Dr. Rachel’s Assignment - Unisabana

Jeysson Sanchez-Suarez

3/7/2019

This R markdown shows the results of genome assembly exercise of Lactococcus lactis from a illumina pair-end files. For the alignment the reference genome was obtained from NCBI database with access number NC_002662.1.

Linux Shell Scripts

A script (variant_calling.sh) with loop instructions for “fastqc”, “trimmomatic”, “alignment”, “bam files creation” and “bcf tools” orders was created and runned. This script was successful for variant calling.

After, the genome assembly was done by ABySS program, using the script “alignment.sh”.

See git-hub repository for scripts details in: https://github.com/jfsanchezs/assignment_drrachel (https://github.com/jfsanchezs/assignment_drrachel)

Finally, the obtained data was analyzing by R scripts.

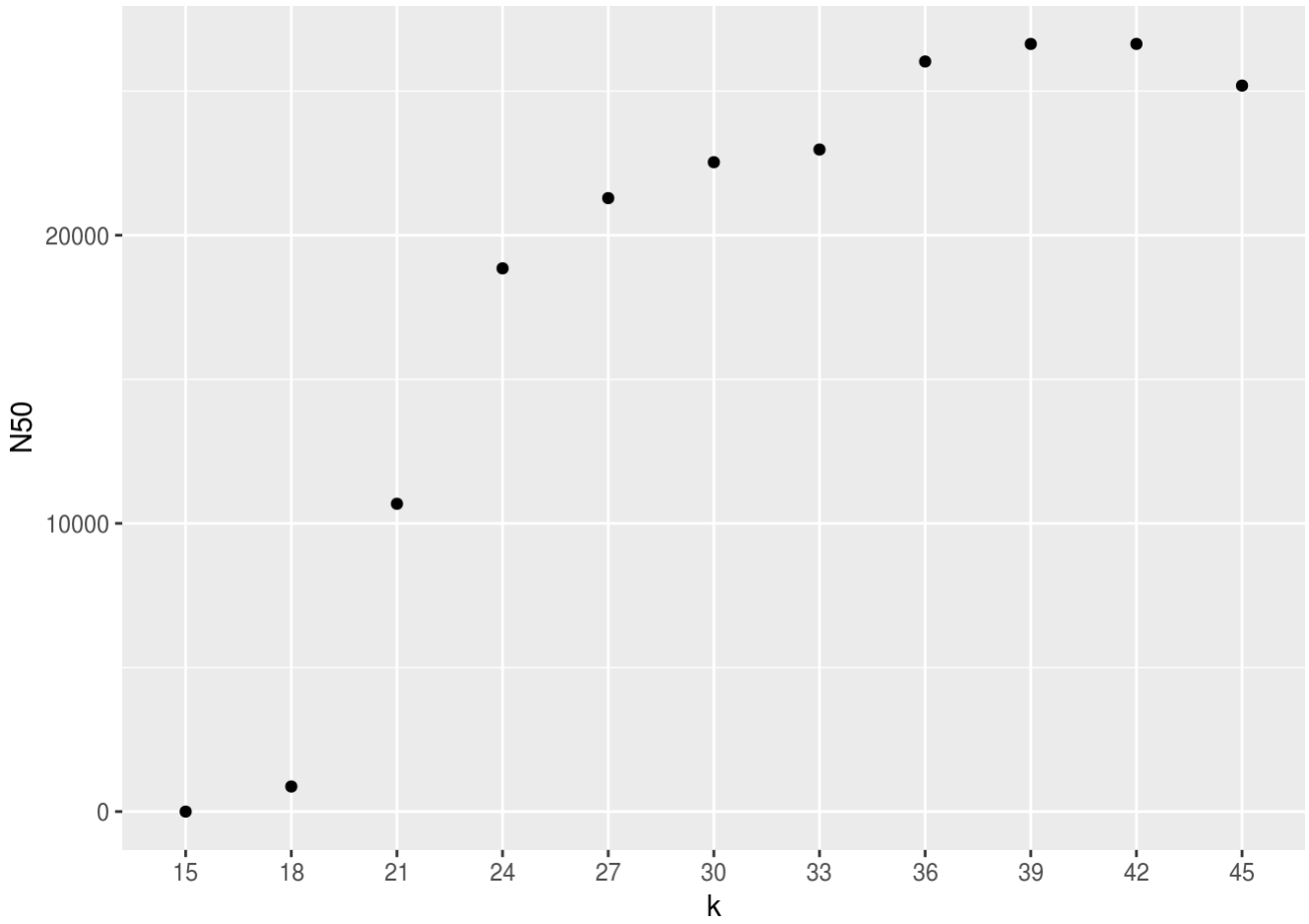
R Scripts Results

```
setwd("~/jeysson/")
#Load package
library(tidyr)

#Load data (create object)
assembly_llactis <- read.csv("data/trimmed/allteststats", sep="\t")

#Separate "k" values, create a new column.
assembly_llactis2 <- separate(assembly_llactis,n,c("name1","n"), sep=":") %>%
  separate(name1,c("k","name1"),sep="-") %>% separate(k,c("name2","k"),sep="k")
```

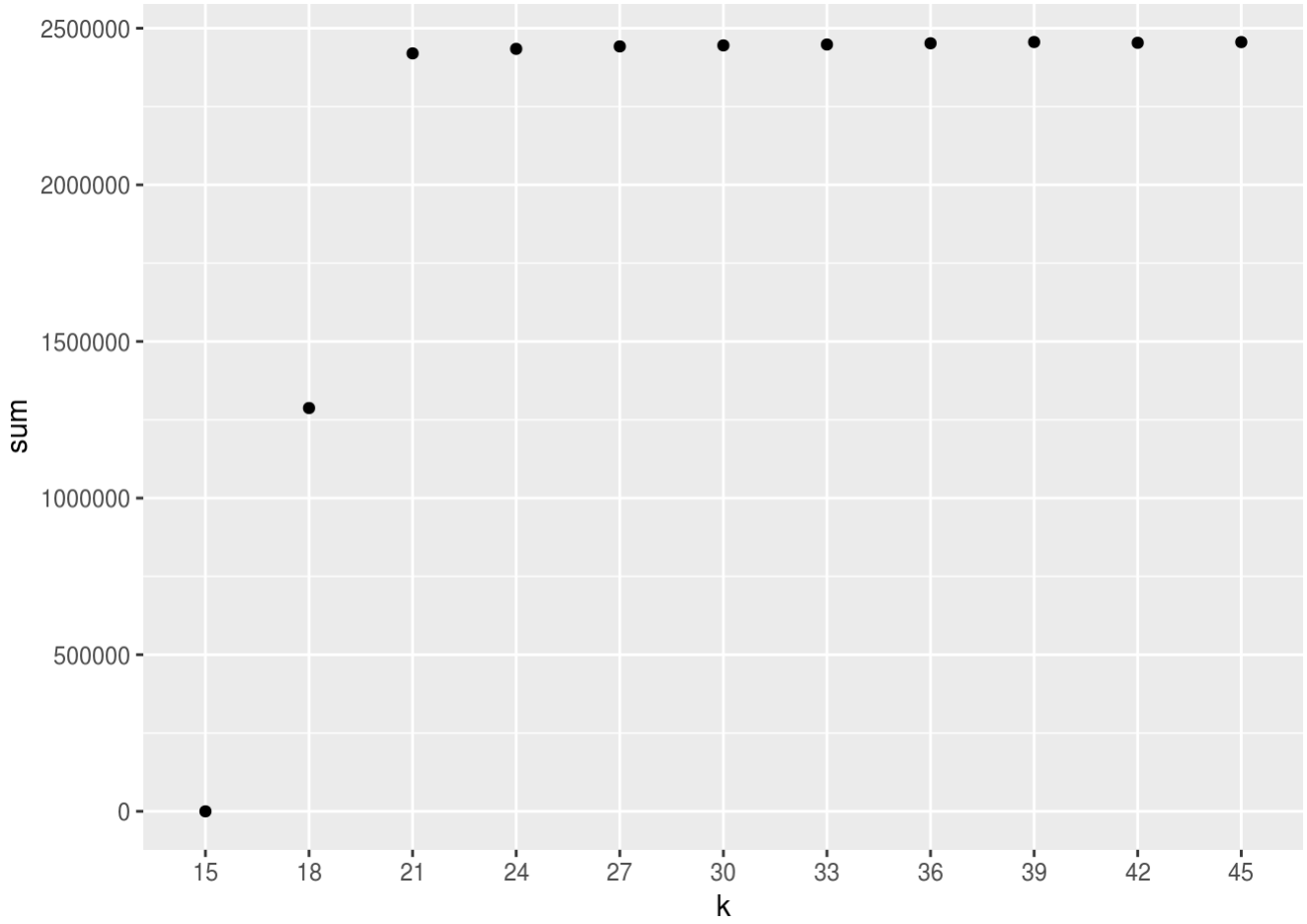
Plots of Genome Assembly Results

The plots were created using ggplot2 program.

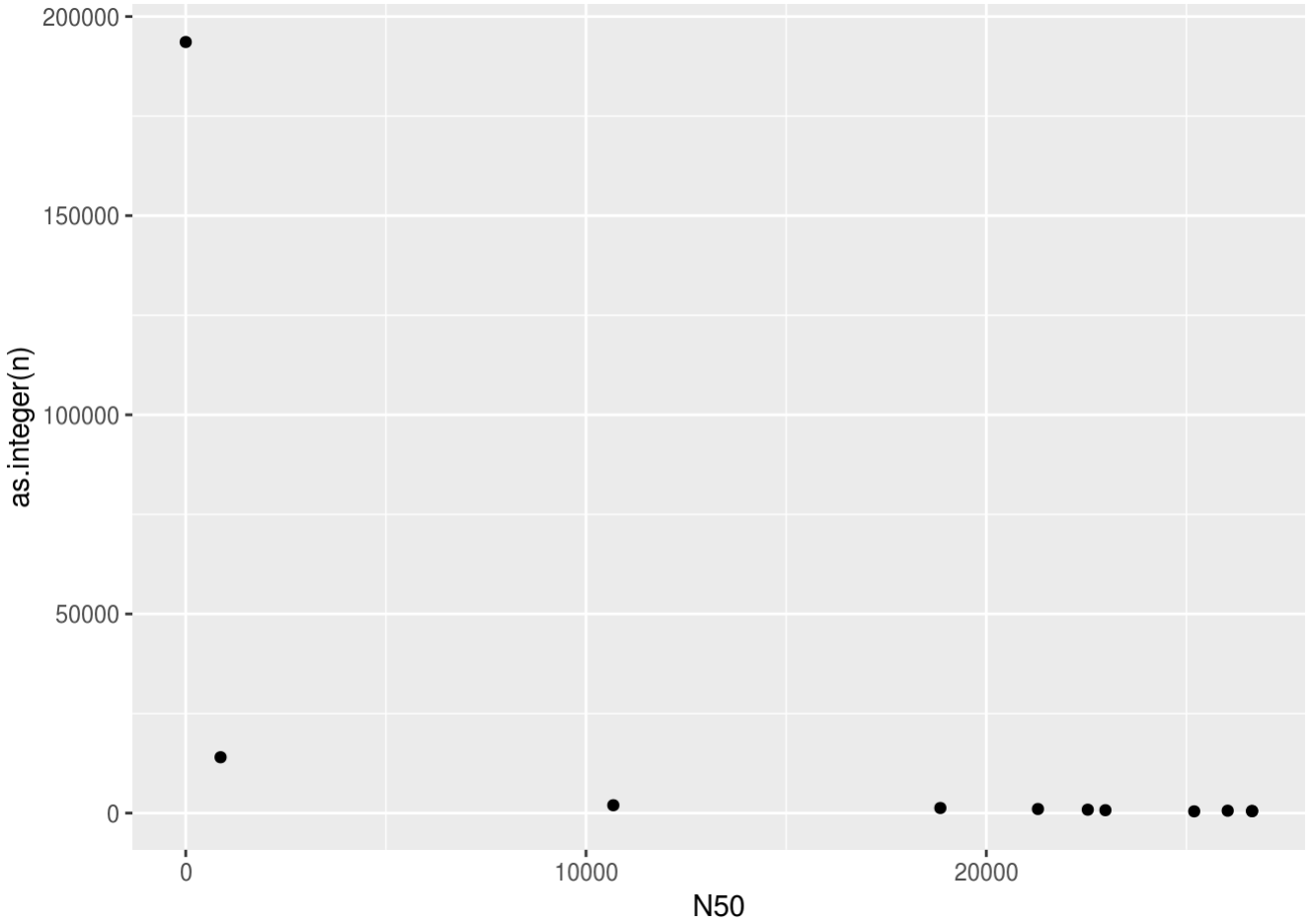
```
#Plot N50 and total length as a function of k
library(ggplot2)
ggplot(assembly_llactis2,aes(k,N50))+geom_line()+geom_point()
```



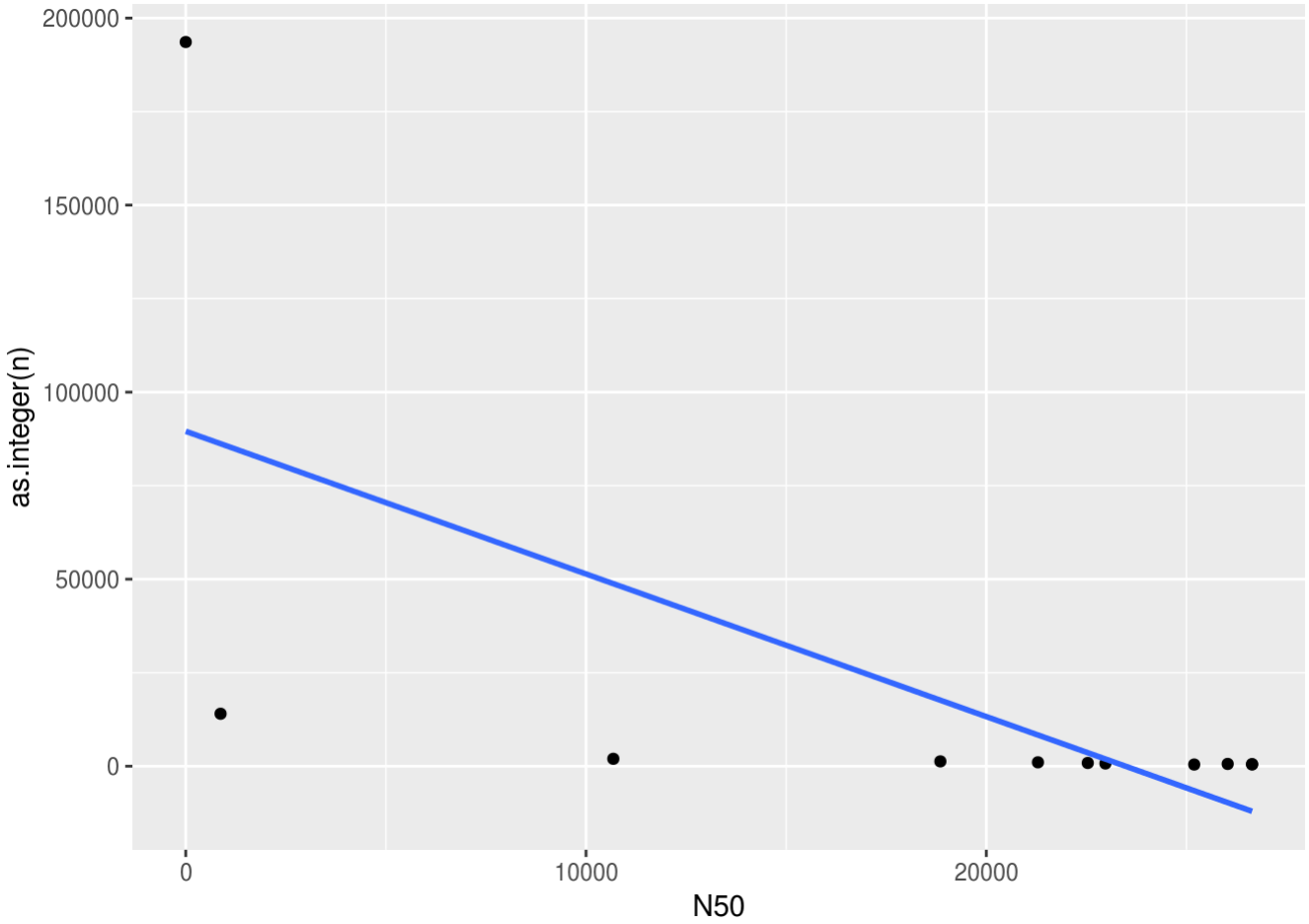

```
ggplot(assembly_llactis2,aes(k,sum))+geom_point()
```




```
ggplot(assembly_llactis2,aes(N50,as.integer(n)))+geom_point()
```




```
ggplot(assembly_llactis2,aes(N50,as.integer(n)))+geom_point()+
  geom_smooth(method="lm", se=FALSE)
```


200000
150000
100000
50000
0
as.integer(n)
0
10000
20000
N50