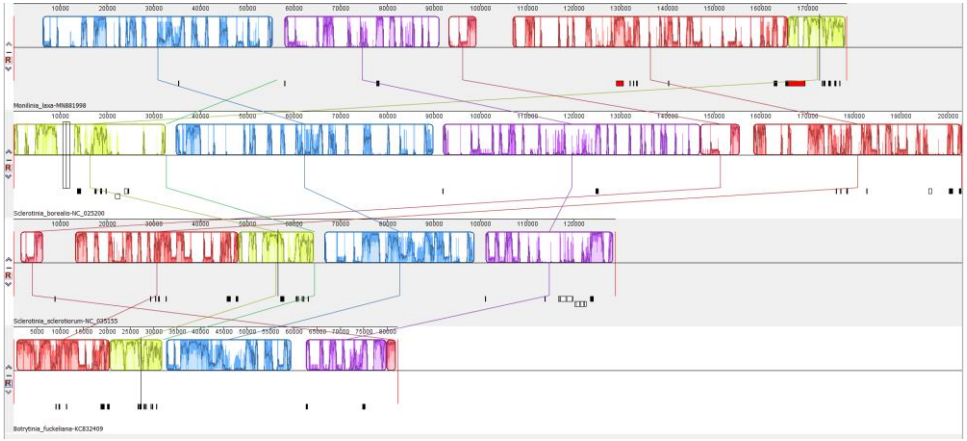
Genomik Çalışmalarda R Kullanımı



Doç. Dr. Hilal ÖZKILINÇ

Çanakkale Onsekiz Mart Üniversitesi
Fen-Edebiyat Fakültesi
Moleküler Biyoloji ve Genetik Bölümü

R Programming

RStudio®

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Yeni Nesil Dizileme-Veri-Veri Isleme

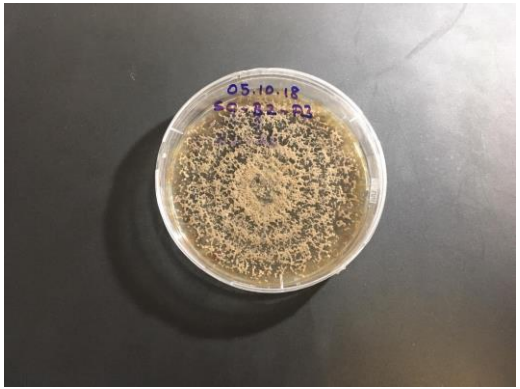
Yapisal Analizler (Gen bolgeleri, gen siralari, genom haritalari …)

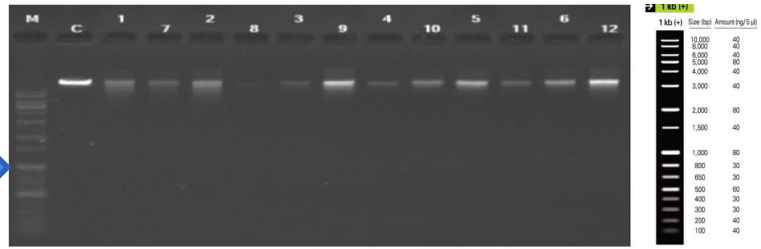
Fonskiyonel Analizler (Regulasyon, transkriptom veri isleme …)

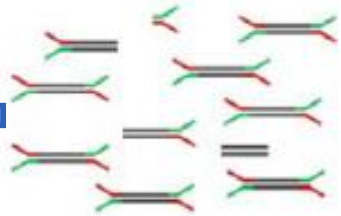
Evrimsel analizler

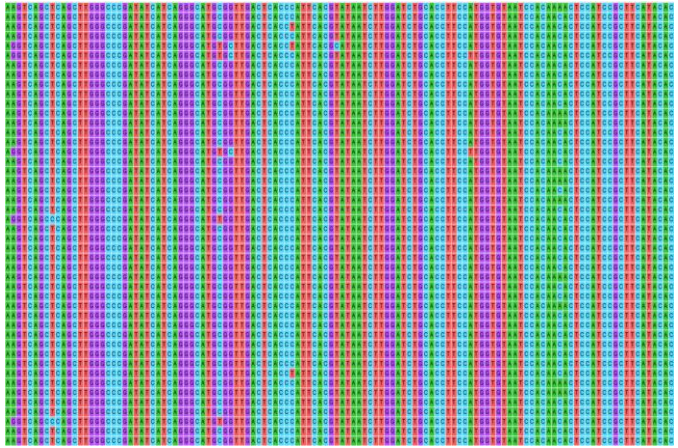
…

Yeni Nesil Dizileme-Veri-Veri Isleme









Yeni Nesil Dizileme-Veri-Veri Isleme

- FastA
- FastQ
- SAM
- BAM
- VCF
- GFF3
- …

Yeni Nesil Dizileme-Veri-Veri Isleme

- Kalite Kontrol
- Assembly (Genom toplama) (Referansa gore veya de novo)
- Anotasyon
- Varyant Tespiti
- Genom Yapilari
- Genom dizi kiyaslamalari
- Genom iliski analizleri
- Genom istatistikleri
- …

BiocManager: Access the Bioconductor Project Package Repository

“Use the BiocManager package to install and manage packages from the Bioconductor project for the statistical analysis and comprehension of high-throughput genomic data.”

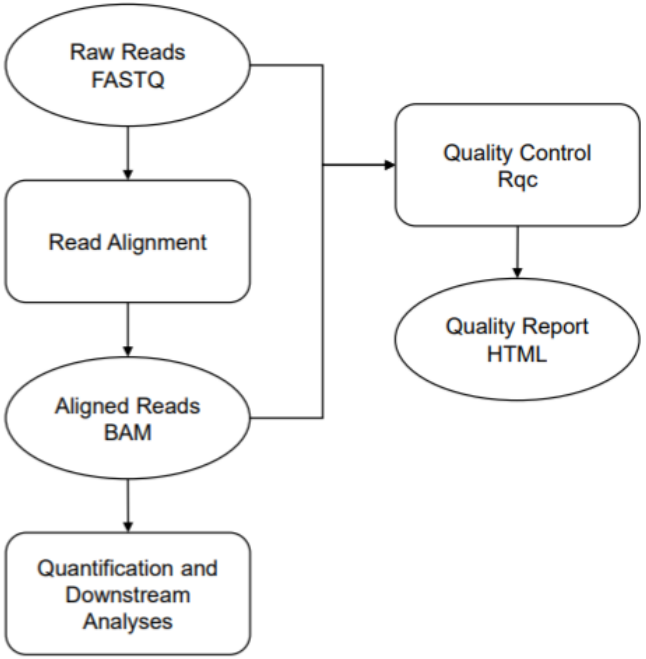
```
BiocManager::install()

#Install or update packages from Bioconductor, CRAN, and GitHub.

BiocManager::repositories()
```

Rqc - Quality Control Tool for High-Throughput Sequencing Data

Rqc**: Quality Control of High-Throughput Sequencing Data in **Bioconductor



de Souza W, Carvalho BS, Lopes-Cendes I (2018). Journal of Statistical Software, Code Snippets, 87(2), 1–14.

Rqc

```
library(Rqc)
folder <- "/Users/Hilal Ozkilinc/Desktop/HLL/FAStQ"
list.files(path = folder, pattern = ".fastq.gz")
rqc(path = folder, pattern = ".fastq.gz")
rqcReadQualityBoxPlot(rqcResultSet)
rqcReport(result[1:10])
```

Rqc 1.24.0 - Quality Control Report

File Information

This table describes input files. `reads` column can be total number of reads (`sample=FALSE`) or sample size.

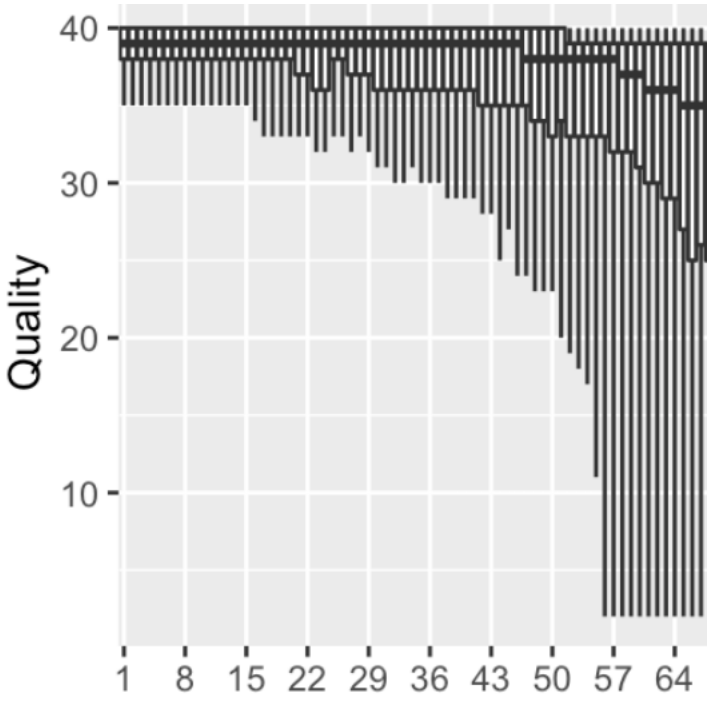
```
kable(perFileInformation(rqcResultSet))
```

filename	pair	format	group	reads	total.reads	path
B5-A4_1.fastq.gz	1	FASTQ	None	1e+06	19913623	/Users/Hilal Ozkilinc/Desktop/HLL/FAStQ

Per Read Mean Quality Distribution of Files

This plot describe an overview of per read mean quality distribution of all files

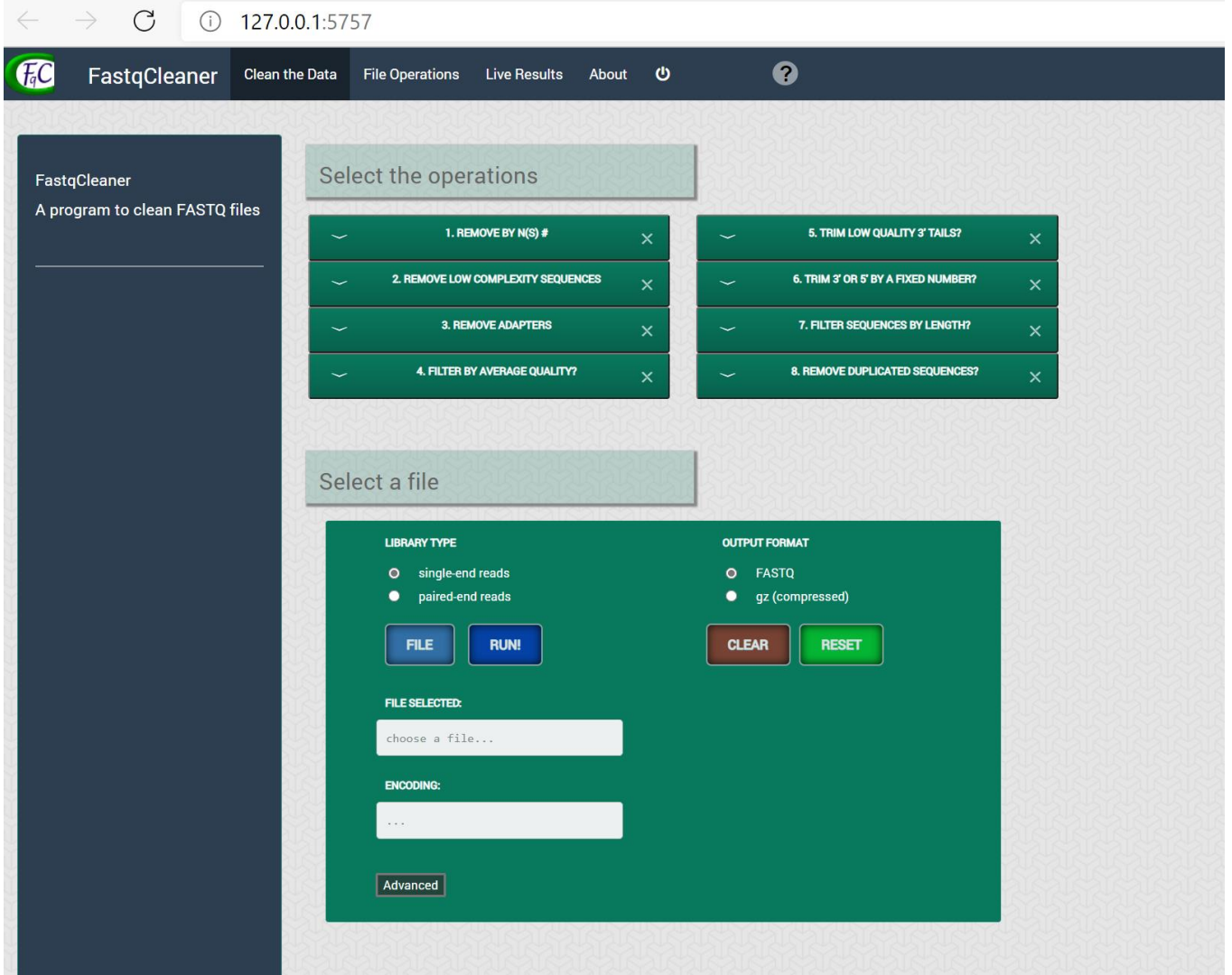
```
rqcReadQualityBoxPlot(rqcResultSet)
```



FastqCleaner

an interactive Bioconductor application for quality-control, filtering and trimming of FASTQ files

```
> library("FastqCleaner")
> launch_fqc()
```



Roser, L.G., Agüero, F. & Sánchez, D.O. 2019. *BMC Bioinformatics* **20,** 361 (2019).

ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data

Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H and Gentleman R (2009). *Bioinformatics*, **25**, pp. 2607-2608

Biostrings: Efficient manipulation of biological strings.

Pagès H, Aboyoun P, Gentleman R, DebRoy S (2020). R package version 2.58.0.

Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import

Morgan M, Pagès H, Obenchain V, Hayden N (2020). R package version 2.6.0,

seqinr: Biological Sequences Retrieval and Analysis

```
> scaf <- read.fasta("scaffolds.fasta")
> length(scaf)
[1] 668
> summary(sapply(scaf, length))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    78.0     82.0    960.5  62945.2  95644.8 734326.0
> sum(sapply(scaf, length))
[1] 42047421
> N <- function(x, t) {
+   x <- sort(x, decreasing = TRUE)
+   cx <- cumsum(x)
+   l <- sum(x)
+   return(as.numeric(x[max(which(cx <= t * l))]))
+ }
> N(sapply(scaf, length), 0.5)
[1] 213066
```

```r
library(seqinr)
x <- "/Users/Hilal Ozkilinc/Desktop/HLL/hll.fas"
read.alignment(x)
read.alignment(file = x, format = "fasta", whole.header = TRUE)

library(GenomicAlignments)
```

55:1 (Top Level) R Script

Console Terminal Jobs

~/

```
agcccaaaatggtttaaattcaaccataac--aaccatagcaatgtaaacaaattcaactcttattatactaggcctggtatt
tttacaagtgtaaataaatccaatataataggtagaagacacttttctgtaacatctccgcgtaggttaattagtgaagagtt
aaggaaattcatatccgaaaaaaaccttaatcccgtttttatatatgaagatttatcagataaagctgttaaatctagagttt
tgaatgatactcggggtcttagtggtatttatttaattttaaacaaagtgactcttgattattatataggatcagcttcaact
ggtagattccatgctagattttctaatcatttattcaattttcatgggagtaaagtagttaaaaatgcagtaaagaaagatgg
tatatcttgttttgcatttataattttagagttgtttcctgagatagtaaacaaagaaaacaataaaaaattattagatttgg
aagacttttacttaaaatctttattaccaaactataatatattaactgaagcgggttcaagctttggttataaacataacgaa
actactagattaaatatgaaaactaattatagtgaagagcgtagactagctattggtagtttaaataaaggtaaaagcttatc
tccaagtactattgaagcaatgaaacaatctgctttaaatagaataaaacctatttattctgaagaaggtattcaaaacatga
aaaagaattctaaagctatactggtttataatatggactatacagtatatggtgaatttcctagtataacagaggcatctaaa
tctttaggttgttctcaaaaaacaatttatagggctttacaaacaccaaaaaagatattaagaagacgttgaattgt------
---------taaatatgtttaaata---ttgttgtaggt--ttgaatttaaattatagtcctacgagtataaatttttatgca
atttgtaa----aaaaaa---------------ataaaaagtaaagtagaatagcccga-cggggatattgaagaaatgttta
aatttctttggcaatgttagtgaatacgatcaaataatttgtagttttgttttat---ttagataaagac-taaccttaataa
aatatagaa---ctaaaag-ttagagatcgtcggttatttatatgatcgcgacaggctgggtcactgacgggtgtctgaaatg
atacttaatgtacagtcgaagtatttagttatag------aatatgactaatagaatatacgaattcaaagttattctagctt
tttataaatgtc---ttttaatttatattaagtata---------------tttgtatgttttaccttacggtcaaatgtcgc
tatgaggtattgacatgtcagtatctacattttgtattctgcctgtttctgggattaaaccagataaaaggtttatggctatg
tttattgggtttattgatggtgatggttattttgatattggtgagcaaaaacaatataataaaataactaaaaccttagttaa
tagtactattagaattcgtttagctagtaatgtcaatgttagagatttacccttgttagaatattttgtggaagttctaggag
taggtaaaatatctaacatgtcagtcggaagagaacaagttagagtcatgttctctaaaaaagacttagtaacggtaatactg
cctttgattaaactttataacctacagtttttaacttctcaacgtgtaaaacagtttgctcttgttaattatatattagaaaa
```

BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs.

Pagès H (2020). R package version 1.58.0,

```r
library(BSgenome)
available.genomes()
BiocManager::install("BSgenome.Scerevisiae.UCSC.sacCer1")
Scerevisiae
seqnames(Scerevisiae)
seqlengths(Scerevisiae)
Scerevisiae$chr1
letterFrequency(Scerevisiae$chr1, "GC", as.prob = TRUE)
dnaseq <- DNAString("ACGTACGT")
matchPattern(dnaseq, Scerevisiae$chr1)
dnaseq == reverseComplement(dnaseq)
dnaseq
pairwiseAlignment()
trim()
```

```
> Scerevisiae
Yeast genome:
# organism: Saccharomyces cerevisiae (Yeast)
# genome: sacCer1
# provider: UCSC
# release date: Oct. 2003
# 17 sequences:
#   chr1  chr2  chr3  chr4  chr5  chr6  chr7  chr8  chr9  chr10
#   chr11 chr12 chr13 chr14 chr15 chr16 chrM
> seqnames(Scerevisiae)
 [1] "chr1"  "chr2"  "chr3"  "chr4"  "chr5"  "chr6"  "chr7"
 [8] "chr8"  "chr9"  "chr10" "chr11" "chr12" "chr13" "chr14"
[15] "chr15" "chr16" "chrM"
> seqlengths(Scerevisiae)
   chr1    chr2    chr3    chr4    chr5    chr6    chr7    chr8
 230208  813136  316613 1531914  576869  270148 1090944  562639
   chr9   chr10   chr11   chr12   chr13   chr14   chr15   chr16
 439885  745446  666445 1078173  924430  784328 1091285  948060
   chrM
  85779
> dnaseq <- DNAString("ACGTACGT")
> matchPattern(dnaseq, Scerevisiae$chr1)
Views on a 230208-letter DNAString subject
subject: CCACACCACACCCACACACCCACACAC...TGTGGTGTGGGTGTGGTGTGTGTGGG
views:
      start   end width
  [1] 57933 57940     8 [ACGTACGT]
> dnaseq == reverseComplement(dnaseq)
[1] TRUE
> dnaseq
8-letter DNAString object
seq: ACGTACGT
```

pegas: Population and Evolutionary Genetics Analysis System

Paradis E. 2010.. *Bioinformatics* **26:** 419–420

Untitled1*

```r
library(pegas)
pop <- read.dna("hll.fas", format = "fasta")
ss <- site.spectrum(pop)
plot(ss)
nuc.div(pop)
tajima.test(pop)
```

7:1 (Top Level) R Script

Console Terminal Jobs

C:/Users/Hilal Ozkilinc/Desktop/HLL/

```
> library(pegas)
> pop <- read.dna("hll.fas", format = "fasta")
> ss <- site.spectrum(pop)
Warning message:
In site.spectrum.DNAbin(pop) :
  3589 sites with more than two states were ignored
> plot(ss)
> nuc.div(pop)
[1] 0.1833571
> tajima.test(pop)
$D
[1] -2.450908

$Pval.normal
[1] 0.01424962

$Pval.beta
[1] 0.0002641515
```

Environment History Connections Tutorial

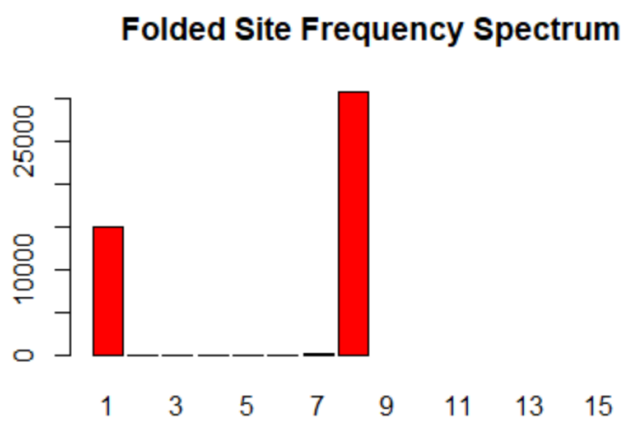
Import Dataset List

R Global Environment

Data	
dnaseq	<Object with null pointer>
scaf	List of 668
Values	
pop	Large DNAbin (1860922 elements, 1.9 M…
ss	'spectrum' int [1:16] 15016 57 6 1 3 8…
Functions	

Files Plots Packages Help Viewer

Zoom Export



Untitled1*

Source on Save | Run | Source

```r
library(pegas)
pop2 <- read.dna("Endo_final.fas", format = "fasta")
h <- haplotype(pop2)
h
d <- dist.dna(h, "N")
nt <- rmst(d)
nt
plot(nt)
```

2:1 (Top Level) R Script

Console | Terminal | Jobs

C:/Users/Hilal Ozkilinc/Desktop/HLL/

```
Haplotypes extracted from: pop2

    Number of haplotypes: 19
         Sequence length: 419

Haplotype labels and frequencies:

    I    II   III    IV     V    VI   VII  VIII    IX     X
    5     1     1     1     1     1     1     1     1     1
   XI   XII  XIII   XIV    XV   XVI  XVII XVIII   XIX
    1     1     1     1     2     1     7     1     1
> d <- dist.dna(h, "N")
> nt <- rmst(d)
> plot(nt)
```

Environment | History | Connections | Tutorial

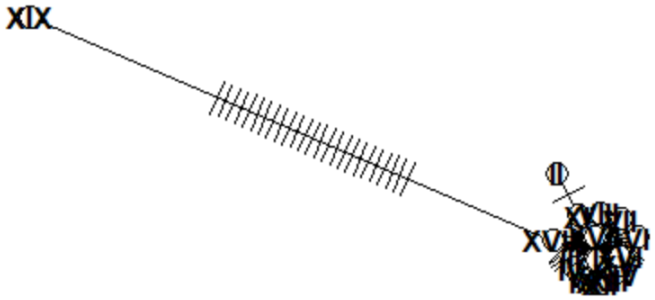
Import Dataset | List

R | Global Environment

tajd	List of 3
Values	
d	'dist' Named num [1:171] 2 0 0 0 1 1 1…
h	'haplotype' raw [1:19, 1:419] c c c c …
nt	'haploNet' num [1:18, 1:3] 1 1 1 1 1 1…
pop	'DNAbin' raw [1:30, 1:419] c c c c ...
pop2	'DNAbin' raw [1:30, 1:419] c c c c ...

Files | Plots | Packages | Help | Viewer

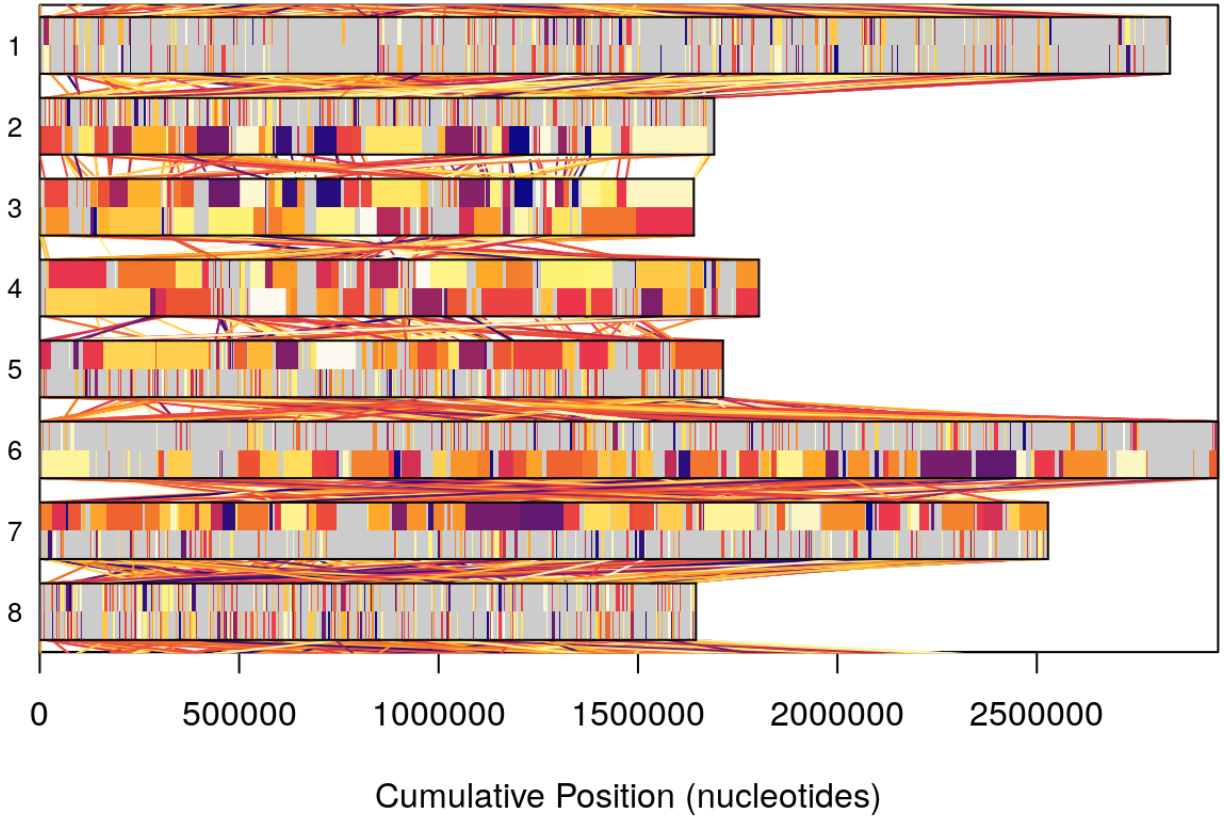
Zoom | Export



DecipheR

Wright ES (2016). "Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R." *The R Journal*, **8**(1), 352-359.

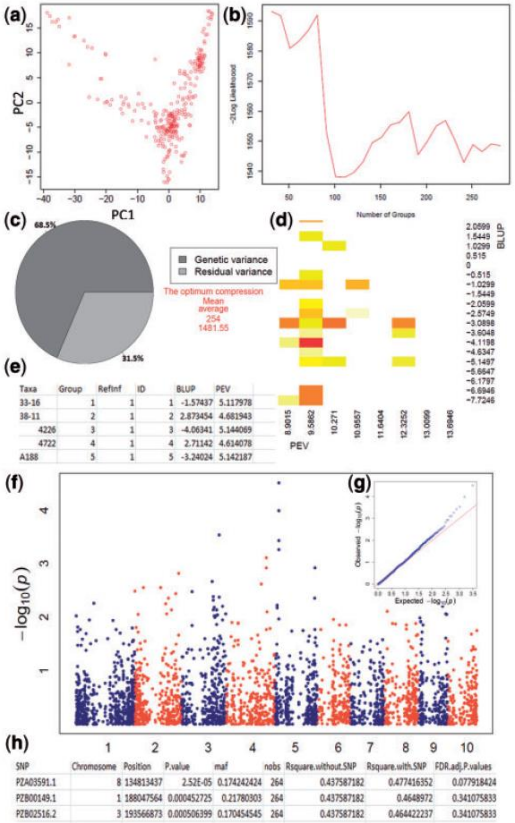
```
FindSynteny()
```



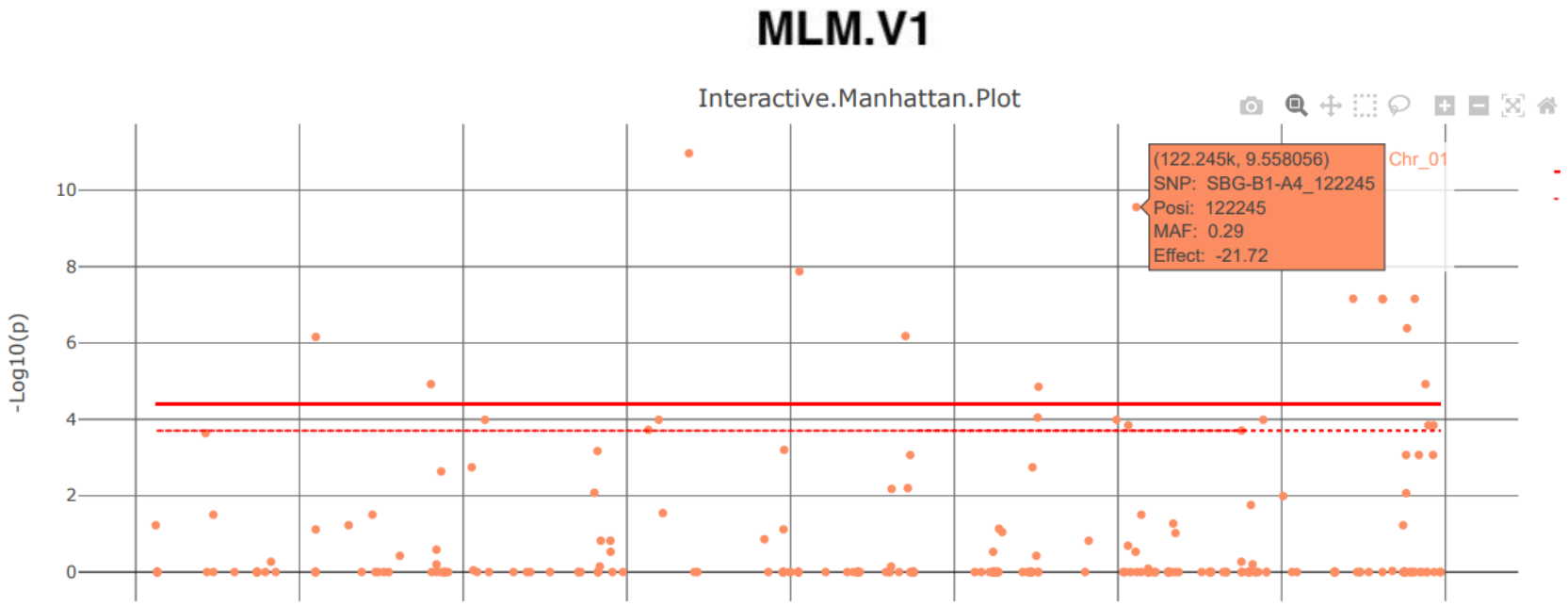
GAPIT: genome association and prediction integrated tool

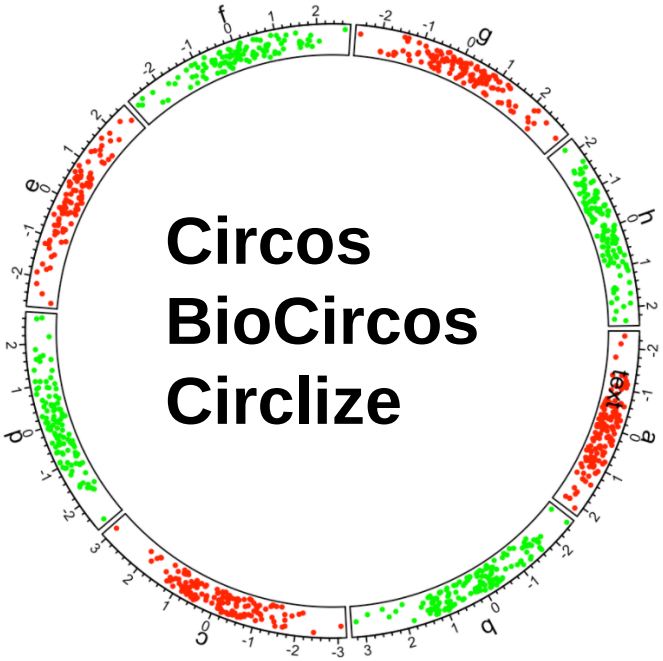
Lipka et al. (2012) Bioinformatics.15;28(18):2397-9.

- GAPIT çok işlevli genom çapında bağdaştırma çalışmalarında kullanılan çok işlevli R paketi
- Genotip ve fenotip türünde veriyi çalıştırmak için HapMap ve numerik formatta veri seti kullanılir.

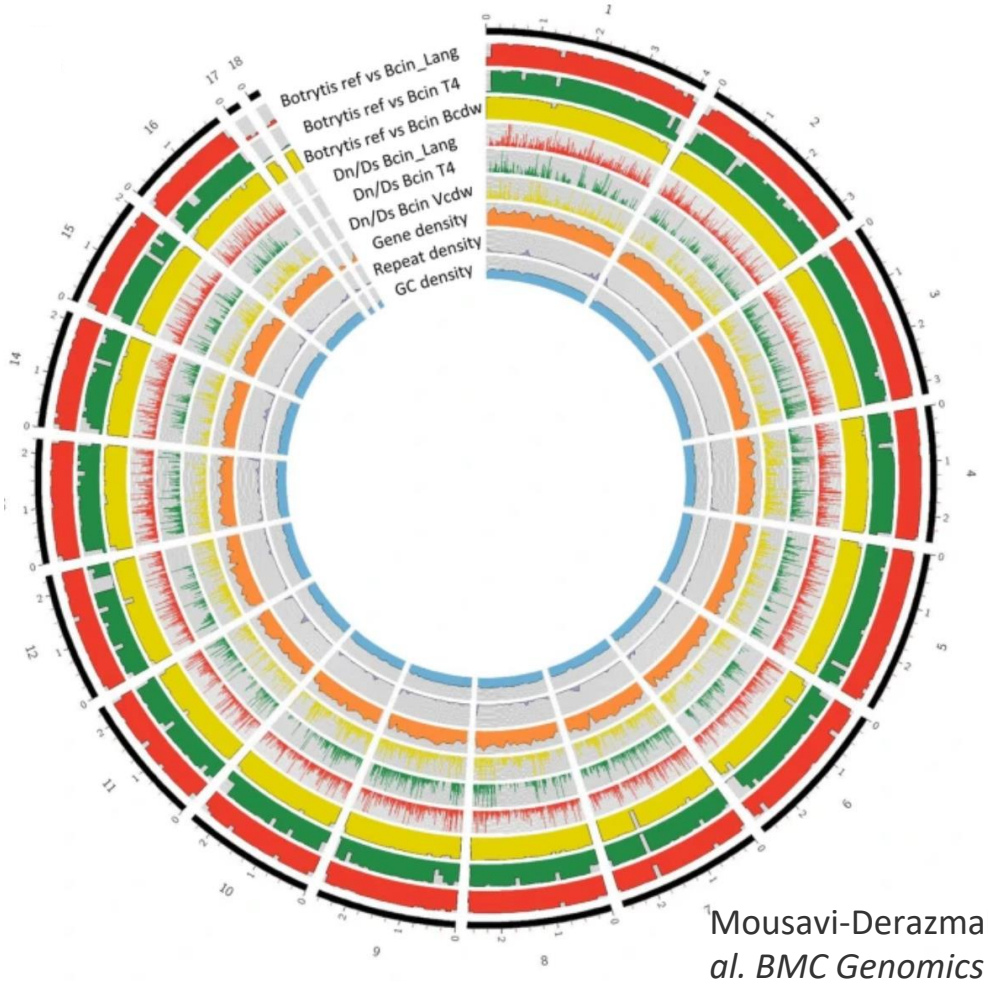


GAPIT (mass, multtest, gplot, compiler, scatterplot3D, LDHeatmap)– Tek komuttan elde edilebilen çıktılardan bazıları





Circos
BioCircos
Circlize



Mousavi-Derazmahalleh *et al. BMC Genomics* **20,** 385 (2019).

…ve daha pekcok paket

- Genomik
- Trankriptomik
- Evrimsel genetic
- Populasyon genetigi
- Filogenetik
- Epigenetik
- …

OzkilincLab
Fungal Evolutionary Genetics

Home Projects Most Recent Publications Lab Head Lab Team Previous Lab Members Photos Events Announcements Lab Meetings

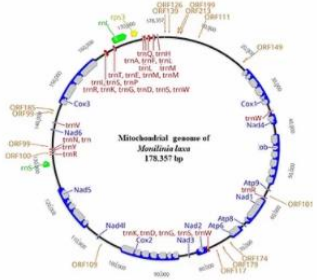
Welcome to the Özkılınç Lab

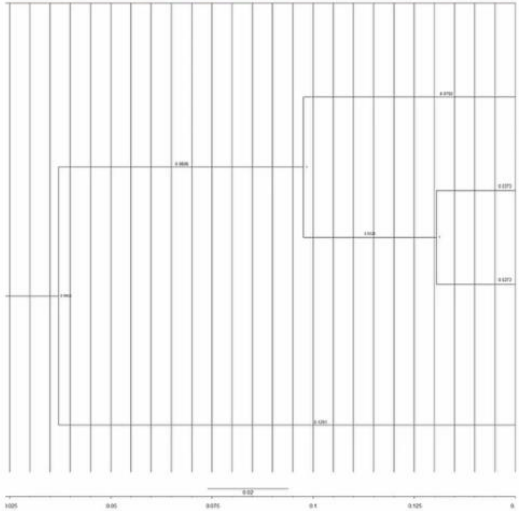
Welcome to our website!

Our group is working at Canakkale Onsekiz Mart University, Faculty of Arts and Sciences, Dept. of Molecular Biology and Genetics, Canakkale/ Turkey

Our group has been directing evolutionary biology questions on fungal plant pathogens by using population genetics, phylogenetics and genomics approaches.

TESEKKURLER





https://ozkilinclab.wixsite.com/website

hilalozkilinc@comu.edu.tr