

Project: Whole genome sequencing and de novo assembly of *Staphylococcus pseudintermedius*: a pangenome approach to unravelling pathogenesis of canine pyoderma

Marcela Corpus Hernández

Escuela Nacional de Estudios Superiores, Unidad León.

Universidad Nacional Autónoma de México.

3 de Diciembre 2021.

Abstract

La mayoría de los pyodermas (llagas) caninos son causados por *Staphylococcus pseudintermedius*, la cual suele encontrarse normalmente en la piel de los perros sin ningún problema. Estos microorganismos se vuelven patógenos cuando ingresan al organismo del perro, pero sus mecanismos no han sido bien estudiados.

Su entendimiento es necesario para desarrollar tratamientos preventivos y remedios. En el artículo se aislaron muestras genómicas de los organismos desde los perros, tanto no infectados como sí, para comprender cómo funciona su patogenicidad.

INTRODUCCIÓN

Se recogieron 22 muestras de *S. pseudintermedius* desde 5 perros saludables y 33 muestras de 33 perros infectados. El tamaño promedio del genoma fue de 2.62 Mbp con gran cantidad de genes asociados a resistencia.

Los objetivos fueron:

- a) Determinar si la tecnología de Nanopore permite secuenciar y ensamblar de novo el genoma completo de *S. pseudintermedius*.

- b) Analizar las diferencias entre los genomas de los *S. pseudintermedius* de perros saludables e infectados.

En este trabajo se llevará a cabo el proyecto con las secuencias e información del artículo original.

MÉTODOS

Se descargó la secuencia de referencia o consenso de *Staphylococcus pseudintermedius* (identificador CP066712.1) y se guardó con el nombre Seq2SPse.fasta.fasta.

Se descargó un archivo SRA con identificador SRR17023946 con el comando **fastq-dump --split-3 SRR17023946** para archivos pareados y se generaron los archivos SRR17023946_1.fastq y SRR17023946_2.fastq.

Se les realizó un análisis de calidad con Fastqc a ambos archivos: **fastqc SRR17023946_*.fastq** y se revisaron los archivos html.

En ambos archivos se encontró buena calidad en general, exceptuando el contenido de secuencia base.

Se realizó un filtrado de secuencias con el comando **fastp -i SRR17023946_1.fastq -I SRR17023946_2.fastq -o SRR1filt.fastq -O SRR2filt.fastq >& Resultfastp.log** y se obtuvo el siguiente resultado con el nombre de **Resultfastp.log**:

Read1 before filtering:

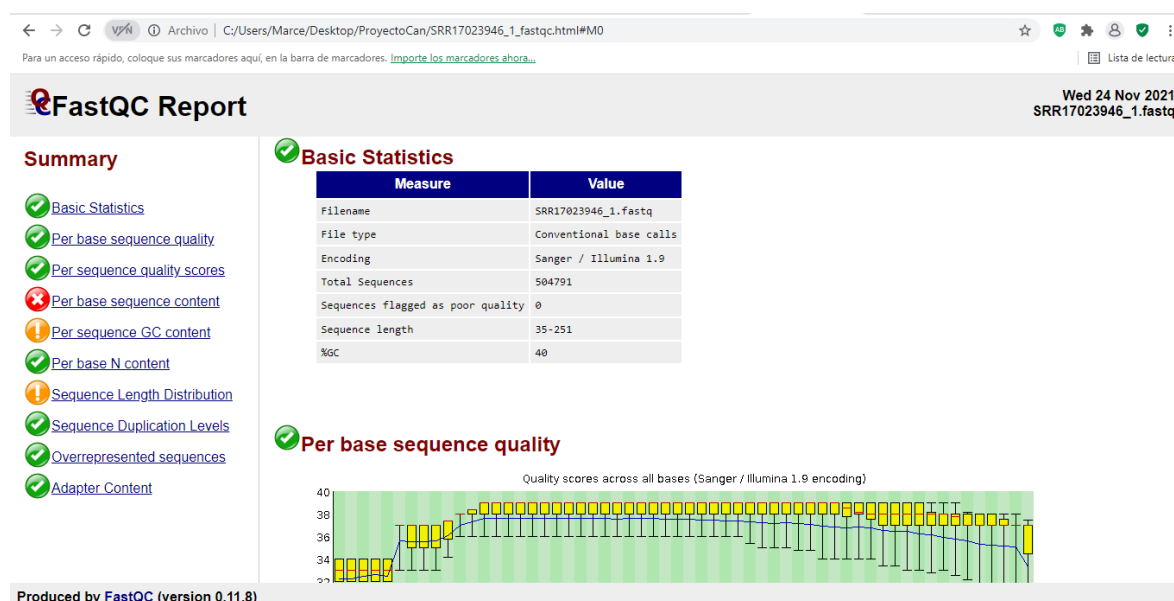


Figura 1.
Resultados
FastQC
Report
SRR17023
946_1

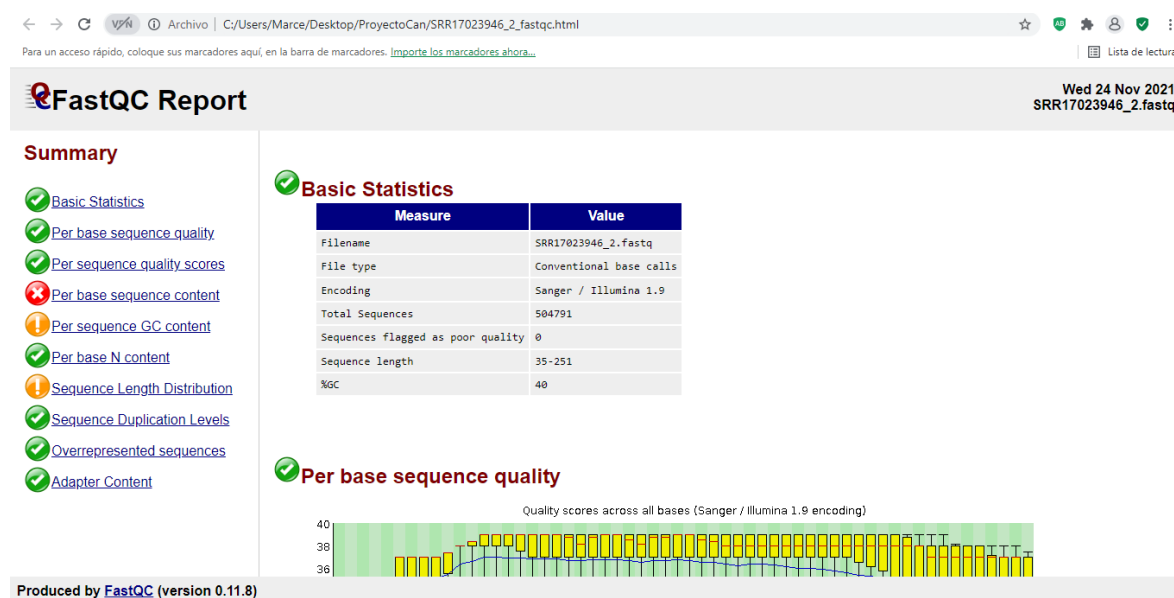


Figura 2.
Resultados
FastQC
Report
SRR17023
946_2

total reads: 504791
total bases: 111902909
Q20 bases: 108323843(96.8016%)
Q30 bases: 106820465(95.4582%)

Read2 before filtering:

total reads: 504791
total bases: 111956976
Q20 bases: 104552901(93.3867%)
Q30 bases: 101836011(90.96%)

Read1 after filtering:

total reads: 503468
total bases: 111523319
Q20 bases: 108040113(96.8767%)
Q30 bases: 106570228(95.5587%)

Read2 after filtering:

total reads: 503468
total bases: 111489368
Q20 bases: 104355176(93.601%)
Q30 bases: 101714808(91.2327%)

Filtering result:

reads passed filter: 1006936
reads failed due to low quality: 2646
reads failed due to too many N: 0
reads failed due to too short: 0

reads with adapter trimmed: 6042
bases trimmed due to adapters:
365918

Duplication rate: 0.590375%

Insert size peak (evaluated by paired-end reads): 250

JSON report: fastp.json

HTML report: fastp.html

fastp -i SRR17023946_1.fastq -I
SRR17023946_2.fastq

fastp v0.22.0, time used: 9 seconds

Los filtrados se guardaron con los
nombres **SRR1filt.fastq** y
SRR2filt.fastq.

Las calidades mejoraron un poco,
elevando los puntajes de los archivos.

Después se armó el índice para el
genoma de referencia con bowtie3
usando el comando **conda activate
bowtie2, bowtie2-build -f
Seq2SPse.fasta.fasta Spseu** y se
generaron los archivos **Spseu.1.bt2,
Spseu.2.bt2, Spseu.3.bt2,
Spseu.4.bt2, Spseu.rev.1.bt2 y
Spseu.rev.2.bt2**. Para mapear los
reads a la referencia se utilizó el
comando **bowtie2 --maxins 1000 -x
Spseu -1 SRR1filt.fastq -2
SRR2filt.fastq -S Spseu.sam** y se
generó el archivo **Spseu.sam** para
después filtrarlo con el comando **awk
'\$3!=""' Spseu.sam >300filt.sam**.

Posteriormente se realizó un blast descargando la siguiente secuencia JF275103.1. Se corrió el comando **makeblastdb -in SpseuJF.fasta -dbtype nucl -out SpseuJFDB**, generando los archivos **SpseuJFDB.ndb**, **SpseuJFDB.nhr**, **SpseuJFDB.nin**, **SpseuJFDB.not**, **SpseuJFDB.nsq**, **SpseuJFDB.ntf** y **SpseuJFDB.nton**.

Luego se corrió el comando **blastn -db SpseuJFDB -query Seq2SPse.fasta.fasta -evaluate 1e-5 -max_target_seqs 4 -outfmt "7 qseqid length eval evalue qcovs qcovshsp" -num_threads 4 -out Spseublast.tab** para generar el archivo **Spseublast.tab**.

```
mcorpus@gaia:~/Proyecto
# BLASTN 2.10.0+
# Query: CP066712.1 Staphylococcus pseudintermedius strain DSP027 chromosome, complete genome
# Database: SpseuJFDB
# Fields: query id, alignment length, evalue, % query coverage per subject
# 1 hits found
CP066712.1      1530      0.0      0
# BLAST processed 1 queries
Spseublast.tab (END)
```

La secuencia JF275103.1 se corrió en Augustus para predicción de genes con los datos *Staphylococcus aureus*, both strands y few alternative script. Se predijeron las siguientes secuencias de código:

```
>JF275103.1:g1.t1
atgatgagtaacgcagctaatttactatTTTTT
gttgcccaaacgcgcgaatcatacctaataaa
atcggtacaacag
cgccagcacgtgctgtagcggaacggtacgaaaa
acgctaaaataatagatacaagaatcgccccga
taacgatatTTTTTcgTTTTtattaccgacaaac
gataataactaatagtgcacaaacgTTTTatgtaag
TTTgtcacttgcatcgcggttgctaagaacaat
gctgcagcaactaacgcgcacggcactcgTTga
aaaaccactaaaggctaacttcaatgcatttcc
tgtaccgaataagtcgtcacctTTtaacaccgc
accgccactTTga
>JF275103.1:g2.t1
```

```
atgttaaacagtgaaattaatgcaatcgTTgct
acaatcggcatacctTTTcaaaccattaaccg
aatgTTTggTcac
TTaaccattgtgctgctgctgTTTTtaataaga
cgTTacctaacgaaatgccgacaccgaatacaa
tgatagtgcCCcatggaatacggcTTTcgact
tTTTTccaattcatgacaccaattTTTggcgTT
aacataatcgccaatgcaattaatgtgatagat
gaagaatcgattgggtgtaacactTTTTcagt
tgaccaagacacgagtaataataatgagatgat
aattaagcgccattctgTTggTTtaactggacc
aagTTctgccagTTgTTgTTtaacgagctccg
taccacTTcaatggcatctgTTcaggtggaa
tgactTTtaacataacaaaaataa
>JF275103.1:g3.t1
Gtggatgttaatgaaataaagactggaattaat
gcagatgctaaacttgTgcacttgcaaatcct
aatga
```

Y las siguientes de proteínas:

```
>JF275103.1:g1.t1
MMSNAANLLFFVAPNAIIIPNKIGTTAPARAVA
DGTKNAKIIDTRIAPITIFFVLLPTNDNTNSAK
RLCKFVTCIAVA
KNNAATNATALVEKPLKANFNAPFPVNKSSPF
```



```
NTAPPL
>JF275103.1:g2.t1
MLNSEINAIVATIGIPFKPIKPNVWSLNHCAAV
VFNKTLPNEMPTPNTMIVPHGIRLSTSFQFMTF
ILGVNIIANAIN
VIDEESIGCNTFSVDQDTSNNNEMIIKRHSVGL
TGPSSASCCLTSSVPPSMASCSGGMTFNITK
>JF275103.1:g3.t1
MDVNEIKTGINADAKLVALANPK
```

Las proteínas se guardaron en archivos fasta por separado para poder enviarlas a Interproscan.

Resultados prot 1:



Figura 3. Visualización archivo **Spseublast.tab**

Figura 4.
Resultados
Interpro,
proteína 1.

InterProScan Search Result	
Title	JF275103.1:q1.t1
Job ID	iprscan5-R20211126-022104-0083-46782142-p2m
Length	117 amino acids
Action	 
Status	✓ finished
Expires	Thu Dec 02 2021
Protein family membership	
None predicted	



Resultados prot2:

Figura 5.
Resultados
Interpro,
proteína 2.

InterProScan Search Result	
Title	JF275103.1:q2.t1
Job ID	iprscan5-R20211126-022119-0720-53497526-p2m
Length	142 amino acids
Action	 
Status	✓ finished
Expires	Thu Dec 02 2021
Protein family membership	
None predicted	

Resultados prot3:

Figura 6.
Resultados
Interpro,
proteína 2.

InterProScan Search Result	
Title	JF275103.1:q3.t1
Job ID	iprscan5-R20211126-022216-0682-31587863-p1m
Length	23 amino acids
Action	 
Status	✓ finished
Expires	Thu Dec 02 2021
Protein family membership	
None predicted	

En ninguno de los casos se encontraron familias proteicas.

RESULTADOS

Finalmente, las secuencias estudiadas para anotación no pudieron predecir una familia o dominio de proteínas compatible, posiblemente porque la secuencia no era un gen.

El resto de la metodología solamente se realizó con un archivo SRA, con lo que se obtuvo buena calidad y fue posible encontrar similitudes por medio de blast.

DISCUSIÓN Y CONCLUSIONES

Solamente se trabajó con un archivo SRA en formato fastq, por lo que los pasos no requirieron de gran consumo de tiempo. Se logró analizar las calidades y realizar las filtraciones para lograr mapear al genoma original publicado, nuestra referencia. Dado que el mapeo se logró, el archivo sí contenía un fragmento del genoma de *Staphylococcus pseudintermedius*.

Para lograr hacer la anotación se utilizó otro archivo en formato fasta, aunque también referente a nuestro organismo. Se lograron hacer las predicciones, pero al buscar los dominios de las posibles proteínas no

se encontraron resultados, lo que podría sugerir que el archivo no representaba una sección codificante.

REFERENCIAS

1. Artículo:
<https://onlinelibrary.wiley.com/doi/10.1111/vde.13040>
2. Información de artículo en NCBI:
<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA685966>
3. Información de Staphylococcus pseudintermedius:
<https://www.ncbi.nlm.nih.gov/genome/3429>
4. Secuencia de referencia de Staphylococcus pseudintermedius:
<https://www.ncbi.nlm.nih.gov/nuccore/CP066712.1?report=fasta>
5. Links de secuencias del proyecto:
https://www.ncbi.nlm.nih.gov/biosample?LinkName=bioproject_biosample_all&from_uid=685966
6. Archivo SRA descargado:
[https://www.ncbi.nlm.nih.gov/sra/SRX13214064\[accn\]](https://www.ncbi.nlm.nih.gov/sra/SRX13214064[accn])
7. Archivo de JF275103.1.
<https://www.ncbi.nlm.nih.gov/nuccore/JF275103.1?report=fasta>
8. Resultados de Augustus
<http://bioinf.uni-greifswald.de/augustus/cabinet?folder=AUG-476778028>
9. Resultados de Interproscan:
<http://www.ebi.ac.uk/interpro/result/InterProScan/#table>
10. Proteína 1 Interproscan resultados:
<http://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20211126-022104-0083-46782142-p2m/>
11. Proteína 2 Interproscan resultados:
<http://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20211126-022119-0720-53497526-p2m/>
12. Proteína 3 Interproscan resultados:
<http://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20211126-022216-0682-31587863-p1m/>