Contents

Diário de bordo da geração dos shinyCircos

Considerar apenas os C147-Phased

Camadas:

1. Cromossomos
2. Coding
3. Non Coding
4. TE

Configuração inicial

```
1 mkdir -pv {tmp,output}
```

Pre processamento

O arquivo `ResultadoFinal_C174-PHASED.gff` possui strand `+`, `-`, `plus` e `minus`. O correto é apenas `+` e `-`.

```
awk 'BEGIN {FS="\t";OFS="\t"} {if ($7=="plus"){$7="+"}else
    if($7=="minus"){$7="-"} {print}}'
    ResultadoFinal_C174-PHASED.gff >
    tmp/ResultadoFinal_C174-PHASED-corrigido.gff
```

Obter os cromossomos

```
# Print sequence length and names (no sequences)
mkdir -pv output
echo "chr;start;end" > output/chr.csv
seqkit fx2tab C174-PHASED.fa -l -n | awk 'BEGIN {OFS = ";"}
    {print $1,0,$2}' >> output/chr.csv
```

Obtendo a densidade dos cromossomos

Exemplos de uso

```
DensityMap.pl -i ./annot-genes/Tgrand_C174P-Annotation.gff3 -ty
    "ncRNA=fused=10" -o 2R.svg -for -v

DensityMap.pl -i ../DensityMap/dmel/dmel.gff3 -o egn.svg -ty
    'gene=fused;exon=fused;ncRNA=fused=10' -gc 12 -sc 40000 -ba
    white -str_s 15 -str_w 25 -sp 35 -sh 50 -title "Density Map
    of Gene, Exon, ncRNA and GC%" -la -15 -ro ceil -v
```

Problemas encontrados até o momento

1. A marcação `##sequence-region` do GFF precisa obrigatoriamente iniciar em 1. Caso contrário o DensityMap não identifica a linha. Foi apresentado uma solução no Pull Request (https://github.com/sguizard/DensityMap/pull/5). Aguardando o autor aceitar.
2. Se o GFF possuir mais de uma região (`##sequence-region`), a marcação da região deve vir imediatamente antes dos seus registros.

Corrigindo os arquivos para o uso do DensityMap

```
# ./annot-genes/Tgrand_C174P-Annotation.gff3

mkdir -pv tmpx
```

```
awk '/##sequence-region/ {print "##gff-version 3" > "tmpx/"$2".gff3"}' ./annot-genes/Tgrand_C174P-Annotation.gff3
awk '/##sequence-region/ {print $1,$2,1,$4 >> "tmpx/"$2".gff3"}' ./annot-genes/Tgrand_C174P-Annotation.gff3
awk '/^chr[0-9]{1,2}[ab]/ {print $0 >> "tmpx/"$1".gff3"}' ./annot-genes/Tgrand_C174P-Annotation.gff3
cat tmpx/*.gff3 > tmp/Tgrand_C174P-Annotation-for-DensityMap.gff3
rm -rv tmpx
```

```
# ResultadoFinal_C174-PHASED.gff
mkdir -pv tmpx
awk 'BEGIN {FS=";";OFS=" "} /^chr[0-9]+[ab]/ {print "##gff-version 3" > "tmpx/"$1".gff3"}' output/chr.csv
awk 'BEGIN {FS=";";OFS=" "} /^chr[0-9]+[ab]/ {print "##sequence-region", $1, "1", $3 >> "tmpx/"$1".gff3"}' output/chr.csv
awk 'BEGIN {FS="\t";OFS="\t"} /^chr[0-9]{1,2}[ab]/ {if ($3 ~ /RNA$/) {$3="ncRNA"; print $0 >> "tmpx/"$1".gff3"}}' tmp/ResultadoFinal_C174-PHASED-corrigido.gff
cat tmpx/*.gff3 > tmp/ResultadoFinal_C174-PHASED-for-DensityMap.gff3
rm -rv tmpx
```

```
# TEs/Tgrand_C174P-EDTA.TEanno.gff3
mkdir -pv tmpx
awk 'BEGIN {FS=";";OFS=" "} /^chr[0-9]+[ab]/ {print "##gff-version 3" > "tmpx/"$1".gff3"}' output/chr.csv
awk 'BEGIN {FS=";";OFS=" "} /^chr[0-9]+[ab]/ {print "##sequence-region", $1, "1", $3 >> "tmpx/"$1".gff3"}' output/chr.csv
awk 'BEGIN {FS="\t";OFS="\t"} /^chr[0-9]{1,2}[ab]/ {$3="TE"; print $0 >> "tmpx/"$1".gff3"}' TEs/Tgrand_C174P-EDTA.TEanno.gff3
cat tmpx/*.gff3 > tmp/Tgrand_C174P-EDTA.TEanno-for-DensityMap.gff3
rm -rv tmpx
```

Obter os CODING (Camada 1)

```
echo "chr;start;end;stack" > output/chart1_mRNA.csv
awk 'BEGIN {FS = "\t";OFS = ";"} /^chr[0-9]{1,2}[ab]/ {if ($3=="mRNA") {print $1,$4,$5,"mRNA"}}' ./annot-genes/Tgrand_C174P-Annotation.gff3 >> output/chart1_mRNA.csv

wc -l output/*_mRNA.csv
```

Obter os CODING com densidade (Camada 1)

```
DensityMap.pl -i tmp/Tgrand_C174P-Annotation-for-DensityMap.gff3
    -ty "mRNA=fused" -o tmp/chart1_mRNA.svg -sc 500000 -ro ceil
    -for -v

echo "chr;start;end;value" > output/chart1_mRNA-DensityMap.csv
awk 'BEGIN {FS = "\t";OFS = ";"} /^chr[0-9]{1,2}[ab]/ {print $1,
    $3, $4-1, $5}' tmp/chart1_mRNA.csv >>
    output/chart1_mRNA-DensityMap.csv
```

Obter os NON CODING (Camada 2)

```
echo "chr;start;end;stack" > output/chart1_ncRNA.csv
awk 'BEGIN {FS = "\t";OFS = ";"} /^chr[0-9]{1,2}[ab]/ {if ($3 ~
    /RNA$/) {print $1,$4,$5,"ncRNA"}}'
    tmp/ResultadoFinal_C174-PHASED-corrigido.gff >>
    output/chart1_ncRNA.csv

grep -P "RNA\t" tmp/ResultadoFinal_C174-PHASED-corrigido.gff |
    wc -l
wc -l output/*_ncRNA.csv
```

Obter os NON CODING com densidade (Camada 2)

```
DensityMap.pl -i
    tmp/ResultadoFinal_C174-PHASED-for-DensityMap.gff3 -ty
    "ncRNA=fused" -o tmp/chart1_ncRNA.svg -sc 1000000 -ro ceil
    -for -v

echo "chr;start;end;value" > output/chart1_ncRNA-DensityMap.csv
awk 'BEGIN {FS = "\t";OFS = ";"} /^chr[0-9]{1,2}[ab]/ {print $1,
    $3, $4-1, $5}' tmp/chart1_ncRNA.csv >>
    output/chart1_ncRNA-DensityMap.csv
```

Obter os TEs (Camada 3)

```
echo "chr;start;end;stack" > output/chart1_TE.csv
awk 'BEGIN {FS = "\t";OFS = ";"} /^chr[0-9]{1,2}[ab]/ {print
    $1,$4,$5,"TE"}' ./TEs/Tgrand_C174P-EDTA.TEanno.gff3 >>
    output/chart1_TE.csv

wc -l ./TEs/Tgrand_C174P-EDTA.TEanno.gff3
wc -l output/*_TE.csv
```

Obter os TEs com densidade (Camada 3)

```
DensityMap.pl -i
    tmp/Tgrand_C174P-EDTA.TEanno-for-DensityMap.gff3 -ty
    "TE=fused" -o tmp/chart1_TE.svg -sc 500000 -ro ceil -for -v

echo "chr;start;end;value" > output/chart1_TE-DensityMap.csv
awk 'BEGIN {FS = "\t";OFS = ";"} /^chr[0-9]{1,2}[ab]/ {print $1,
    $3, $4-1, $5}' tmp/chart1_TE.csv >>
    output/chart1_TE-DensityMap.csv
```

Anotações

```
awk '/^chr[0-9]{1,2}[ab]/ {print $3}' <
    ResultadoFinal_C174-PHASED.gff | sort | head
awk '/^chr[0-9]{1,2}[ab]/ {if ($3 ~ /RNA$/) {print $3}}'
    ResultadoFinal_C174-PHASED.gff | sort | uniq

# Obter tipos de TEs
awk '/^chr[0-9]{1,2}[ab]/ {print $3}'
    ./TEs/Tgrand_C174P-EDTA.TEanno.gff3 | sort | uniq
```