.. seqfileN
```

- 其中：
 - -o 用来指定输出文件的所在目录。
 - --(no)extract 用来控制是否解压缩输出的 .zip 文件。
 - -t 用来选择程序运行的线程数，即同时处理的文件数目。这样可以提高 fastqc 的运行速度，但也会占用更多的内存资源。
 - -f 用来强制指定输入文件格式，默认会自动检测。
 - -c 用来指定污染物文件，用于检测序列中是否含有不期望的序列。
 - seqfile1.. seqfileN 表示可以输入多个序列文件，支持 fastq, bam 和 sam 格式。

2.3.2 Trimmomatic

Trimmomatic 是一个快速的多线程命令行工具，于 2014 年首次发表在 Bioinformatics 期刊上，它可以用来整理和裁剪 Illumina (FASTQ) 数据以及删除 adapter。

Trimmomatic 有两种过滤模式，分别对应单末端 (SE) 和双末端 (PE) 测序数据，同时支持 gzip 和 bzip 压缩文件。它还支持 phred-33 和 phred-64 格式互相转化，目前多数 Illumina 测序数据为 phred-33 格式。

- Trimmomatic 的主要功能包括：
 - 去除 adapter 序列以及测序中其他特殊序列。
 - 采用滑动窗口的方法，切除或者删除低质量碱基。
 - 去除头 (尾) 部低质量以及 N 碱基过多的 reads。
 - 截取固定长度的 reads。
 - 丢掉小于一定长度的 reads。
 - phred 质量值转换。

Trimmomatic 命令格式 Trimmomatic 的命令格式根据使用的是双端模式 (PE) 还是单端模式 (SE) 而不同。一般来说，命令格式为：

```
1 java -jar <trimmomatic.jar> PE|SE [-threads <threads>] [-phred33|-  
    phred64] [-trimlog <logFile>] <input 1> [<input 2>] <output 1> [<  
    output 2>] <step 1> ...
```

- 其中：
 - `-jar <trimmomatic.jar>` 是运行 Trimmomatic 的 Java 命令，需要指定 `trimmomatic.jar` 文件的路径。
 - `PE|SE` 如果是双端模式，需要提供两个输入文件和两个输出文件；如果是单端模式，只需要提供一个输入文件和一个输出文件。
 - `[-threads <threads>]` 是可选参数，用于设置线程数，默认为 1。
 - `[-phred33-phred64]` 是可选参数，用于设置碱基质量值的编码系统，默认为 phred64。自 v0.32 版本之后，Trimmomatic 可以自动识别是 phred33 还是 phred64。
 - `[-trimlog <logFile>]` 是可选参数，用于设置日志文件的路径和名称。日志文件记录了每个 reads 的修剪情况。
 - `<input 1> [<input 2>]` 是输入文件的路径和名称，可以是压缩或非压缩的 FASTQ 文件。如果是双端模式，需要提供两个输入文件；如果是单端模式，只需要提供一个输入文件。
 - `<step 1> ...` 是指定要执行的修剪步骤和参数。每个步骤之间用空格隔开。目前支持以下几种步骤：

- * ILLUMINACLIP: 去除接头序列
- * SLIDINGWINDOW: 滑动窗口裁剪低质量碱基
- * MAXINFO: 基于信息熵裁剪低质量碱基
- * LEADING: 去除开头低质量碱基
- * TRAILING: 去除结尾低质量碱基
- * CROP: 裁剪 reads 到指定长度
- * HEADCROP: 去除 reads 开头指定长度
- * MINLEN: 过滤掉小于指定长度的 reads

2.3.3 Bowtie2

Bowtie 是一个超快的，存储高效的短序列片段比对程序。它能够将短的 DNA 序列片段 (reads) 比对到人类基因组或其他较大的参考基因组上。它有两个版本：Bowtie 和 Bowtie2。Bowtie2 是 Bowtie 的升级版，能够比对更长的 reads，支持局部比对和剪接性比对。

Bowtie 的使用也需要先对参考基因组建立索引，然后再进行比对。比对的结果是一个 SAM 或 BAM 格式的文件，可以用 Samtools 进行后续的处理。

2.3.4 Samtools

Samtools 是一个用于处理 SAM 或 BAM 格式的比对结果的软件包，它可以实现以下功能：

- 转换 SAM 和 BAM 格式，例如：

```
1 samtools view -bS in.sam > out.bam
```

- 排序和合并 BAM 文件，例如：

```
1 samtools sort in.bam -o out.sorted.bam
```

- 统计比对结果，例如：

```
1 samtools flagstat in.bam
```

2.3.5 Bedtools

Bedtools 是一个用于处理基因组数据的软件，它可以对 BED、BAM、GFF 等格式的文件进行各种操作，如求交集、并集、覆盖度、分组统计等。Bedtools 的主要功能有：

- **genomecov**: 计算基因组的覆盖度，即某些特征覆盖了基因组的哪些部分。
- **groupby**: 对文件或流按照指定的列进行分组，并对另一列进行统计，类似于数据库的“group by”语句。
- **intersect**: 计算两个文件或流中的特征的交集，即哪些特征在两个文件或流中都存在。
- **merge**: 合并两个文件或流中的特征，即将有重叠的特征合并为一个特征。
- **sort**: 对文件或流按照某些列进行排序，以便于其他操作。

2.3.6 MACS2

MACS 是一款用于分析 ChIP-seq 数据的软件，它可以用来寻找转录因子或组蛋白修饰在基因组上的结合位点。MACS 的基本原理是利用 ChIP-seq 数据中的片段长度和富集度来估计结合位点的位置和显著性。

MACS 的使用需要安装 Python 和一些依赖包，具体的安装方法可以参考官方文档。MACS 的主要命令是 `macs2 callpeak`，它可以用来从 ChIP-seq 数据中调用结合位点。基本语法是：

命令格式

```
1 macs2 callpeak -t <ChIP-seq文件> -c <对照文件> -f <文件格式> -g <基  
   因组大小> -n <输出文件名> [其他参数]
```

- 其中：

- `-t` 和 `-c` 参数分别指定 ChIP-seq 文件和对照文件，它们可以是 BAM, SAM, BED 或 ELAND 格式。
- `-f` 参数指定文件格式。
- `-g` 参数指定基因组大小
- `-text` 参数指定输出文件名
- 其他参数可以根据需要调整，例如 `-text` 参数可以设置 p 值阈值，`-B` 参数可以输出 bedGraph 格式的信号强度文件等。

2.3.7 MEME***

MEME 是一款用于分析蛋白质，DNA 和 RNA 中的序列 Motif 的软件，它可以用来寻找转录因子结合位点的共有序列特征。MEME 的基本原理是利用最大期望算法（Expectation Maximization）来从一组序列中识别出重复出现的 Motif。MEME 提供了在线版和本地版。

命令格式：

```
1 meme <序列文件> -o <输出目录> [其他参数]
```

- 其中：

- `<序列文件>` 指定输入的序列文件，它可以是 FASTA 格式或 MEME 格式。
- `<输出目录>` 指定输出结果的目录，它必须不存在或为空。
- 其他参数可以根据需要调整，例如 `-nmotifs` 参数可以设置要发现的 Motif 个数，`-minw` 参数可以设置 Motif 的最小长度，`-maxw` 参数可以设置 Motif 的最大长度等。

2.3.8 ChIPseeker

ChIPseeker 是一个 R 包，用于 ChIP 峰值的注释、比较和可视化。它实现了检索峰值周围最近的基因、注释峰值的基因组区域、估计 ChIP 峰值数据集之间重叠的显著性的统计方法，并将 GEO 数据库纳入其中，以供比较。

命令格式

```
1 anno <- annotatePeak(peak, tssRegion=c(-3000, 3000), TxDb=txdb)
```

其中:

- TxDb 可以使用 `makeTxDbFromGFF()` 命令从基因组注释文件获得。
- peak 是 Peak Calling 产生的 .bed 文件。

3 实验流程及结果

3.1 使用 FastQC 分析数据

1. 创建目录用于存放输出结果: `mkdir fastqc.test`
2. 使用 FastQC 命令: `fastqc -o fastqc.test/ -t 4 SRR14325856.fastq.gz`
3. 输出文件:

```
1 fastqc.test/  
2 |-- SRR14325853_fastqc.html  
3 |-- SRR14325853_fastqc.zip  
4 |-- SRR14325854_fastqc.html  
5 |-- SRR14325854_fastqc.zip  
6 |-- SRR14325856_fastqc.html  
7 |-- SRR14325856_fastqc.zip
```

FastQC 结果:

- 每序列 GC 含量略微偏离正态分布。
- 过表达序列为 adapter.

3.2 使用 Trimmomatic 去除 adapter 序列

1. 在终端中输入以下命令, 文件路径均使用绝对路径:

```
1 java -jar /home/software/Trimmomatic-0.39/trimmomatic-0.39.jar  
SE -phred33 -trimlog test.log.txt /Bioinfo/ZhZidong/  
SRR14325856.fastq /Bioinfo/ZhZidong/SRR14325856.processed.  
fastq ILLUMINACLIP:/home/software/Trimmomatic-0.39/adapters/  
TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 MINLEN:25
```

2. 生成 `SRR14325856.processed.fastq` 文件。

- 根据 FastQC 结果:

- 每序列 GC 含量略微偏离正态分布。
- 序列长度分布集中于 49 bp，相比原数据，多了 25 bp 长度的序列。
- 多增的 25 bp 序列，可能是因为这个序列是接头或者其他污染物，并且与读段有足够高的匹配分数。

3.3 BWA 比对

3.4 Bowtie 比对

3.4.1 建立参考基因组索引

使用 Bowtie2 软件进行比对之前，也需要先建立索引，索引的作用和原理与 BWA 软件类似，都是为了加快比对的速度，减少内存的占用。建立索引的方法是将参考基因组分割成多个子序列，然后对每个子序列建立一个 Burrows-Wheeler 变换 (BWT) 和后缀数组 (SA)，记录每个 k-mer 的出现位置。比对的时候，Bowtie2 会将 reads 也分割成多个子序列，然后在 BWT 和 SA 中查找匹配的位置，从而找到最佳的比对位置。

具体命令：

```
1 bowtie2-build bowtie_index/c_elegans.PRJNA275000.WS286.genomic.fa  
   bowtie_index/c_elegans.PRJNA275000.WS286.genomic.bowtie
```

得到文件：

```
1 bowtie_index/  
2 |-- c_elegans.PRJNA275000.WS286.genomic.bowtie.1.bt2  
3 |-- c_elegans.PRJNA275000.WS286.genomic.bowtie.2.bt2  
4 |-- c_elegans.PRJNA275000.WS286.genomic.bowtie.3.bt2  
5 |-- c_elegans.PRJNA275000.WS286.genomic.bowtie.4.bt2  
6 |-- c_elegans.PRJNA275000.WS286.genomic.bowtie.rev.1.bt2  
7 |-- c_elegans.PRJNA275000.WS286.genomic.bowtie.rev.2.bt2  
8 |-- c_elegans.PRJNA275000.WS286.genomic.fa
```

3.4.2 Bowtie 比对

具体命令：

```
1 bowtie2 -p 10 -x ~/bowtie_index/c_elegans.PRJNA275000.WS286.genomic.  
   bowtie -U ~/SRR14325856.fastq -S ~/bowtie_result/c_elegans_ChIP-  
   Seq.sam
```

生成文件：c_elegans_ChIP-Seq.sam

3.5 Samtools 统计比对结果

3.5.1 Bowtie 比对结果

具体命令：

```
1 samtools view -S -b ~/bowtie_result/c_elegans_ChIP-Seq.sam |  
  samtools flagstat - > ~/bowtie_result/flagstat.txt
```

输出结果：

```
1 6492282 + 0 in total (QC-passed reads + QC-failed reads)  
2 0 + 0 secondary  
3 0 + 0 supplementary  
4 0 + 0 duplicates  
5 3405738 + 0 mapped (52.46\% : N/A)  
6 0 + 0 paired in sequencing  
7 0 + 0 read1  
8 0 + 0 read2  
9 0 + 0 properly paired (N/A : N/A)  
10 0 + 0 with itself and mate mapped  
11 0 + 0 singletons (N/A : N/A)  
12 0 + 0 with mate mapped to a different chr  
13 0 + 0 with mate mapped to a different chr (mapQ>=5)
```

3.6 Peak Calling

利用计算的方法找到 ChIP-seq 或 ATAC-seq 中 reads 富集的基因组区域。

具体命令：

```
1 macs2 callpeak -t ~/bwa_result/c_elegans_ChIP-Seq.sam -f SAM -g ce -  
  n test
```

生成文件：

```
1 peak_calling/  
2 |-- motif_analysis  
3 |-- test_model.pdf  
4 |-- test_model.r  
5 |-- test_peaks.narrowPeak  
6 |-- test_peaks.xls  
7 |-- test_summits.bed
```

3.7 Motif 分析

首先需要获得 fasta 格式的 peak 文件，使用 Bedtools：

```
1 bedtools getfasta -fi ~/bwa_index/c_elegans.PRJNA275000.WS286.
   genomic.fa -bed ~/peak_calling/test_peaks.narrowPeak -fo ./peak.
   fasta
```

由于在线版 MEME 需要排队，以及运行时间长，故使用 TBtools 中内嵌的本地版 MEME.

3.8 Peak 注释

使用 R 包 ChIPseeker 进行 peak 注释。

具体命令：

```
1 library(ChIPseeker)
2 library(GenomicFeatures)
3 txdb <- makeTxDbFromGFF(file = file.choose(), format = "gff3")
4 peak <- readPeakFile(file.choose())
5 peakAnno <- annotatePeak(peak,
6                           TxDb=txdb,
7                           tssRegion=c(-1000, 1000))
```

导出 Gene List 用于进一步的 Gene Ontology 分析。

```
1 df <- as.data.frame(peakAnno)
2 gene <- df[,14]
3 write.table(gene, file = "D:\\gene_list.txt", sep = "", row.names =
   FALSE, col.names = FALSE, quote = FALSE)
```

4 结果与图表

4.1 FastQC Result

4.1.1 Original Data

4.1.2 Processed Data

4.2 Peak calling

4.3 Motif Analysis

4.4 Peak Annotation

4.5 Gene Ontology

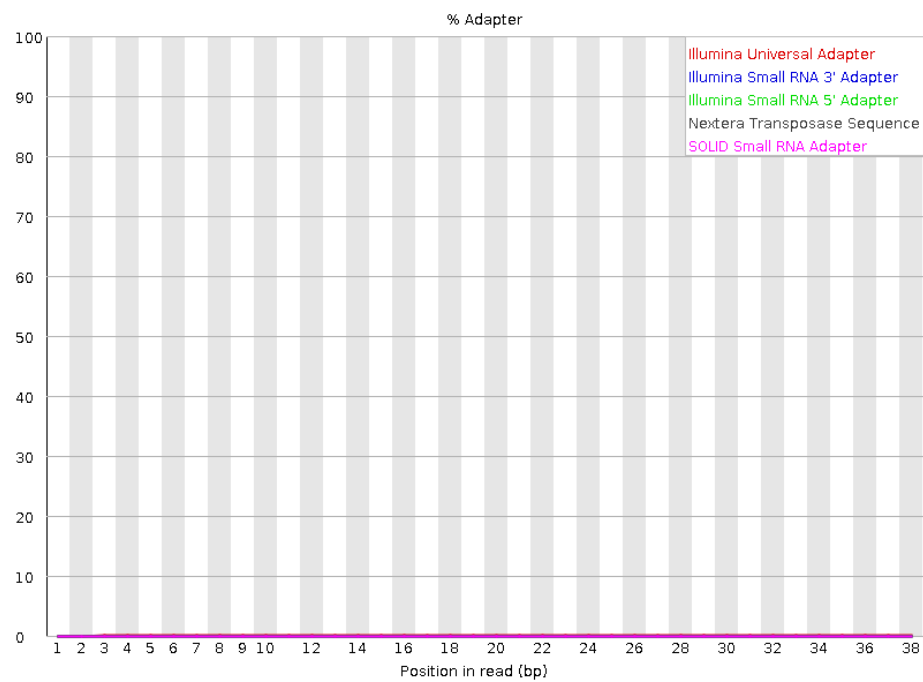


图 1: Adapter content

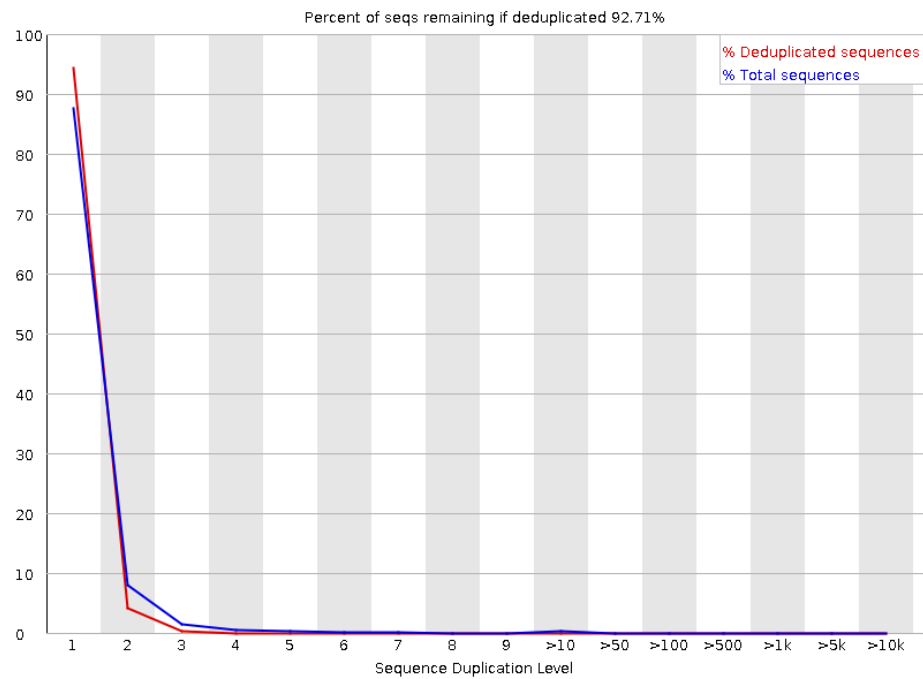


图 2: Duplication levels

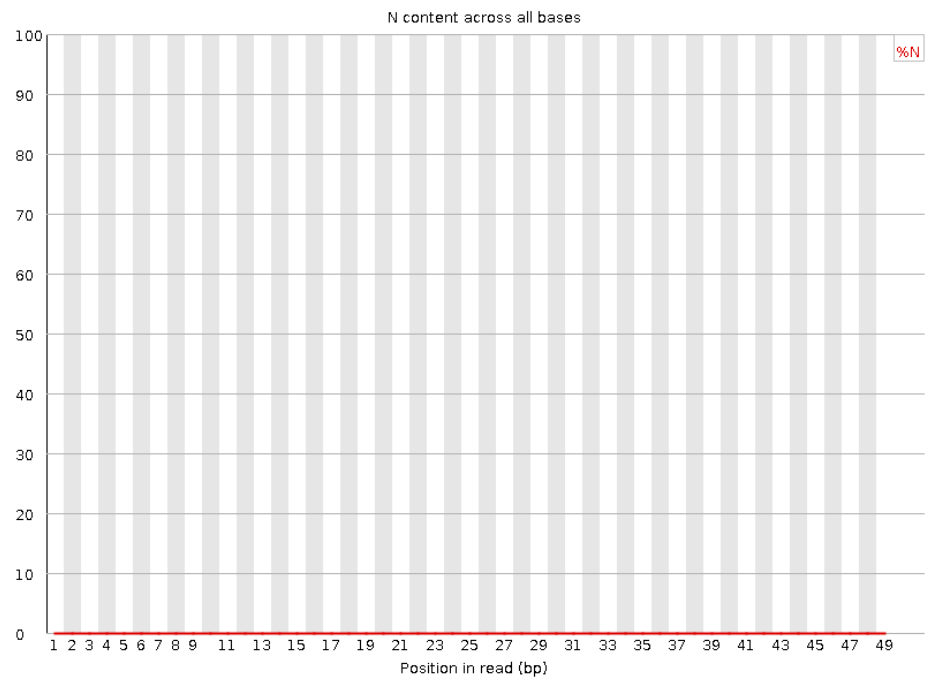


图 3: Per base N content

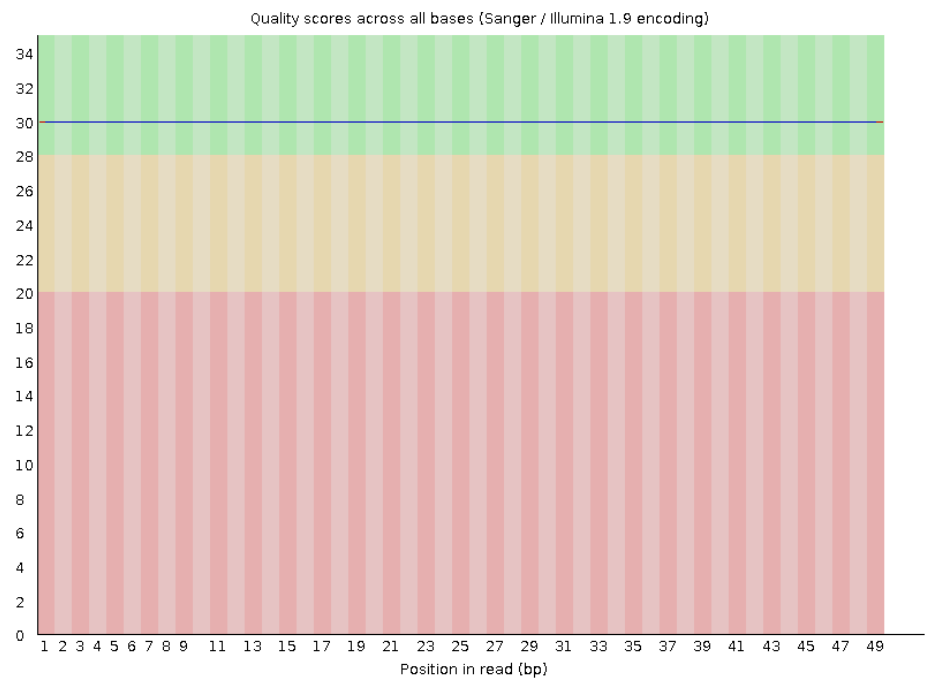


图 4: Per base quality

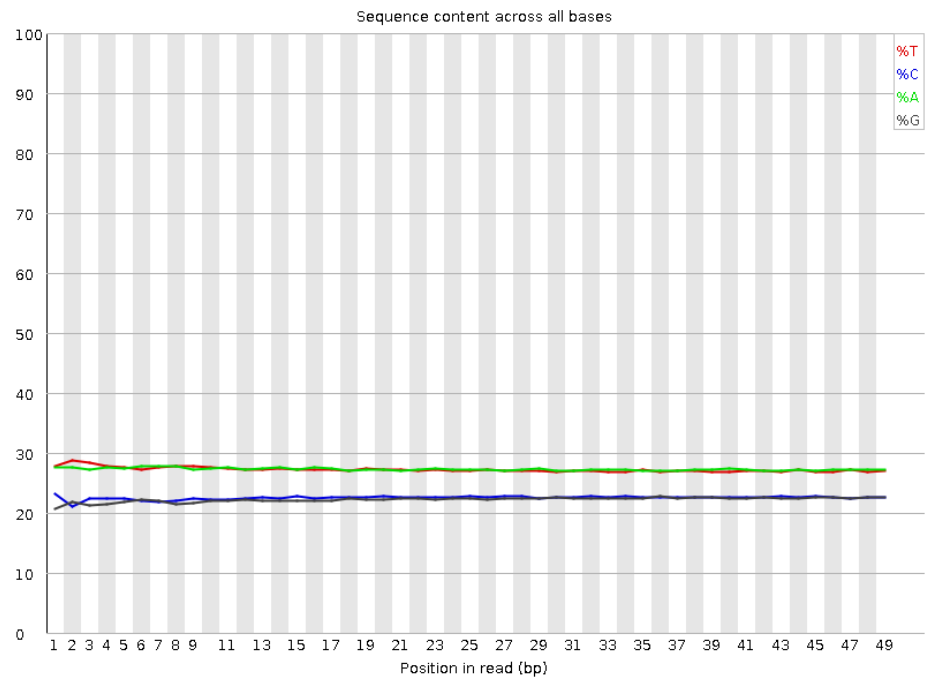


图 5: Per base sequence content

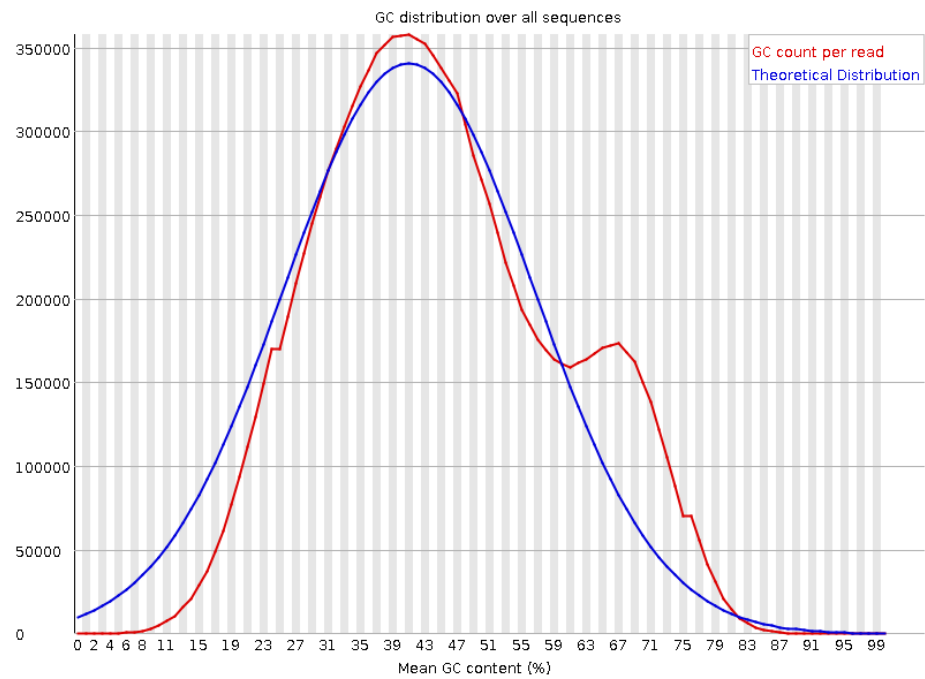


图 6: Per sequence GC content

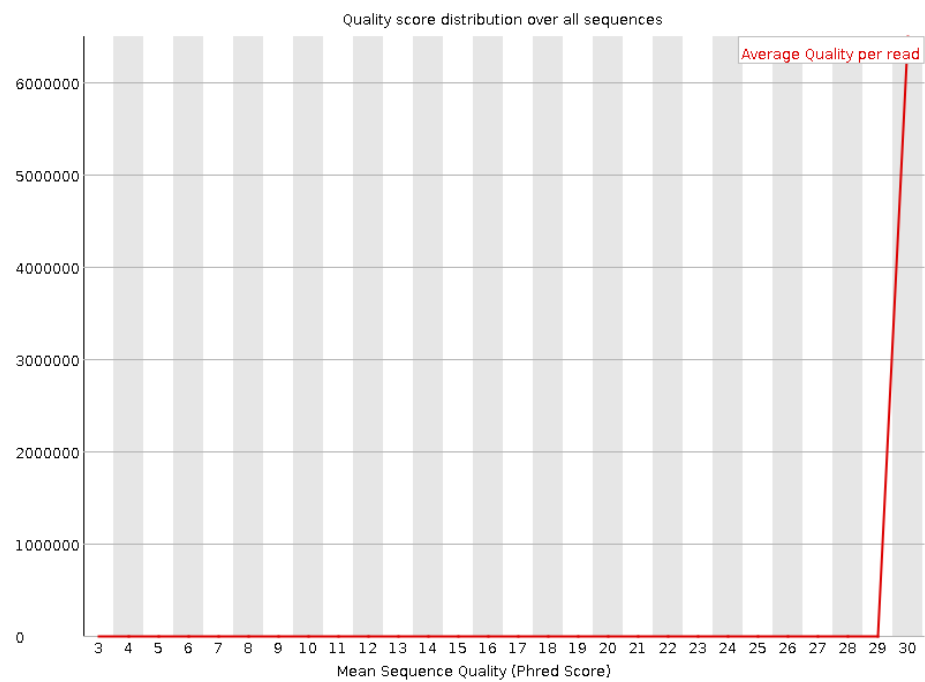


图 7: Per sequence quality

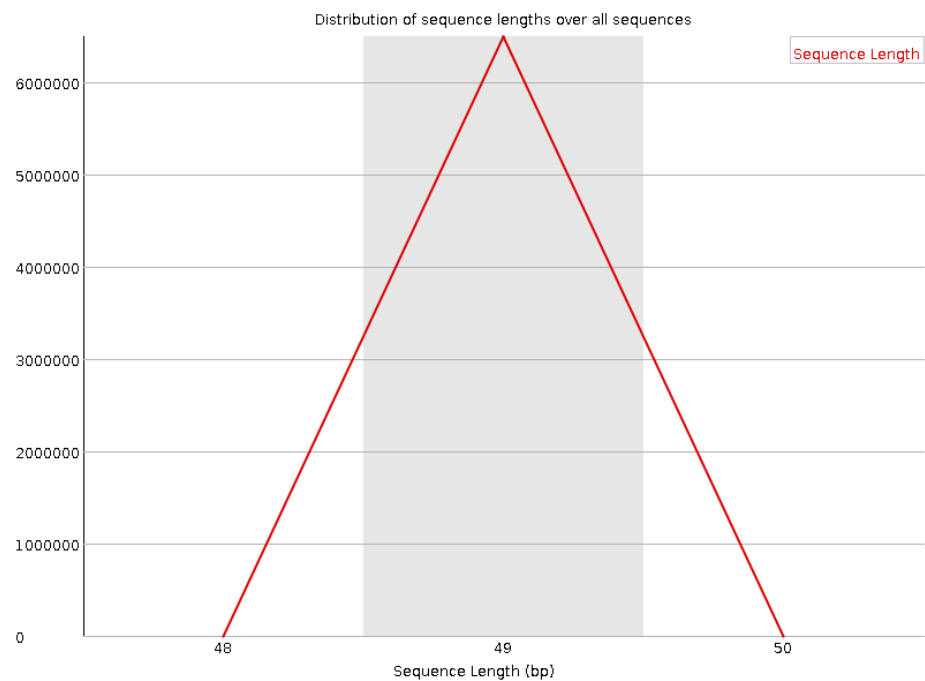


图 8: Sequence length distribution

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGTGAAATCTCGTAT	11589	0.17850426090548746	TruSeq Adapter, Index 13 (97% over 36bp)

图 9: Overrepresented sequences

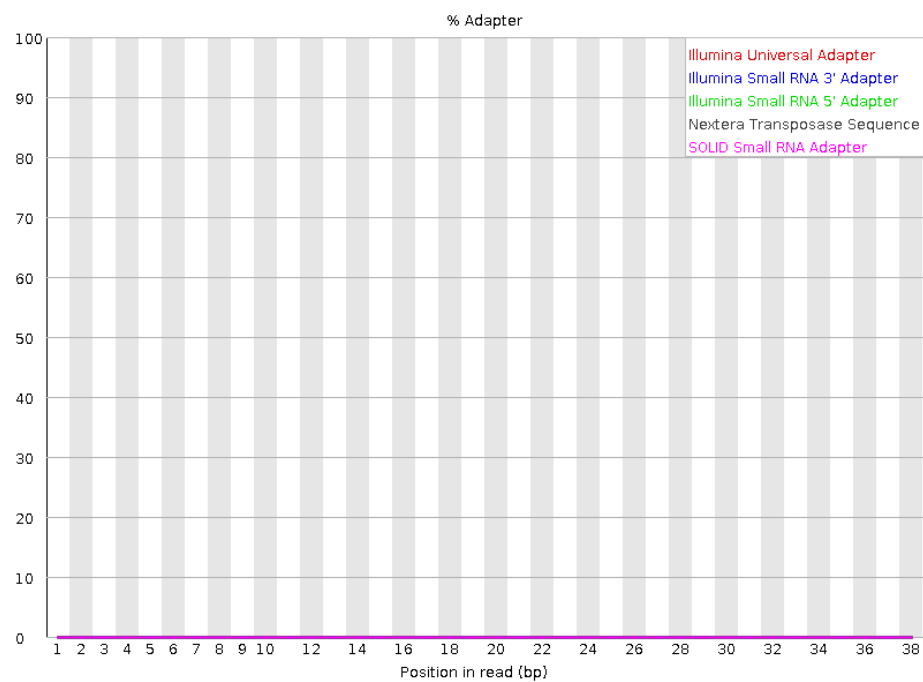


图 10: Adapter content (Processed)

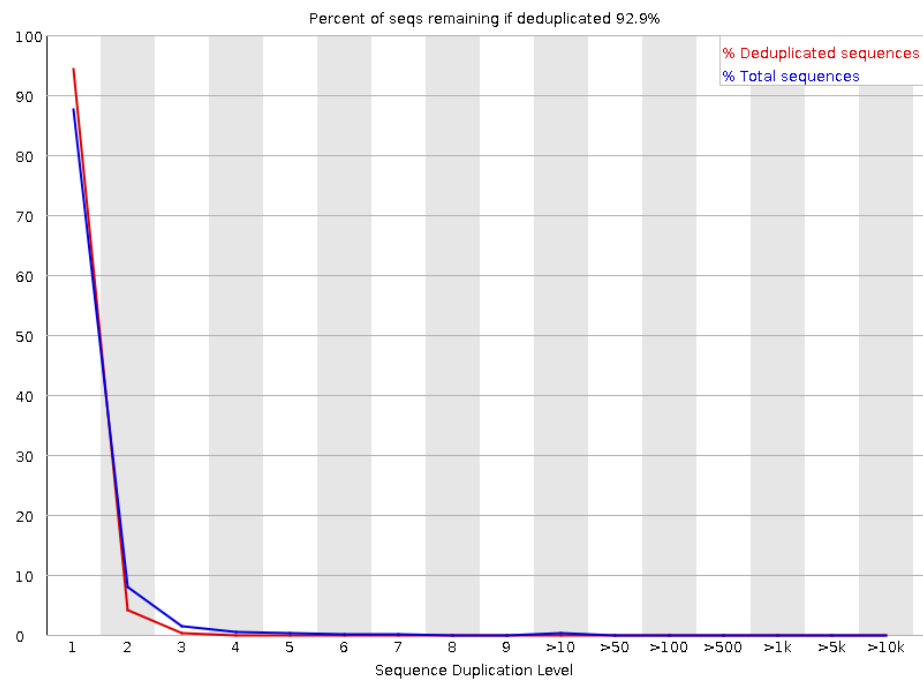


图 11: Duplication levels (Processed)

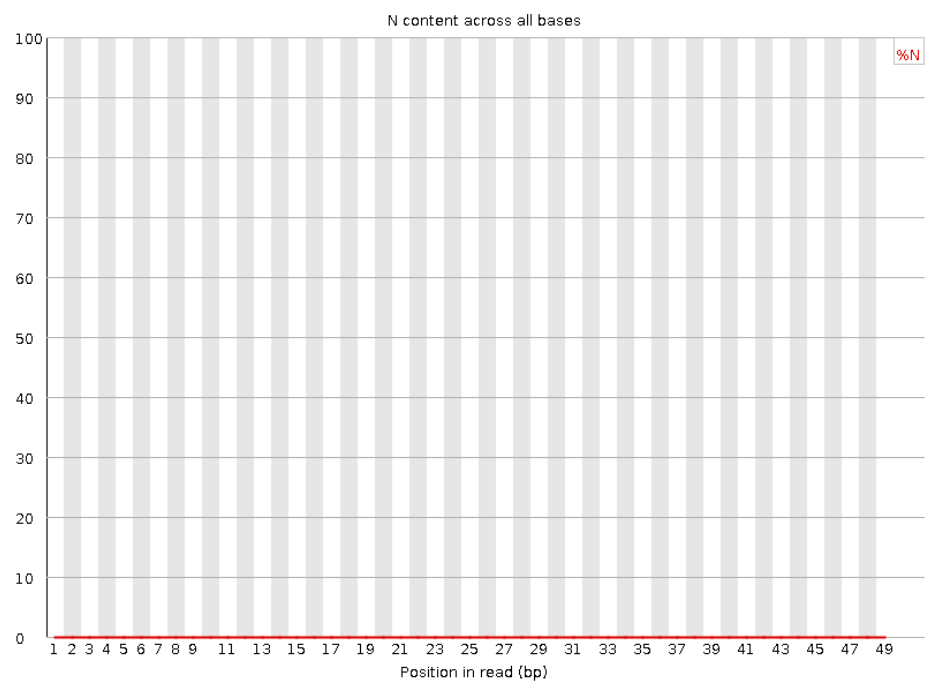


图 12: Per base N content (Processed)

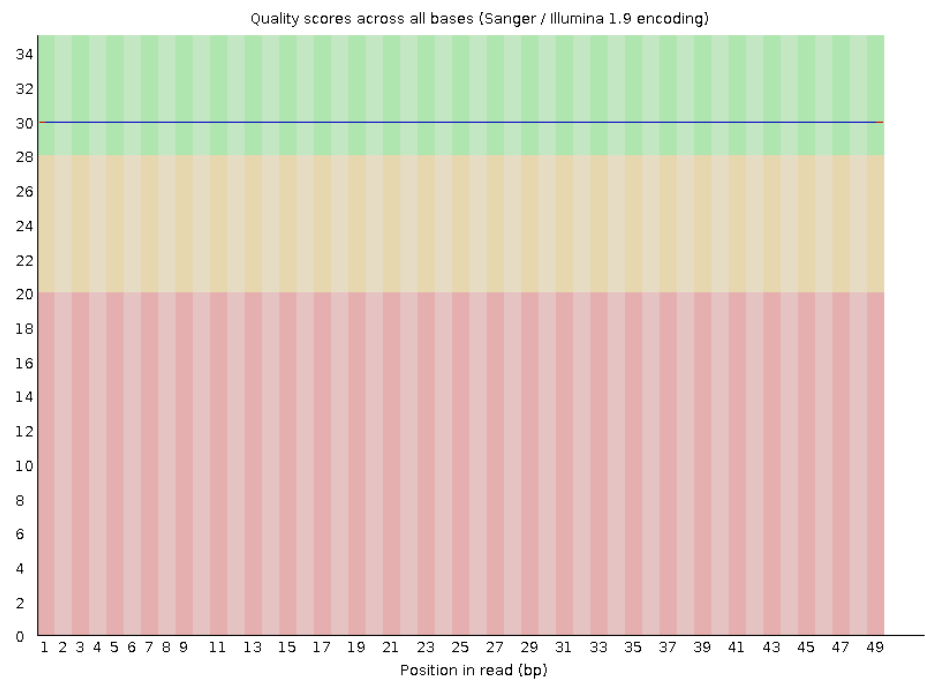


图 13: Per base quality (Processed)

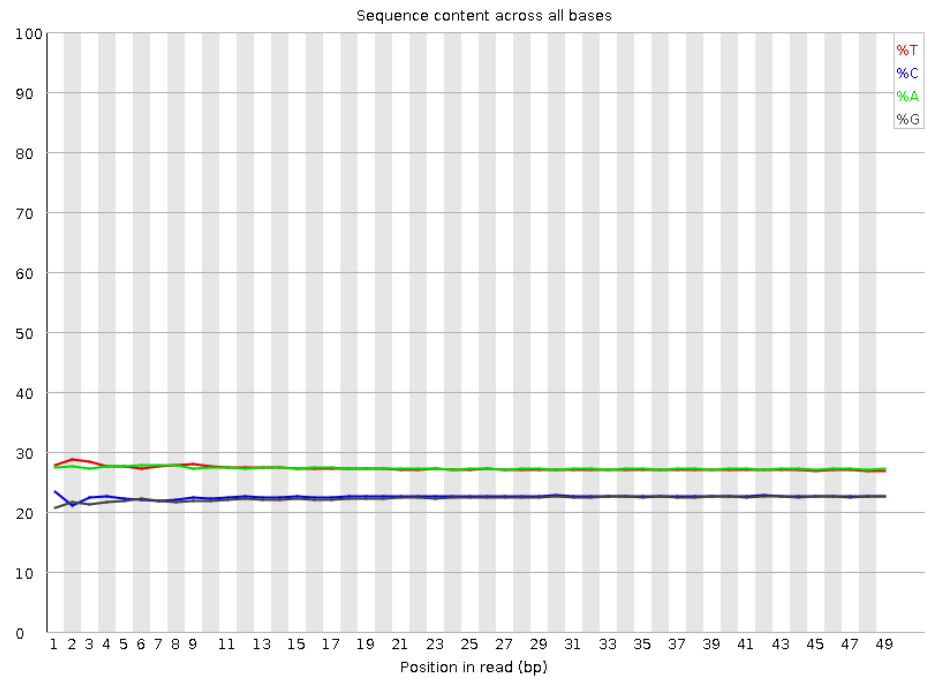


图 14: Per base sequence content (Processed)

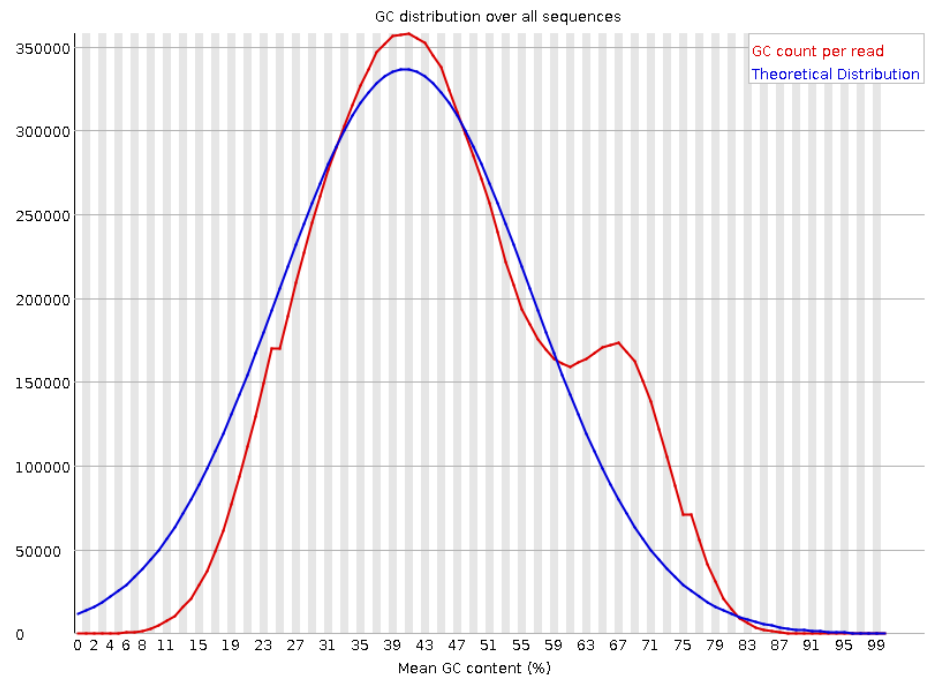


图 15: Per sequence GC content (Processed)

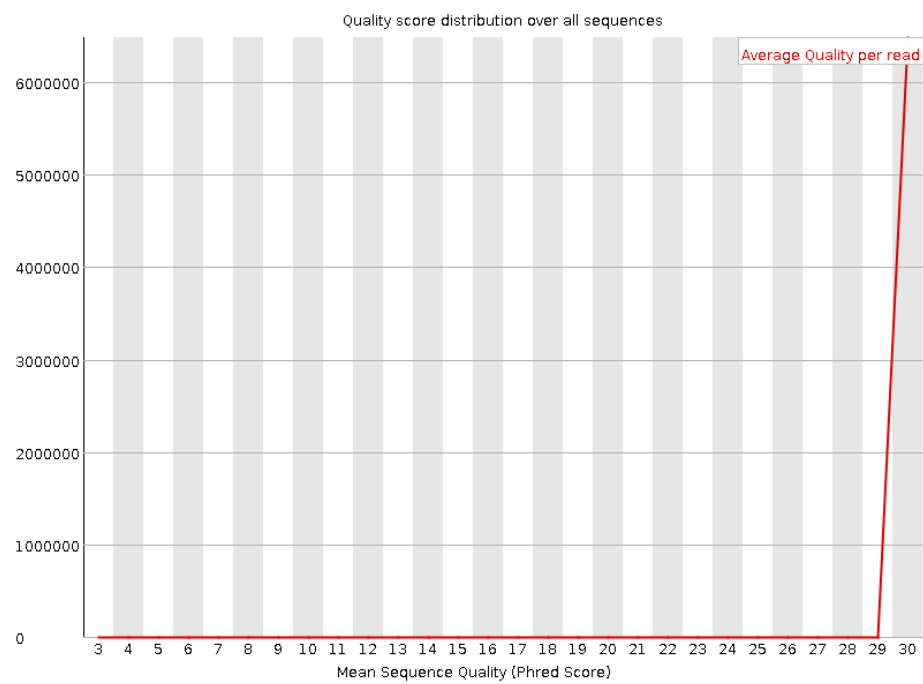


图 16: Per sequence quality (Processed)

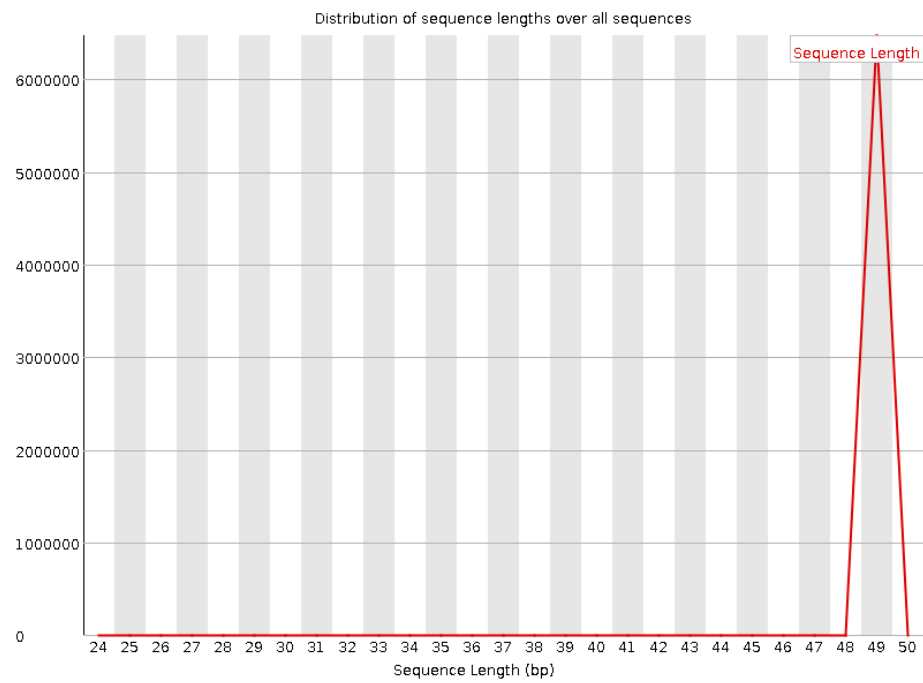


图 17: Sequence length distribution (Processed)

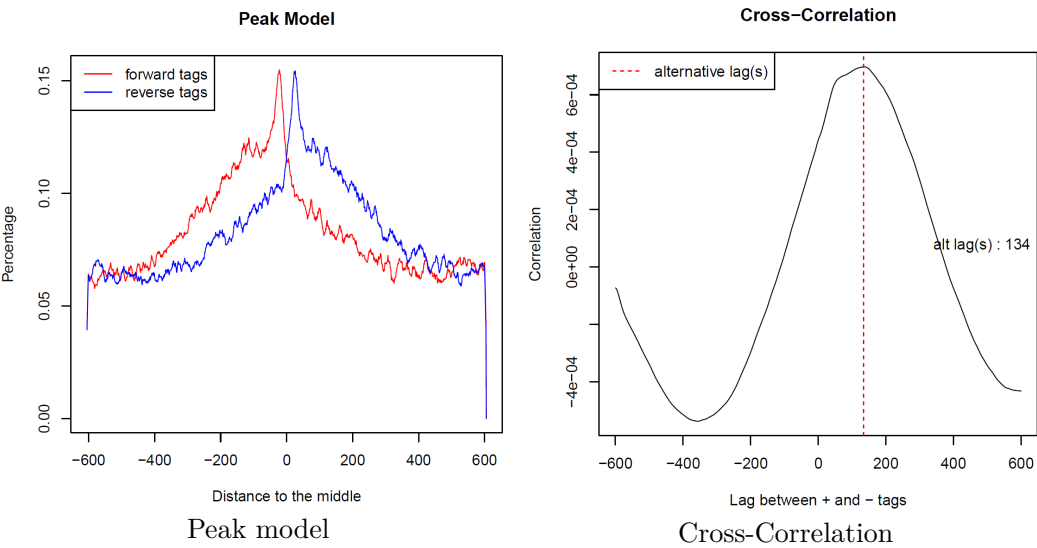


图 18: Peak calling 结果

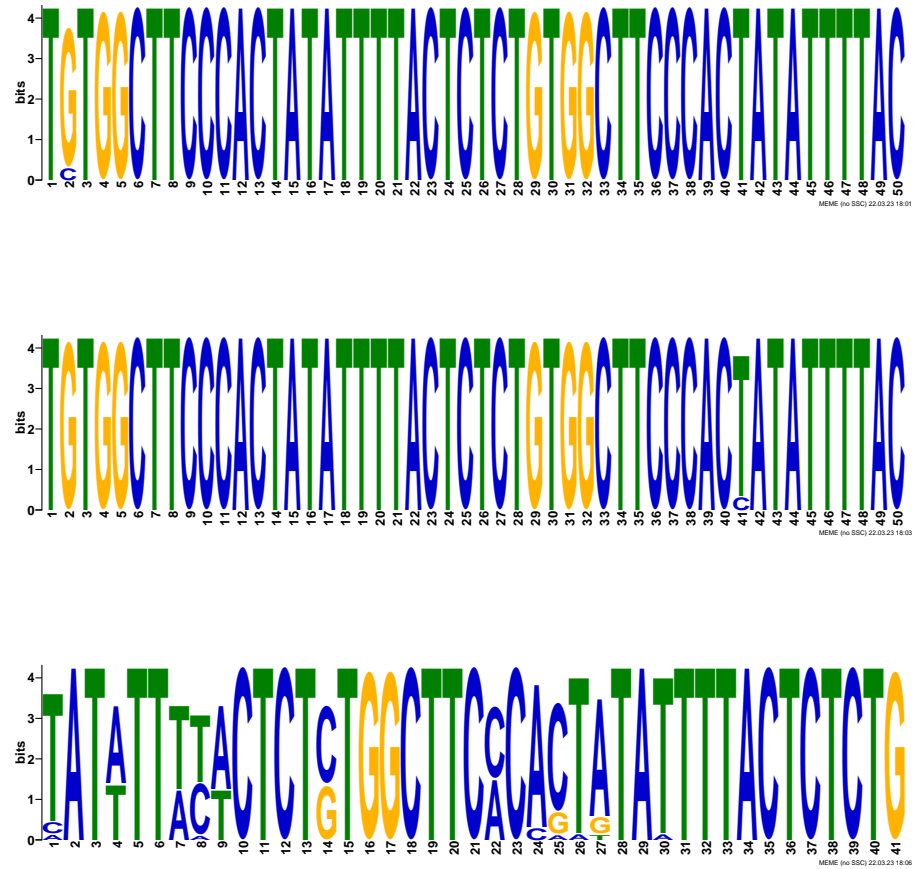


图 19: MEME 输出结果

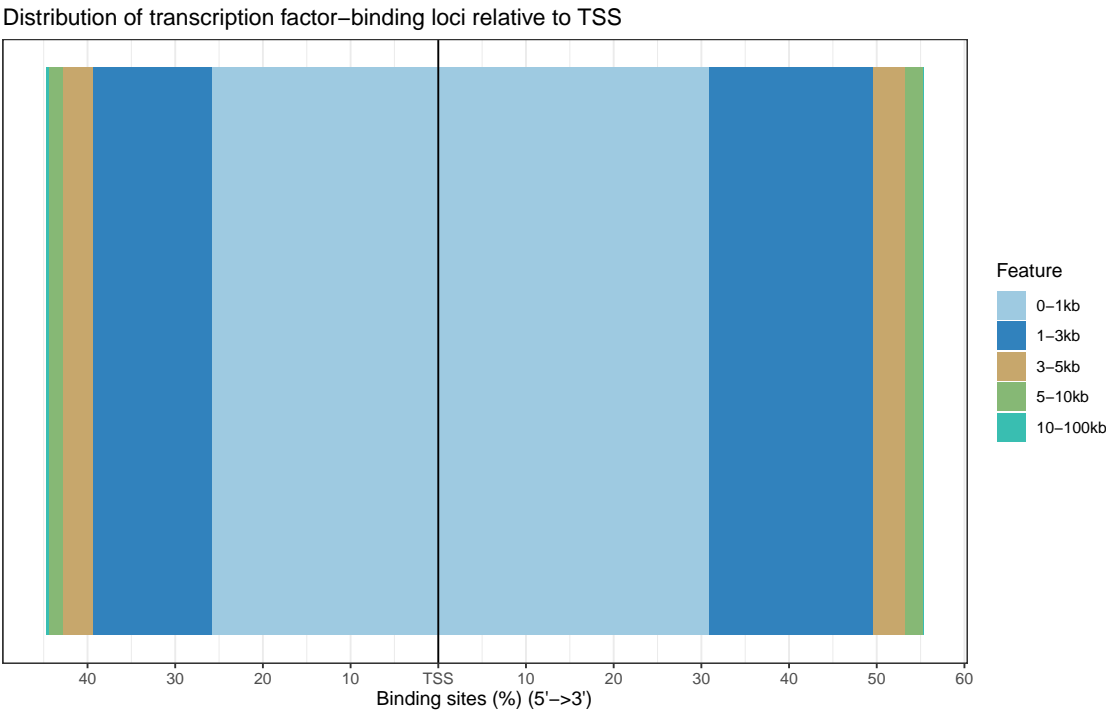


图 20: Distribution of transcription factor-binding loci relative to TSS

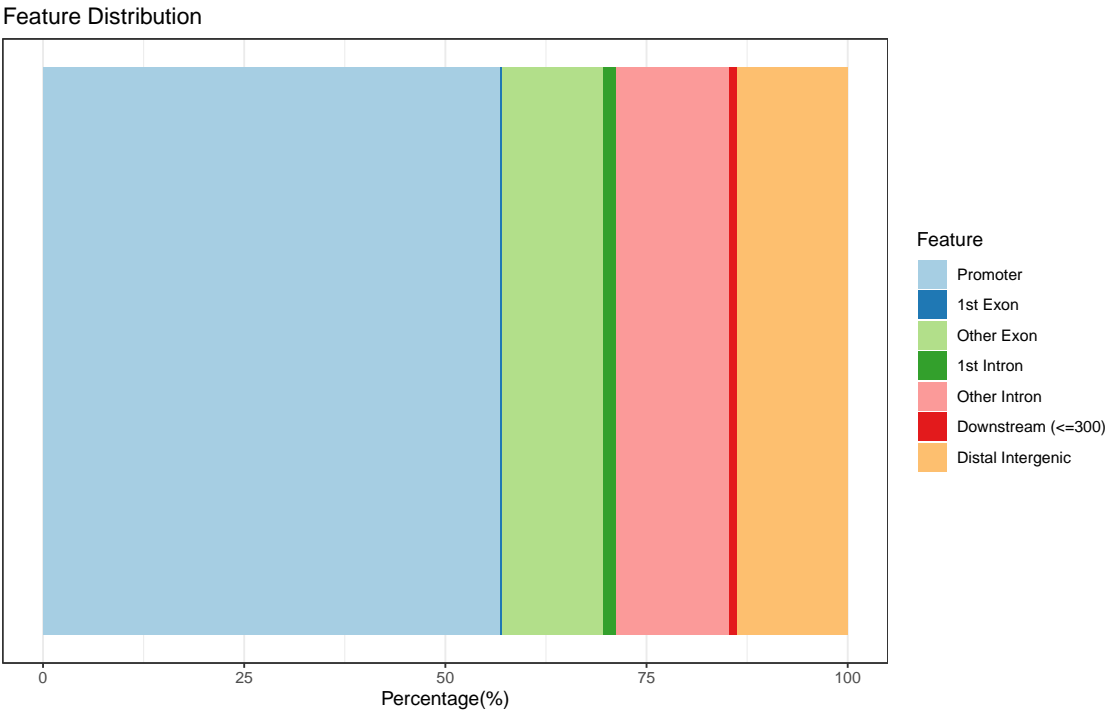


图 21: Feature Distribution in Percentage

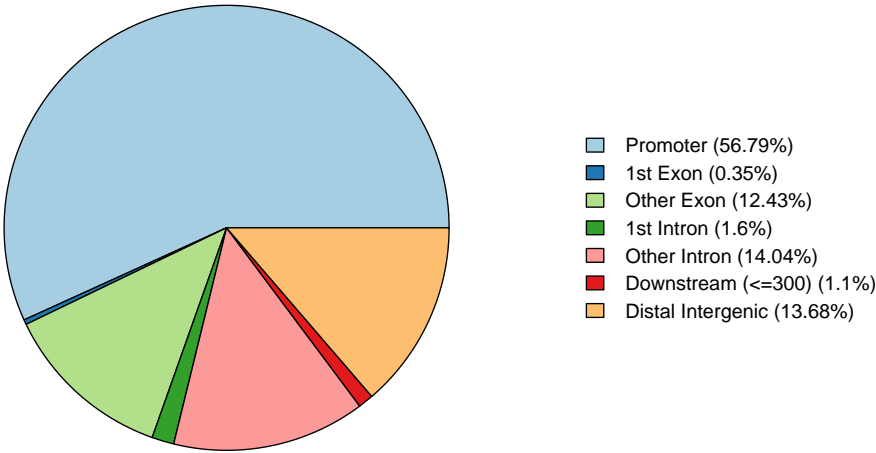


图 22: Feature Distribution in Pie Chart

