Download the genome file of Fusarium graminearum from ensembl genome

In [2]:
```
! wget ftp://ftp.ensemblgenomes.org/pub/release-51/fungi/fasta/fusarium_graminearum
```

```
--2024-04-14 18:07:43--  ftp://ftp.ensemblgenomes.org/pub/release-51/fungi/fasta/fus
arium_graminearum/dna/Fusarium_graminearum.RR1.dna.toplevel.fa.gz
           => ‘fusa_genome.fasta.gz’
Resolving ftp.ensemblgenomes.org (ftp.ensemblgenomes.org)... 193.62.193.161
Connecting to ftp.ensemblgenomes.org (ftp.ensemblgenomes.org)|193.62.193.161|:21...
connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/release-51/fungi/fasta/fusarium_graminearum/d
na ... done.
==> SIZE Fusarium_graminearum.RR1.dna.toplevel.fa.gz ... 11624507
==> PASV ... done.    ==> RETR Fusarium_graminearum.RR1.dna.toplevel.fa.gz ... done.
Length: 11624507 (11M) (unauthoritative)

Fusarium_graminearu 100%[===================>]  11.09M  6.34MB/s    in 1.7s

2024-04-14 18:07:47 (6.34 MB/s) - ‘fusa_genome.fasta.gz’ saved [11624507]
```

the size and format of the downloaded file

In [3]:
```
!du -h fusa_genome.fasta.gz
```

```
12M     fusa_genome.fasta.gz
```

Decompress this file while keeping the compressed version as wel

In [4]:
```
! gunzip -c fusa_genome.fasta.gz > fusa_genome.fasta
```

In [5]:
```
! du -h fusa_genome.fasta.gz fusa_genome.fasta
```

```
12M     fusa_genome.fasta.gz
37M     fusa_genome.fasta
```

Identify the sequences present in this file along with their identifiers

In [6]:
```
! grep '>' fusa_genome.fasta && echo "Number of sequences : $(grep -c '>' fusa_geno
```

```
>1 dna:chromosome chromosome:RR1:1:1:11760891:1 REF
>2 dna:chromosome chromosome:RR1:2:1:8997558:1 REF
>3 dna:chromosome chromosome:RR1:3:1:7792947:1 REF
>4 dna:chromosome chromosome:RR1:4:1:9395062:1 REF
>Mt dna:chromosome chromosome:RR1:Mt:1:95638:1 REF
>HG970330 dna:supercontig supercontig:RR1:HG970330:1:5846:1 REF
Number of sequences : 6
```

The size of the genome of Fusarium graminearum

In [7]:
```
!  grep -v '>' fusa_genome.fasta | tr -d '\n'| wc -c
```

```
38047942
```

The number of A/T/G/C bases in the genome

In [8]:
```
! grep -c '[ATCG]' fusa_genome.fasta
```

```
634136
```

In [9]:
```
%%bash
echo "Retrieve the number of A/T/G/C bases from the genome and deduce the overall G

# Calculate the total number of nucleotides
Total_Nucleotide=$(grep -v '>' fusa_genome.fasta | tr -d '\n' | wc -c)

# Calculate the number of G bases
Number_of_Ng=$(grep -v '>' fusa_genome.fasta | tr -d -c 'G' | wc -c)

# Calculate the number of C bases
Number_of_Nc=$(grep -v '>' fusa_genome.fasta | tr -d -c 'C' | wc -c)

# Calculate the sum of G and C counts
GC_sum=$((Number_of_Ng + Number_of_Nc))

# Calculate the overall GC content
GC_content=$(bc -l <<< "scale=2; ($GC_sum / $Total_Nucleotide) * 100")

echo "The overall GC content of Fusarium graminearum is: $GC_content%"
```

```
Retrieve the number of A/T/G/C bases from the genome and deduce the overall GC conte
nt of Fusarium graminearum.
The overall GC content of Fusarium graminearum is: 48.00%
```

We want to use the restriction enzyme BamH1, whose sequence is GGATCC. To guide the insertion of the reporter gene into this site.

The potential restriction sites are located in the genome.

In [10]:
```
%%bash
potential_restriction=$(grep --only-matching GGATCC fusa_genome.fasta | wc -l)
echo $potential_restriction
```

```
6706
```

The distance between two BamH1 sites (in Kb) "average"

In [11]:
```
%%bash
dd=$(expr $Total_Nucleotide / $potential_restriction )
echo "$dd"
```

```
/
```

Download the structural annotation of Fusarium graminearum

In [12]:
```
! wget ftp://ftp.ensemblgenomes.org/pub/release-51/fungi/gff3/fusarium_graminearum/
```

```
--2024-04-14 18:07:50--  ftp://ftp.ensemblgenomes.org/pub/release-51/fungi/gff3/fusa
rium_graminearum/Fusarium_graminearum.RR1.51.chr.gff3.gz
           => ‘fusa_annot.gff3.gz’
Resolving ftp.ensemblgenomes.org (ftp.ensemblgenomes.org)... 193.62.193.161
Connecting to ftp.ensemblgenomes.org (ftp.ensemblgenomes.org)|193.62.193.161|:21...
connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/release-51/fungi/gff3/fusarium_graminearum ..
. done.
==> SIZE Fusarium_graminearum.RR1.51.chr.gff3.gz ... 2010033
==> PASV ... done.    ==> RETR Fusarium_graminearum.RR1.51.chr.gff3.gz ... done.
Length: 2010033 (1.9M) (unauthoritative)

Fusarium_graminearu 100%[===================>]   1.92M  2.14MB/s    in 0.9s

2024-04-14 18:07:52 (2.14 MB/s) - ‘fusa_annot.gff3.gz’ saved [2010033]
```

The features described in this GFF3 file

In [13]:
```
! zcat fusa_annot.gff3.gz | grep -v '#'|cut -f 3|sort|uniq
```

```
CDS
RNase_MRP_RNA
RNase_P_RNA
SRP_RNA
biological_region
chromosome
exon
five_prime_UTR
gene
lnc_RNA
mRNA
ncRNA_gene
rRNA
snRNA
snoRNA
tRNA
three_prime_UTR
```

Types of different features are you counting

In [14]:
```
! zcat fusa_annot.gff3.gz | grep -v '#'|cut -f 3|sort|uniq|wc -l
```

```
17
```

The genes annotated on this genome

In [15]:
```
! zcat fusa_annot.gff3.gz|grep -w gene|grep 'ID=gene'|wc -l
```

```
14898
```

Genes and mRNA annotated on each of the chromosomes

In [16]:
```
! zcat fusa_annot.gff3.gz|grep -w gene|grep 'biotype=protein_coding'|cut -f1,3|sort
```

```
4390 1	gene
4390 1	mRNA
3648 2	gene
3648 2	mRNA
3085 3	gene
3085 3	mRNA
3022 4	gene
3022 4	mRNA
```

The BED format is a simpler alternative to the GFF format for describing genomic objects, consisting of 4 columns separated by tabs:

* Column 1: chromosome
* Column 2: start of the genomic element described
* Column 3: stop of the genomic element
* Column 4: identifier or name of the genomic element

We want to extract from the previous GFF file the lines corresponding to genes and convert this information to BED format. This file should be sorted:

* by chromosome
* by coordinate on the chromosome.

For this, I will use combinations of grep/cut/sed/sort commands and generate the file fusa_genes.bed.

In [26]:
```
! zcat fusa_annot.gff3.gz | fgrep -v '#' | cut -f 1,4,5,9 | fgrep 'ID=gene' | cut -
```

In [27]:
```
! head -n5 fusa_genes.bed
```

```
1	6089	11000	FGRAMPH1_01G00001
1	11394	12168	FGRAMPH1_01G00003
1	19069	19668	FGRAMPH1_01G00005
1	21463	22193	FGRAMPH1_01G00007
1	23519	25108	FGRAMPH1_01G00009
```

Add a prefix 'FusaChrom' to the chromosome names using sed directly in the file fusa_genes.bed.

In [28]:
```
! sed -i 's/^/FusaChrom/' fusa_genes.bed
! head -n5 fusa_genes.bed
```

```
FusaChrom1	6089	11000	FGRAMPH1_01G00001
FusaChrom1	11394	12168	FGRAMPH1_01G00003
FusaChrom1	19069	19668	FGRAMPH1_01G00005
FusaChrom1	21463	22193	FGRAMPH1_01G00007
FusaChrom1	23519	25108	FGRAMPH1_01G00009
```

Retrieve the gene sizes in a file gene_size.tab (column1: gene name, column2: gene size), using the created BED file, and calculate the gene sizes with awk.

In [30]:
```
! awk 'BEGIN{OFS="\t"} {genesize=$3-$2+1;print $4,genesize}' fusa_genes.bed > gene_
```

```
FGRAMPH1_01G00001       4912
FGRAMPH1_01G00003       775
FGRAMPH1_01G00005       600
FGRAMPH1_01G00007       731
FGRAMPH1_01G00009       1590
```

Calculate the average size of the genes

In [31]:
```
! awk 'BEGIN{sum=0}{sum+=$2}END{mean=sum/NR;print "The average size of the genes :
```

```
The average size of the genes : 1686.88
```

Genes larger than 5kb are considered particularly large. I will count how many are equal to or larger than 5kb, and how many are smaller than 5kb

In [33]:
```
%%bash
awk 'BEGIN{smaller5kb=0;larger5kb=0} {if ($2 < 5000) smaller5kb+=1; else \
larger5kb+=1}END {printf "Found %d genes smaller than 5kb and %d genes larger than
5kb\n",smaller5kb,larger5kb}' gene_size.tab
```

```
Found 14542 genes smaller than 5kb and 356 genes larger than 5kb
```

I wish to extract the coordinates of promoters for these genes (2000 bp upstream of the genes). I'll create a file named "fusa_prom.bed" containing the coordinates of the promoters of these genes, and we'll name these promoters using the gene name followed by the suffix_prom.

In [34]:
```
%%bash
awk 'BEGIN{OFS="\t"}{print $1,$2-2000,$2-1,$4"_prom"}' fusa_genes.bed > fusa_prom.b
awk 'BEGIN{OFS="\t"}{print $1,$2-2000,$2-1,$4}' fusa_genes.bed | sed 's/$/_prom/' >
head -n 5 fusa_prom.bed
```

```
FusaChrom1      4089    6088    FGRAMPH1_01G00001_prom
FusaChrom1      9394    11393   FGRAMPH1_01G00003_prom
FusaChrom1      17069   19068   FGRAMPH1_01G00005_prom
FusaChrom1      19463   21462   FGRAMPH1_01G00007_prom
FusaChrom1      21519   23518   FGRAMPH1_01G00009_prom
```