Clavibacter sequencing coverage in metagenomics samples at subspecies resolution

Erik Díaz

10/23/2022

Clavibacter michiganensis es una bacteria gram positiva patógena que afecta negativamente el rendimiento de algunos cultivos como chile, tomate, trigo, papa y maíz. Su correcta identificación empleando ténicas moleculares no es trivial debido a su baja abundancia comparada con otras bacterias y porque posee un genoma compacto.

Un enfoque empleado comúnnmente para estimar la presencia y la abundancia de Clavibacter michiganensis es la amplificación de algunos marcadores genéticos. No obstante, aunque esta técnica es adecuada para identificar este patógeno a nivel de especie, si se quiere identificar la subespecie o cepa particular que afecta algún cultivo existen técnicas alternativas como la secuenciación del microbioma que pueden ayudar a atacar este problema. Un reto de de la identificación de taxa a nivel de subespecie en muestras metagenómicas es que los clasificadores taxonómicos es que usualmente se emplean carecen poder de resolusión a nivel de subespecie.

Una estrategia adecuadoa para estimar la abundancia de diferentes subespecies o cepas de Clavibacter michiganensis en muestras metagenómicas es mapear las lecturas de secuenciación de la especie contra genomas de referencia ensamblados a partir de aislados. El objetivo de este proyecto fue estimar la cobertura de secuenciación de cinco subespecies de Clavibacter michiganensis, capturando en total el la diversidad genética de 28 cepas de esta bacteria.

En las siguientes secciones describiré el enfoque analítico para estimar la cobertura de cada seubespecie/cepa en cinco datasets de secuenciación de metagenoma extraidos de plantas de chile, tomate, papa, trigo y maíz.

1) Alineamos las lecturas de secuenciación que el clasificador taxonómico Kraren asignó como Clavibacter contra los 28 genomas de referencia. Este script se empleo de manera independiente para cada uno de los cinco data sets

```
#!/bin/sh

for k in *.fna

    do
        for i in *-clav-1.fq
        do
            echo "working with file $i"
            base=$(basename $i -clav-1.fq)
            echo "base name is $base"
            i=${base}-clav-1.fq
            j=${base}-clav-2.fq
            bwa mem -M -t 20 $k $i $j > ${base}_$k.sam
        done
    done
```

2) Convertimos los archivos de alineamiento (SAM) a formato BAM y los ordenamos por coordenadas genómicas

```
#!/bin/sh

for i in *.sam
    do
        echo "working with file $i"
        base=$(basename $i .sam)
        echo "base name is $base"
        i=${base}.sam
        samtools view -bS $i | samtools sort > ${base}.bam
    done
```

3) Estimamos la cobertura en ventanas del genoma donde la cobertura no es cero empleando el formato BedGraph y lo exportamos en formato tabular (TSV)

```
#!/bin/sh

for i in *.bam
do
    bedtools genomecov -ibam $i -bg > ${i}.tsv
done
```

4) Concatenamos los alineamientos de cada biblioteca de secuenciación contra cada genoma resultando en una base de datos de dimensiones número de biblioteca x númer de genoma de referencia.

```
#!/bin/sh

awk 'FNR==1{if (NR==1) print "filename", $0; next} {print FILENAME, $0}' *.tsv > capsicum_coverage.tsv
```

EMpleando el lenguaje de programación R y apoyandonos de los paquetes tidyverse y windowscanr Estimamos y visualiamos la cobertura mediana por cada ventana de 10KB colapsando la información de todas las bibliotecas de secuenciación de cada dataset

5) Escribimos una función para cargar los datos y manipular las columnas de manera que extraigamos la información adecuada como ID de cada bilioteca, contra cual genoma se alinearon las lecturas, las coordenandas de la ventana genómica y la cobertura estimada.

```
load_coverage_data <- function(i) {
  read_tsv(i, col_names = FALSE) %>%
    rename(library_ref = X1, start = X2,
           end =  X3, coverage = X4) %>%
    separate(library_ref, into = c("genome", "chromosome"),
             sep = "\ ") %>%
    separate(genome, into = c("library", "stuff"),
             sep = "_Clavibacter_michiganensis_subsp_") %>%
    separate(stuff, into = c("ref_genome", "format"),
             sep = ".fna.bam.tsv") %>%
    select(library, ref_genome, chromosome, start, end, coverage)
}
```

6) Escribimos una función para estimar la mediana de la cobertura de todas las bibliotecas de cada dataset por posición y por cromosoma

```
median_cove <- function(i) {
  i %>%
    group_by(chromosome, start) %>%
    summarise(mean_cove = mean(coverage))
}
```

7) Escribimos una función para estimar la cobetura de secuenciación en ventanas de 10kb

```
sliding_windows <- function(i) {
  winScan(x = i,
          groups = "chromosome",
          position = "start",
          values = c("mean_cove"),
          win_size = 10000,
          win_step = 5000,
          funs = c("median"))

}
```

8) Escribimos una función para vizualizar la cobertura en las ventanas de 10kb

```
genome_wide_cove <- function(i) {
  i %>%
    ggplot(aes(win_start/1000000, mean_cove_median, color = chromosome)) +
    geom_line() +
    labs(y = "Genomic position (Mb)", x = "Mean 10-kb coverage") +
    scale_color_viridis_d() +
    facet_wrap(~name, scales = "fixed", ncol = 4) +
    theme(panel.background = element_rect(fill = NA),
          axis.title.x=element_text(size = 8),
          axis.text.x=element_text(size = 8),
          axis.text.y=element_text(size = 8),
          axis.title.y=element_text(size = 8),
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8),
          strip.background = element_rect(fill = "#f1f2f6"),
          legend.position = "none")
}
```

9) Para estimar visualmente cuáles regiones del genoma de Clavibacter michiganensis poseen variabilidad en la cobertura, lo cuál podría ayudar a separar las lecturas que son específicas a cada genoma de referencia, estimamos la varianza de cada ventana y la visualizamos en las coordenadas genómicas.

```
Variance_p_windows <- function(i){
  i %>%
    group_by(win_start) %>%
    summarise(variance_ok = var(mean_cove_median, na.rm = T)) %>%
    ggplot(aes(win_start/1000000, variance_ok)) +
    geom_line() +
    theme(panel.background = element_rect(fill = NA),
          axis.title.x=element_text(size = 8),
          axis.text.x=element_text(size = 8),
          axis.text.y=element_text(size = 8),
          axis.title.y=element_text(size = 8),
          legend.text = element_text(size = 8),
          legend.title = element_text(size = 8),
          strip.background = element_rect(fill = "#f1f2f6"),
          legend.position = "none")
}
```

10) Empleamos la función para cargar los datos del data set de chile (como ejemplo)

```
pepper_data_in <- load_coverage_data("capsicum_coverage.tsv")
```

11) Partimos el data set en una lista empleando el genoma de referencia como parámetro

```
sliding_win_analysis_pepper <- pepper_data_in %>%
  group_split(ref_genome) %>%
  setNames(unique(pepper_data_in$ref_genome))
```

12) Empleamos la función para colapsar la cobertura de todas las bibliotecas por ventana genómica sobre la lista de dataframes

```
pepper_median_cove_sum <- lapply(
  sliding_win_analysis_pepper, median_cove)
```

13) Estimamos la cobertura en ventanas de 10KB con la función que definimos anteriormente

```
pepper_covergae_median_ok <- lapply(
  pepper_median_cove_sum, sliding_windows)
```

14) Unimos las listas en un solo dataframe para visualizar los datos empleando la sintaxis de tidyverse

```
bind_pepper_list <- pepper_covergae_median_ok %>%
  map_dfr(~ .x %>% as_tibble(), .id = "name")
```

15) Por último empleamos las funciones de visualización

```
pepper_viz <- genome_wide_cove(bind_pepper_list)

pepper_viz_var <- Variance_p_windows(bind_pepper_list)
```

En la siguiente sección se mostrarán las visualizaciones de cobertura de secuenciación en cada genoma de referencia. Estos resultados permitirán primero una exploración visual, de cuáles regiones podrían contener lecturas de secuenciación que se originaron de alguna subespecia o cepa específica. De igual manera, la visualización de la varianza de cobertura de secuenciación podrá ayudar a identificar regiones del genoma de Clavibacter michiganensis que pueden tener abundancias diferentes en las muestras metagenómicas.

Datos de chile:

```
knitr::include_graphics('/Users/erik/clavibacter/pepper_clavi_cove.pdf')
```

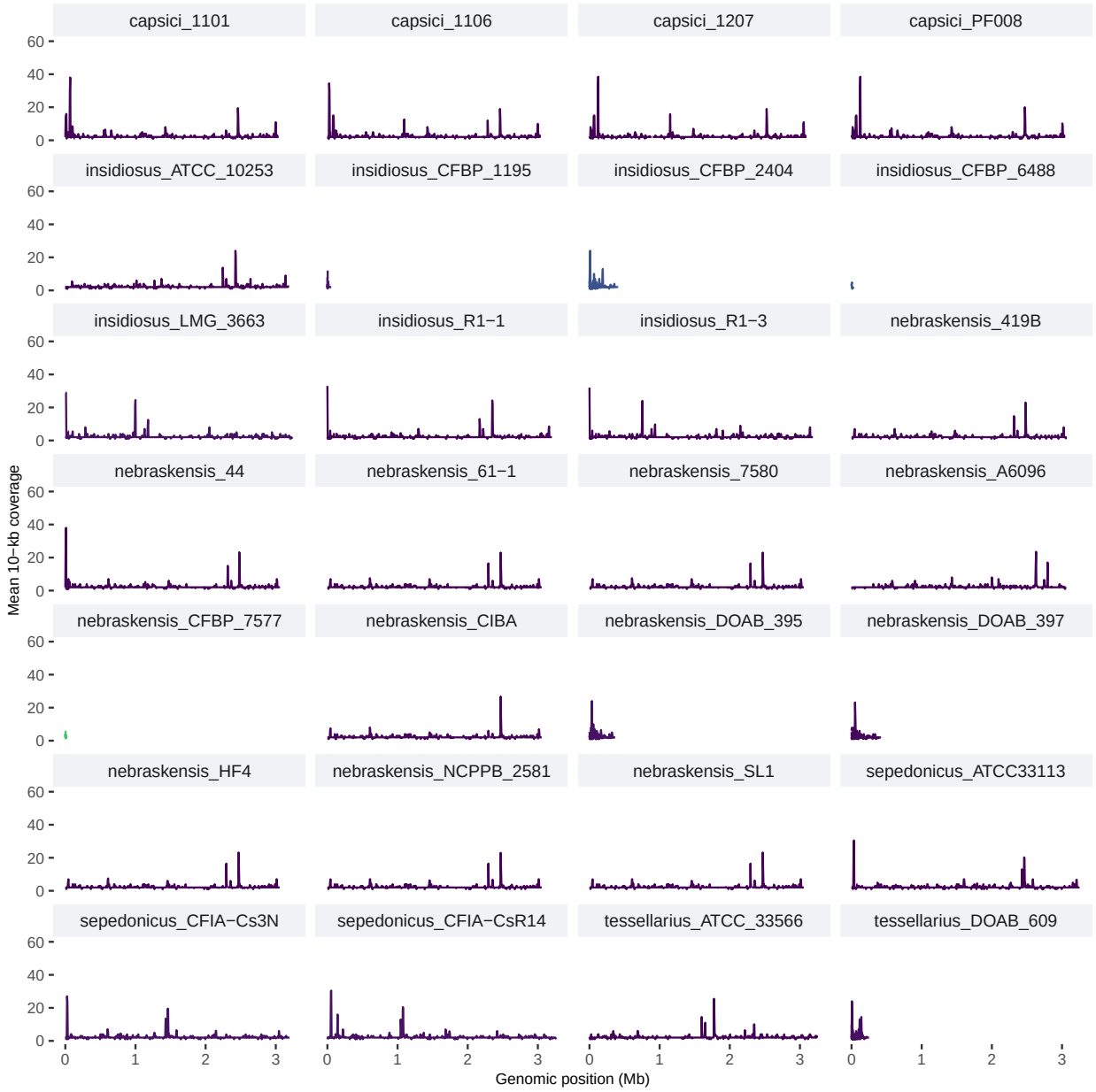

Figure 1: Cobertura de secuenciación en ventanas de 10KB revelando diferencias claras de abundancia de algunas cepas de Clavibacter michiganensis sobre otras

```
knitr::include_graphics('/Users/erik/clavibacter/variance_cove_chile.pdf')
```

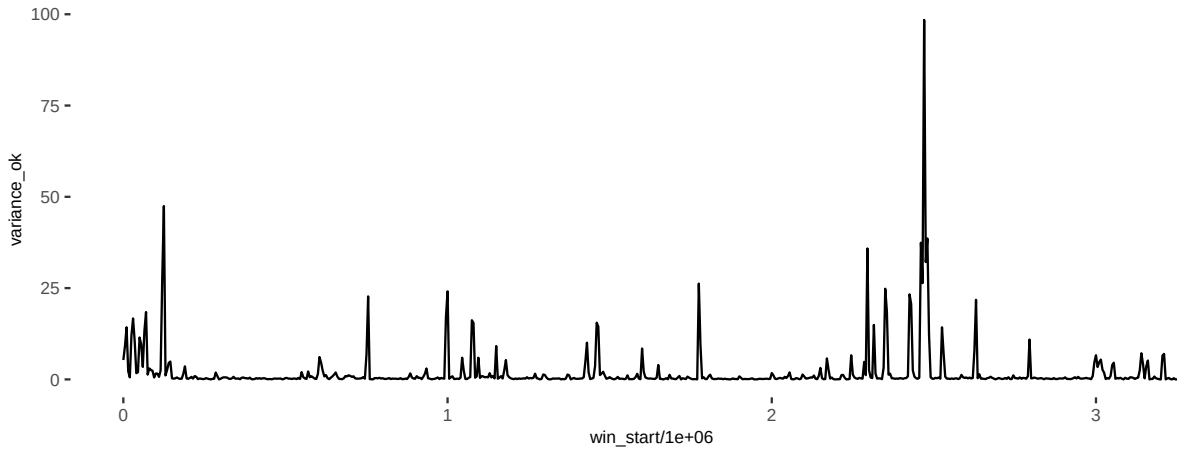

Figure 2: Varianza de cobertura de secuenciación en ventanas de 10KB revelando regiones del genoma de Clavibacter michiganensis de donde posiblemente se originan lectuas que permiten diferenciar entre subespecies o cepas

Datos de tomate:

```
knitr::include_graphics('/Users/erik/clavibacter/tomato_clavi_cove.pdf')
```

```
knitr::include_graphics('/Users/erik/clavibacter/variance_cove_tomate.pdf')
```

Datos de trigo:

```
knitr::include_graphics('/Users/erik/clavibacter/wheat_clavi_cove.pdf')
```

```
knitr::include_graphics('/Users/erik/clavibacter/variance_cove_trigo.pdf')
```

Datos de papa:

```
knitr::include_graphics('/Users/erik/clavibacter/potato_clavi_cove.pdf')
```

```
knitr::include_graphics('/Users/erik/clavibacter/variance_cove_papa.pdf')
```

Datos de maiz:

```
knitr::include_graphics('/Users/erik/clavibacter/maize_clavi_cove.pdf')
```

```
knitr::include_graphics('/Users/erik/clavibacter/variance_cove_maiz.pdf')
```

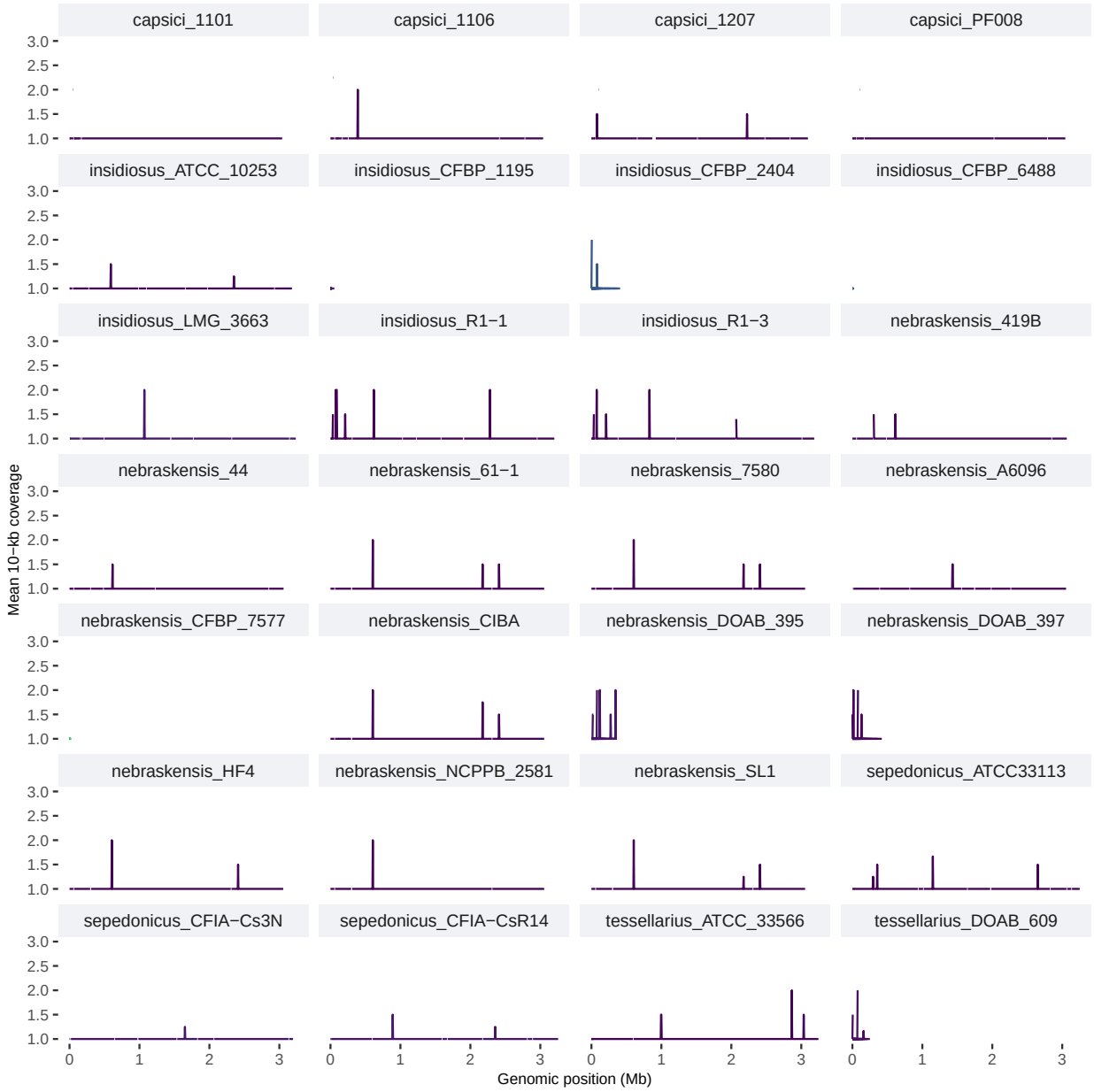

Figure 3: Cobertura de secuenciación en ventanas de 10KB revelando diferencias claras de abundancia de algunas cepas de Clavibacter michiganensis sobre otras

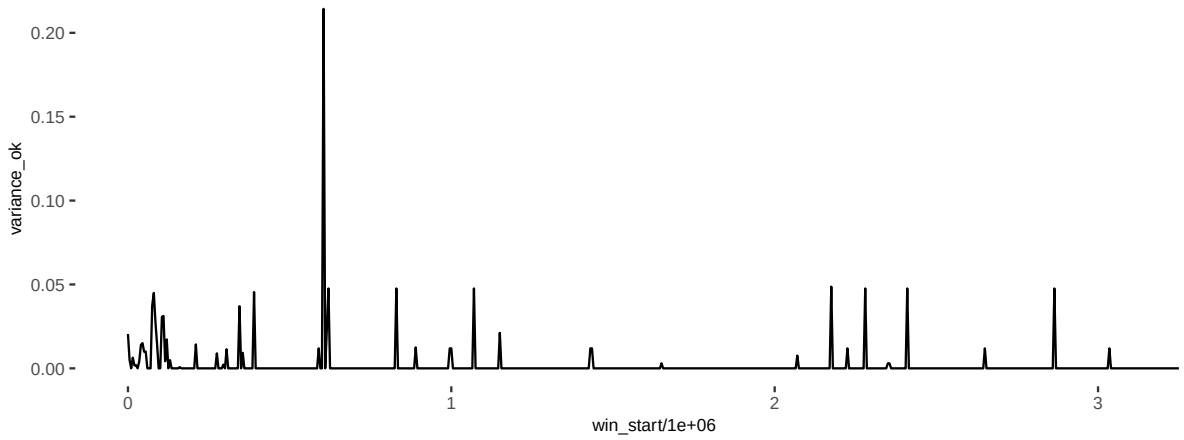

Figure 4: Varianza de cobertura de secuenciación en ventanas de 10KB revelando regiones del genoma de Clavibacter michiganensis de donde posiblemente se originan lectuas que permiten diferenciar entre subespecies o cepas

Una pregunta pertinente para descartar que las regiones con diferente cobertura correspondan a regiones homólogas entre los diferentes genomas de referencia se pretende alinear todos los genomas y estimar un mapa de homología. De esta manera se contribuirá a reducir la ambigüedad del origen de las lecturas de secuenciación cuando las diferencias entre los genomas de referencias son sutíles.

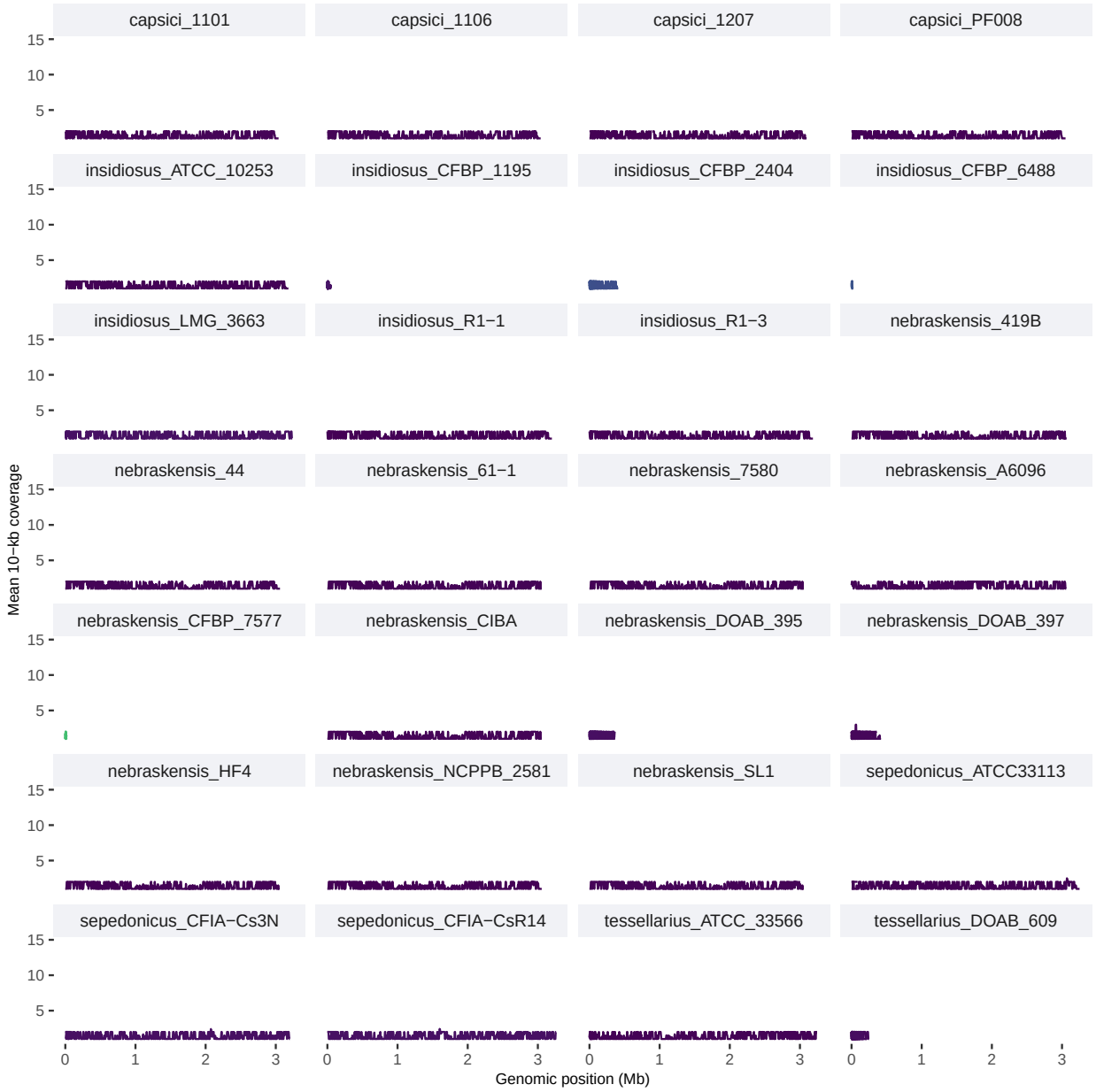

Figure 5: Cobertura de secuenciación en ventanas de 10KB revelando diferencias claras de abundancia de algunas cepas de Clavibacter michiganensis sobre otras

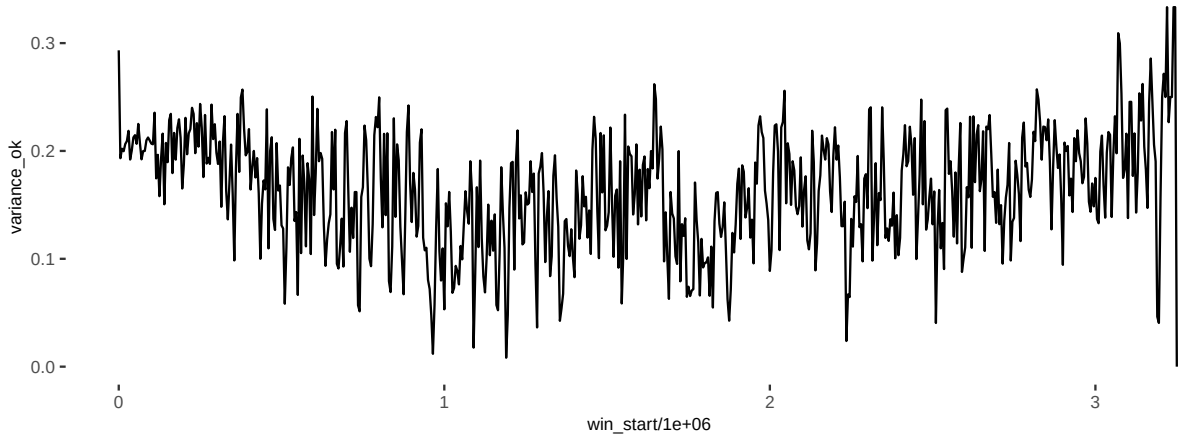

Figure 6: Varianza de cobertura de secuenciación en ventanas de 10KB revelando regiones del genoma de Clavibacter michiganensis de donde posiblemente se originan lectuas que permiten diferenciar entre subespecies o cepas

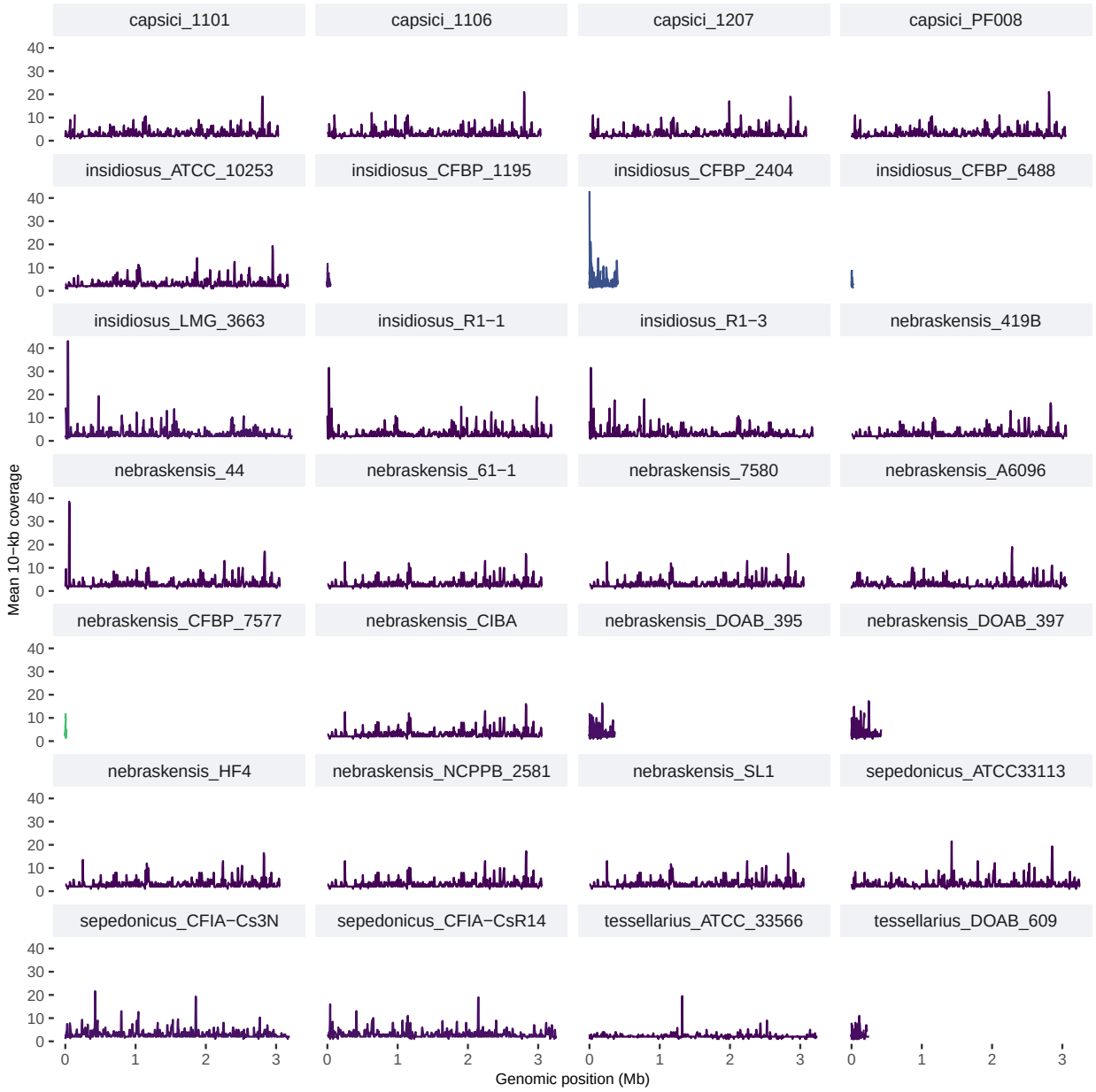

Figure 7: Cobertura de secuenciación en ventanas de 10KB revelando diferencias claras de abundancia de algunas cepas de Clavibacter michiganensis sobre otras

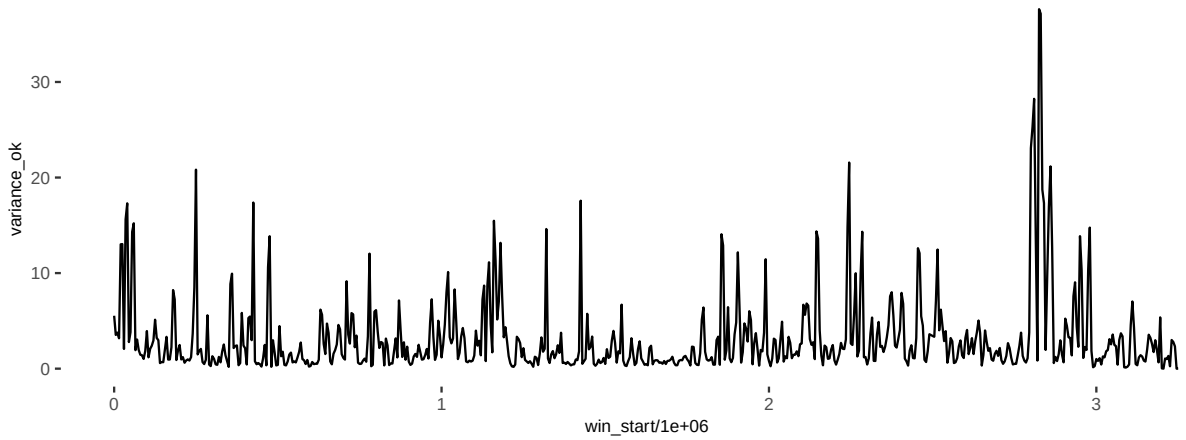

Figure 8: Varianza de cobertura de secuenciación en ventanas de 10KB revelando regiones del genoma de Clavibacter michiganensis de donde posiblemente se originan lectuas que permiten diferenciar entre subespecies o cepas

Figure 9: Cobertura de secuenciación en ventanas de 10KB revelando diferencias claras de abundancia de algunas cepas de Clavibacter michiganensis sobre otras

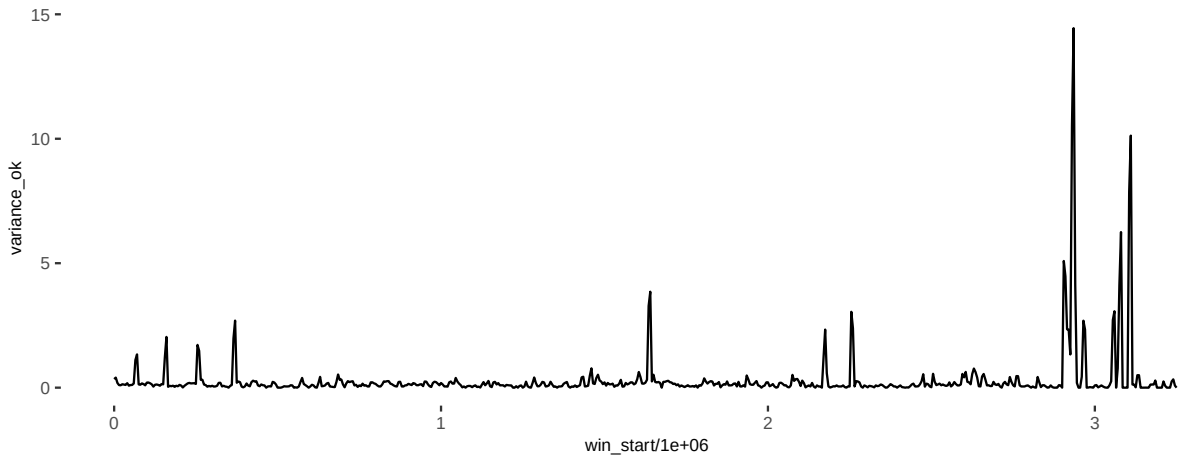

Figure 10: Varianza de cobertura de secuenciación en ventanas de 10KB revelando regiones del genoma de Clavibacter michiganensis de donde posiblemente se originan lectuas que permiten diferenciar entre subespecies o cepas