Occultercut

Occultercut software download and original research paper

The Occultercut software can be downloaded from

wget https://sourceforge.net/projects/occultercut/

The original publication can be accessed from

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4943192/

Syntax format for Occultercut

```
OcculterCut -f <fasta> -a <gff file> -i 30,50,70
```

where the fasta file corresponds to the gapfilled sacffold sequences, and the gff file corresponds to the maker gene annotation file. A small sample of how the gene annotation file looks like is depicted below:

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.4.4
```

```
Meumgff <- read_csv("C:/Users/dr382/Desktop/Meumgff.csv",col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   `scaffold-number` = col_character(),
##   Source = col_character(),
##   Feature = col_character(),
##   Start = col_integer(),
##   End = col_integer(),
##   Score = col_character(),
##   Strand = col_character(),
##   Frame = col_character(),
##   Attribute = col_character()
## )
```

```
##a subset of the data as example
head(Meumgff,n=5)
```

```
## # A tibble: 5 x 9
##       `scaffold-number` Source Feature Start   End Score Strand Frame
##                   <chr>  <chr>   <chr> <int> <int> <chr>  <chr> <chr>
## 1 scaffold353|size25484  maker     CDS 15559 17922     .      +     .
## 2 scaffold353|size25484  maker     CDS 10044 12026     .      +     .
## 3 scaffold353|size25484  maker     CDS  5817  6863     .      +     .
## 4 scaffold353|size25484  maker     CDS  8607  9386     .      +     .
## 5 scaffold353|size25484  maker     CDS  4584  5341     .      +     .
## # ... with 1 more variables: Attribute <chr>
```

More details on the gff file format can be found at https://uswest.ensembl.org/info/website/upload/gff.html

Sample data

The three genomes used here are from the sigatoka complex *M.eumusae*, *M.musae*, and *M.fijiensis*. For further details on the genomes and download of fasta files for the analysis, the link to the original publication is http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005904.

Identifying AT-rich regions using Occultercut

M.eumusae

```
./OcculterCut -f Meum_gapfiller_out.gapfilled.final.fa -a eumusae
# uncomment the following two lines if you would like an eps file output of your plot
set terminal png
set output 'M.eumusae.png'
set xlabel "GC (%)"
set ylabel "Proportion of genome"
set sample 1000
set xrange[0:100]
set yrange[0:]
set boxwidth 1
set style fill solid
set key off
set style line 1 lt 1 lc rgb "#0000FF" lw 3
set style line 2 lt 2 lc rgb "#32CD32" lw 3
Cauchy(x,xo,wi) = (1./pi) * wi / ((x - xo)**2 + wi**2)
set arrow 1 from 47.6, graph 0 to 47.6, graph 1 front nohead lc rgb "#0000CD"
plot 'compositionGC.txt' w boxes, 0.625358*Cauchy(x, 40.7133,1.60869) + 0.374642
*Cauchy(x, 52.3131, 0.807679) ls 2
set yrange[0:GPVAL_Y_MAX]
replot
```

M.musae

```
./OcculterCut -f Mmus_gapfiller_out.gapfilled.final.fa -a musae
# uncomment the following two lines if you would like an eps file output of your plot
set terminal png
set output 'M.musicola.png'
set xlabel "GC (%)"
set ylabel "Proportion of genome"
set sample 1000
set xrange[0:100]
set yrange[0:]
set boxwidth 1
set style fill solid
set key off
set style line 1 lt 1 lc rgb "#0000FF" lw 3
set style line 2 lt 2 lc rgb "#32CD32" lw 3
Cauchy(x,xo,wi) = (1./pi) * wi / ((x - xo)**2 + wi**2)
set arrow 1 from 47.6, graph 0 to 47.6, graph 1 front nohead lc rgb "#0000CD"
plot 'compositionGC.txt' w boxes, 0.625358*Cauchy(x, 40.7133,1.60869) + 0.374642
```

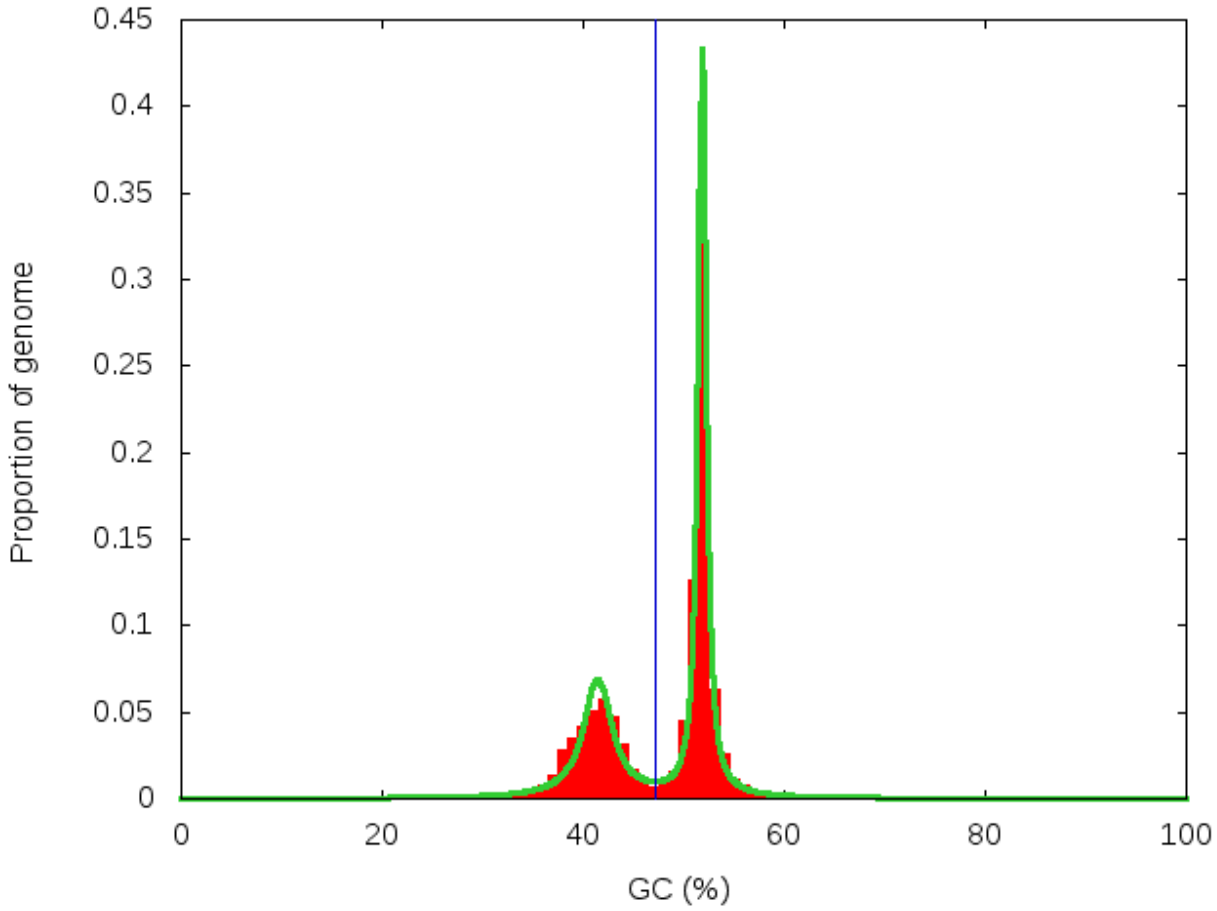

Figure 1: GC content plot of the *M.eumusae* genome.

```
*Cauchy(x, 52.3131, 0.807679) ls 2
set yrange[0:GPVAL_Y_MAX]
replot
```

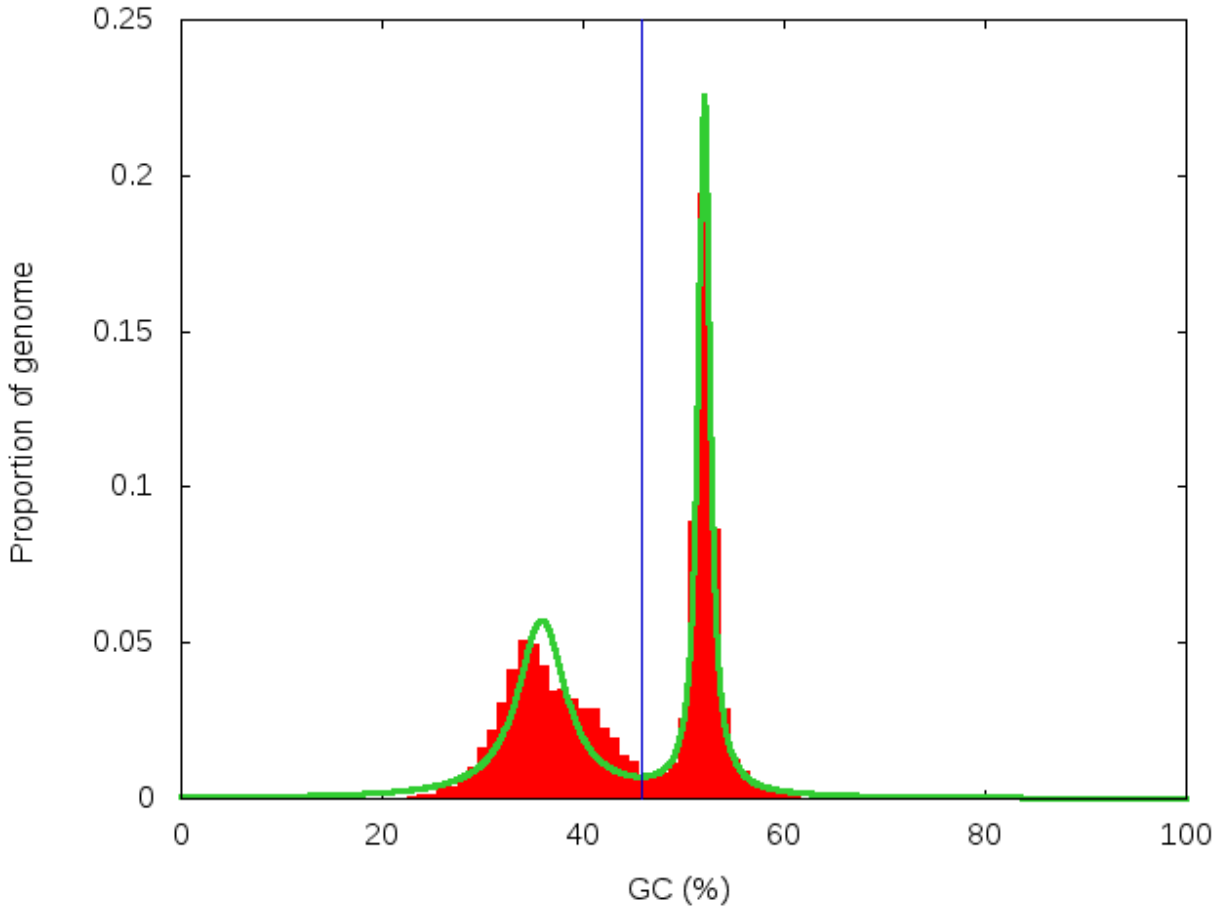

Figure 2: GC content plot of the *M.musae* genome.

M.fijiensis

```
./OcculterCut -f Mycosphaerella_fijiensis_v2.fasta -a Mfij.gff
# uncomment the following two lines if you would like an eps file output of your plot
set terminal png
set output 'M.fijiensis.png'
set xlabel "GC (%)"
set ylabel "Proportion of genome"
set sample 1000
set xrange[0:100]
set yrange[0:]
set boxwidth 1
set style fill solid
set key off
set style line 1 lt 1 lc rgb "#0000FF" lw 3
```

```
set style line 2 lt 2 lc rgb "#32CD32" lw 3
Cauchy(x,xo,wi) = (1./pi) * wi / ((x - xo)**2 + wi**2)
set arrow 1 from 47.6, graph 0 to 47.6, graph 1 front nohead lc rgb "#0000CD"
plot 'compositionGC.txt' w boxes, 0.625358*Cauchy(x, 40.7133,1.60869) + 0.374642
*Cauchy(x, 52.3131, 0.807679) ls 2
set yrange[0:GPVAL_Y_MAX]
replot
```

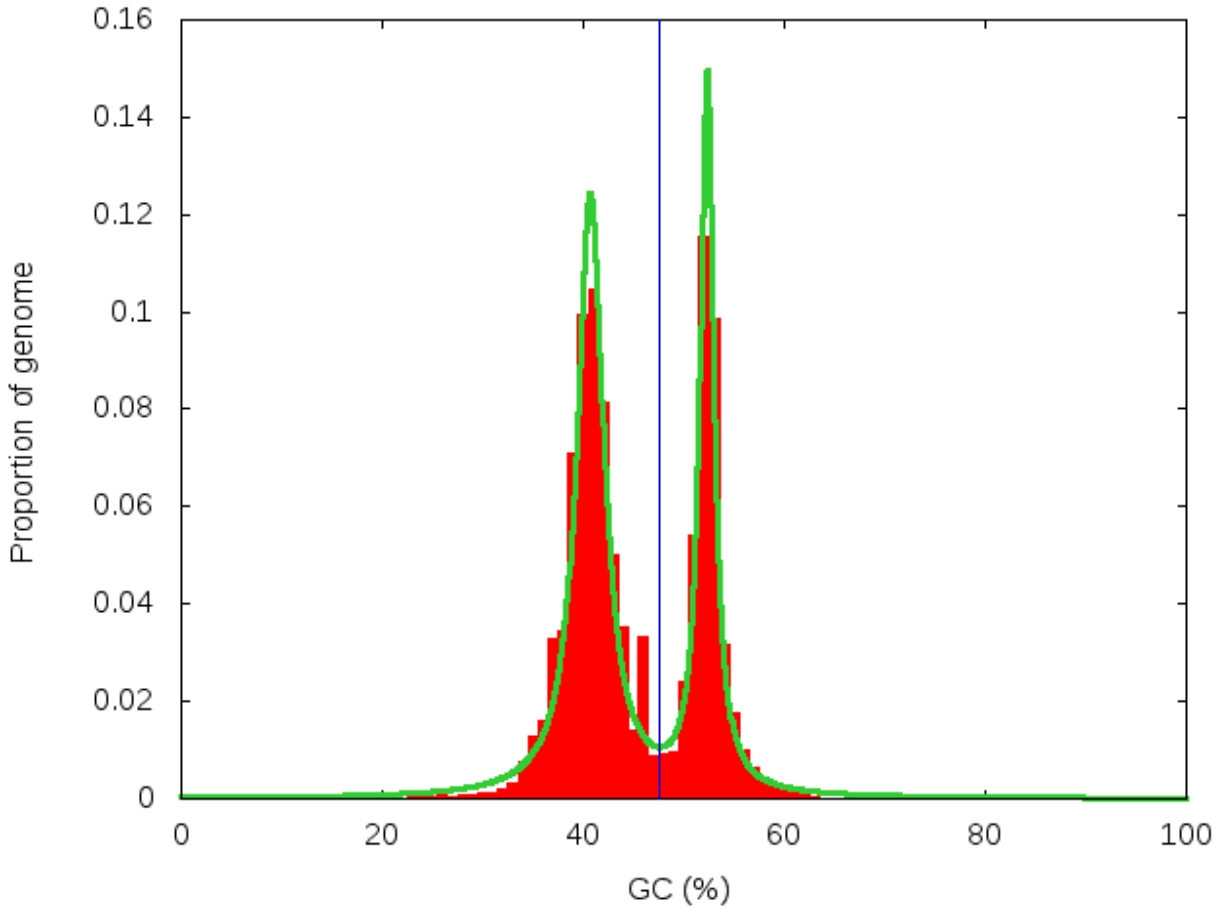

Figure 3: GC content plot of the *M.fijiensis* genome.

Figures 1, 2, and 3 represent the GC content plots of the genomes from Sigatoka complex. The plots display their diversity in the GC-contents of peaks, shape, and height of the peaks. Classification of genome segments into AT-rich or GC-equilibrated is represented by vertical blue lines and done only when the peak separation is >=10% GC and a minimum Cauchy distribution can be identified between the two peaks (green line shows a mixture of two Cauchy distributions). Red bars indicate the proportion of the genome classified as segments of different GC-content. Percentages on the left indicate percentage of the genome classified as AT-rich and to the right indicates GC-equilibrated.

Role of Repeat Induced Mutations in AT-content

Repeat induced point mutations (RIP), a fungal specific defense mechanism first reported in Neurospora crassa involves transitions from C:G to T:A nucleotides that occurs at the pre-meiotic stage of sexual reproduction (Galagan et al., 2003). An observable impact of these nucleotide transitions is the depletion of GC content (AT-rich regions) causing wide variation in the GC content of fungal genomes. This in turn targets repetitive

DNA, and the presence of RIP counterbalances the ability of transposable elements (TEs) to invade genomes resulting in smaller sized genomes. RIP mutations are also known to impact single-copy regions and fungal pathogenicity related genes. There have been previous studies focusing on the impact of RIP on repeat regions which indicate that large blocks of mutated transposable elements tend to increase the AT-isochore. However, studying the impact of RIP using annotated repeat elements does not shed light on RIP-degraded repeats. Here, we focus on identifying AT-rich regions within whole genomes of the banana and solanaceous pathogens allowing additional inferences to be made. Genome segmentation of the genomes into AT-rich regions and GC-equilibrated genomes was done using OcculterCut and the surveyed genomes are bimodal in nature implying that they contain AT-rich regions. The genomes of P.musicola, and P.fijiensis have AT-rich regions in higher proportions than GC-equilibrated regions (Figures 2 and 3).

```
##Processing Occultercut output files to get the necessary information using awk and sed

##Get all genes in from groupedgenes file
awk '$3 == "gene" { print $0 }' groupedGenes.gff3 >all_genes

##Get genes in AT rich regions
grep "R0" all_genes >AT-richgenes

##Get only complete genes
grep -iw "complete" distances.txt >completegenes

###remove ID= and everything after .R0 to get the necessary ids
cut -f 9  AT-richgenes >AT-richgene_ids
sed -i 's/.R0;[^ ]* *//g' AT-richgene_ids
sed -i 's|[ID=]||g' AT-richgene_ids

##Match gene ids to complete genes and get final list
grep -f "AT-richgene_ids" "completegenes" >ATrich_complete_Meum


##Check if the genes that are At-rich are effectors
##This could be done using the compliantFasta file prepared for Orthomcl analysis
###then have the ids match to the effector file to pull out matching ids
grep -f "ATrich_complete_Meum" "/compliantfasta/Meum.fasta" >Meum_effectors_ATrich
```