Mapping of Cas12a PAMs and base editing sites in the C.phytofermentans genome

Andrew Tolonen

may23

Introduction

The goal of this script is to map the distribution of Cas binding and base editing sites in the C.phytofermentans ISDg genome.

Methods

Setup and file I/O

```
rm(list = ls());
setwd("/home/tolonen/Github/actolonen/Seq_analysis_R");

library(tidyverse);
library(plotly);
library(curl);
library(seqinr); # read.fasta
library(pepliner); # fasta_tidier

mytheme = theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12),
                axis.title.x = element_text(size = 16), axis.title.y = element_text(size = 16),
                aspect.ratio =1/1.61,
                panel.grid.minor=element_blank(), panel.grid.major=element_blank());

# get genome FNA file from genbank
fnafilegz = curl_download("https://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Lachnoclostridium_phyt
```

Step 1: format sequence data for analysis

```
# convert file to data.frame with 2 cols: Gene name, sequence
geneseqs = read.fasta(file = "./genome.fna.gz", seqtype="DNA", as.string = TRUE, strip.desc = TRUE, whol
  fasta_tidier() %>% # convert to data.frame
  mutate(Gene_name = str_extract(string=ID, pattern="Cphy_[\\d]+")) %>% # make column of gene names
  select(Gene_name, Sequence);

# count number of PAMS in each gene
geneseqs = geneseqs %>%
  mutate(Number_PAMs = str_count(string=Sequence, pattern="ttt[agc]"));

# count number of PAMs associated with stop codons in each gene
```

Step 2: plot data

```
# plot distribution of number of PAMs per gene

myplot = ggplot(geneseqs, aes(x=Number_PAMs)) +
 geom_histogram(binwidth=2, fill="#68A2AD")+
 ylab("Cas12a PAMs per gene")+
 xlab("Number genes")+
 mytheme;

myplot
```

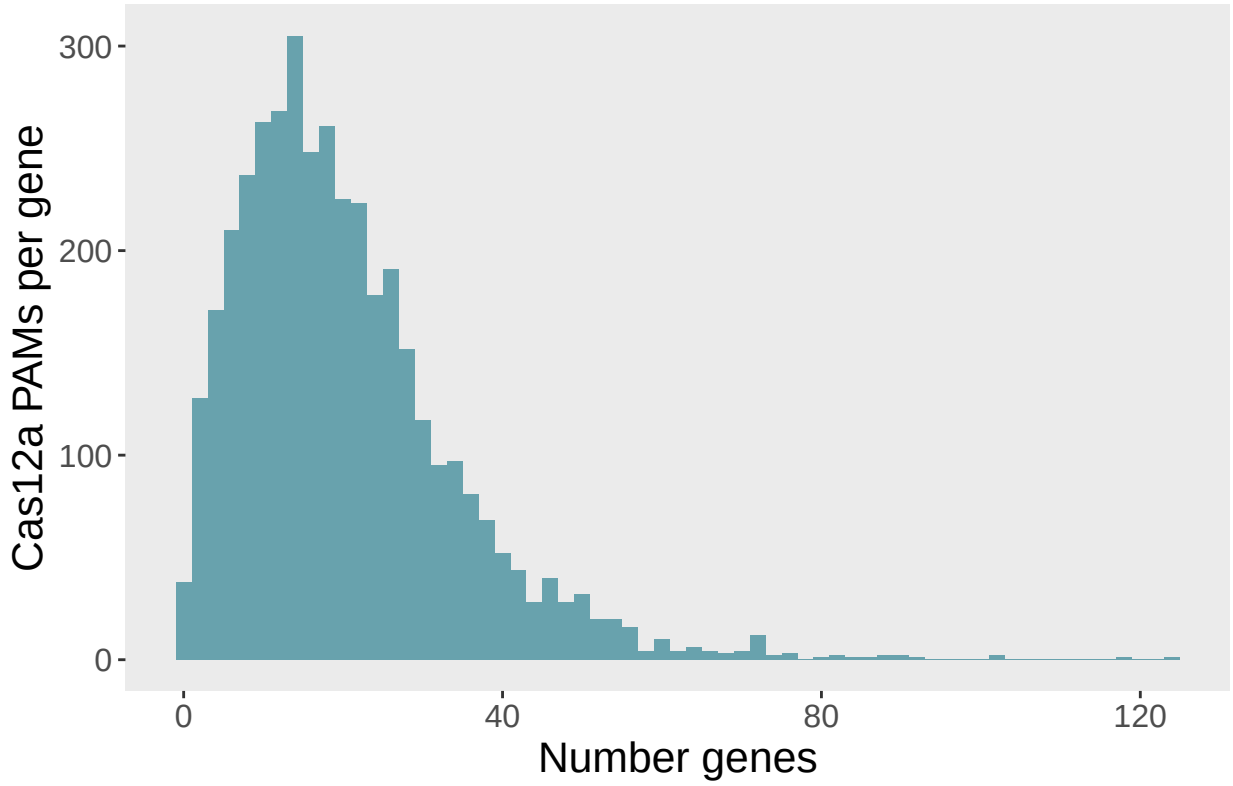

Fig 1. Number of Cas12a PAMs (5'-TTTV-3') per gene in the C. phytofermentans ISDg genome. Among the 3902 genes in the genome, 13 genes have no PAM sites.