



**Aplicación de secuenciación masiva paralela para el descubrimiento de virus  
fitopatógenos en agua de riego del ramal “El Pueblo” No. 12.**

Guevara Guillén, Fiama Estefany

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría de Investigación en Biotecnología Vegetal

Trabajo de titulación, previo a la obtención del título de Magíster en Biotecnología  
Vegetal

Flores Flor, Francisco Javier, Ph.D

10 de mayo de 2021



### Document information

Analyzed document	TESIS_GUEVARAF_vp2.docx (D110675510)
Submitted	7/22/2021 9:31:00 PM
Submitted by	
Submitter email	fjflores2@espe.edu.ec
Similarity	7%
Analysis address	fjflores2.espe@analysis.orkund.com

### Sources included in the report

<b>SA</b>	<b>Tesis Fiama plain text.txt</b> Document Tesis Fiama plain text.txt (D29631919)	22
<b>W</b>	URL: <a href="https://www.apsnet.org/edcenter/disandpath/viral/introduction/Pages/PlantVirusesEspanol.asp">https://www.apsnet.org/edcenter/disandpath/viral/introduction/Pages/PlantVirusesEspanol.asp</a> x Feiched: 7/22/2021 9:32:00 PM	1
<b>SA</b>	<b>Universidad de las Fuerzas Armadas ESPE / Eduardo Moncayo_Tesis Final.docx</b> Document Eduardo Moncayo_Tesis Final.docx (D'3529759) Submitted by: fjflores2@espe.edu.ec Receiver: fjflores2.espe@analysis.orkund.com	1
<b>SA</b>	<b>28bbe78ba879806d5e7301d44f10a4abe0f7df39.docx</b> Document 28bbe78ba879806d5e7301d44f10a4abe0f7df39.docx (D78237706)	1
<b>W</b>	URL: <a href="https://repositorio.espe.edu.ec/bitstream/21000/13366/1/T_ESPE_057321.pdf">https://repositorio.espe.edu.ec/bitstream/21000/13366/1/T_ESPE_057321.pdf</a> Feiched: 12/4/2019 12:56:11 AM	5
<b>SA</b>	<b>PROYECTO DE INVESTIGACION JHON FINAL 192018 URKUND.doc</b> Documen: PROYECTO DE INVESTIGACION JHON FINAL 192018 URKUND.doc (D41037250)	1



Escaneo el código QR para:  
FRANCISCO  
JAVIER FLORES  
FLOR

.....

Flores Flor, Francisco Javier, Ph.D

Director

C.C.: 1713443479



VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y  
TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Aplicación de secuenciación masiva paralela para el descubrimiento de virus fitopatógenos en agua de riego del ramal "El Pueblo" No. 12.**" fue realizado por la señorita **Guevara Guillén, Fiana Estefany** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 10 de mayo de 2021

Firma:



.....  
Flores Flor, Francisco Javier, Ph.D

Director

C.C.: 1713443479



**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y  
TRANSFERENCIA DE TECNOLOGÍA  
CENTRO DE POSGRADOS**

**RESPONSABILIDAD DE AUTORÍA**

Yo Guevara Guillén, Fiana Estefany, con cédula de ciudadanía n° 171959789-8, declaro que el contenido, ideas y criterios del trabajo de titulación: **“Aplicación de secuenciación masiva paralela para el descubrimiento de virus fitopatógenos en agua de riego del ramal “El Pueblo” No. 12.”** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 10 de mayo de 2021

Firma

.....  
**Guevara Guillén, Fiana Estefany**

C.C.: 171959789-8



**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y  
TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

**AUTORIZACIÓN DE PUBLICACIÓN**

Yo, **Guevara Guillén, Fiama Estefany**, con cédula de ciudadanía n°171959789-8, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **“Aplicación de secuenciación masiva paralela para el descubrimiento de virus fitopatógenos en agua de riego del ramal “El Pueblo” No. 12.”** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolqui, 10 de mayo de 2021

**Firma**

.....  
**Guevara Guillén, Fiama Estefany**

**C.C.: 171959789-8**

## **DEDICATORIA**

Mi camino hasta este momento no habría sido igual sin el apoyo y motivación de mi familia. Sus consejos y cariño siempre me han ayudado a superar cada obstáculo, me han enseñado a creer en mí y a luchar por mis metas. A ellos va dedicado este trabajo.

## AGRADECIMIENTO

Agradezco a mi familia por ser ese ejemplo de perseverancia y disciplina, quienes siempre han apoyado mi carrera y me han motivado a ser mejor cada día. Gracias por su amor y comprensión incondicional.

Seguir el camino del mundo de la investigación no es fácil pero la inspiración, apoyo y el referente de grandes mentores ha sido pieza clave en mi carrera. Agradezco principalmente al Dr. Francisco Flores y a la MSc. Alma Koch por abrirme las puertas del Laboratorio de Microbiología Molecular desde mis inicios como estudiante de pregrado. Les agradezco confiar en mi trabajo, por brindarme sabios consejos no solo en el ámbito, por todas las experiencias y enseñanzas a lo largo de estos años, gracias por toda la dedicación que dejan en su trabajo. También agradezco al Grupo de Investigación en Microbiología y Ambiente (GIMA) de la carrera de Ingeniería en Biotecnología de la Universidad de las Fuerzas Armadas-ESPE del que he sido parte y que ha colaborado con equipos y reactivos usados en mi investigación.

También quiero agradecer al MSc. William Viera y a la Ing. Andrea Sotomayor del Programa de Fruticultura del Instituto de Investigaciones Agropecuarias (INIAP)-Granja Tumbaco y a su Proyecto “Biocontrol for Sustainable Farming Systems, Ecuador” del Ministerio de Asuntos Exteriores y Comercio (MFAT) de Nueva Zelanda, por la colaboración en cuanto a las muestras usadas. Finalmente, agradezco a la Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia, CEDIA, por el financiamiento para el desarrollo del proyecto “CEPRA-XI-2017-11: ANÁLISIS NGS DE PATÓGENOS” dentro del cual se desarrolló mi investigación y que también facilitó el uso de su clúster para computación de alto rendimiento durante el procesamiento de datos.

## ÍNDICE DE CONTENIDOS

Índice de tablas.....	10
Índice de figuras .....	11
Resumen .....	12
Abstract .....	13
Capítulo 1: Introducción .....	14
Tema .....	14
Formulación del Problema .....	14
Objetivos de la Investigación .....	15
Objetivo general .....	15
Objetivos Específicos .....	15
Marco Teórico .....	16
Generalidades de los virus .....	16
Virus en fuentes de agua .....	16
Métodos para concentrar virus.....	17
Secuenciación de alto rendimiento .....	19
Tecnologías de secuenciación masiva .....	20
Aplicaciones de Secuenciación masiva (HTS) .....	21
Metagenómica de virus (viroma) .....	22
Ensamblaje de genomas.....	23
Hipótesis.....	24
Capítulo 2: Materiales Y Métodos .....	25
Zona de estudio.....	25
Duración de la investigación .....	25
Procedimientos.....	25

Toma de muestras .....	25
Concentración viral .....	26
Extracción de ARN.....	26
RT-PCR .....	26
Secuenciación masiva.....	27
Análisis bioinformático.....	28
Comparación de herramientas.....	29
Validación del pipeline.....	31
Identificación de <i>Hibiscus latent Fort Pierce virus</i> (HLFPV).....	31
Capítulo 3: Resultados.....	32
Lugar de muestreo .....	32
Concentración viral.....	33
Amplificación de muestras mediante PCR convencional .....	34
Secuenciación masiva y análisis bioinformático.....	34
Comparación de pre-procesamiento y ensambladores .....	35
Ensamblaje y anotación de Genomas .....	44
Identificación molecular de HLFPV .....	46
Capítulo 4: Discusión .....	48
Capítulo 5: Conclusiones .....	51
Capítulo 6: Recomendaciones .....	52
Capítulo 7: Bibliografía.....	53

## ÍNDICE DE TABLAS

<b>Tabla 1.</b> <i>Combinaciones de herramientas bioinformáticas</i> .....	30
<b>Tabla 2.</b> <i>Número de lecturas obtenidas</i> .....	36
<b>Tabla 3.</b> <i>Número de contigs totales</i> .....	37
<b>Tabla 4.</b> <i>Parámetro N50</i> .....	37
<b>Tabla 5.</b> <i>Parámetro L50</i> .....	38

## ÍNDICE DE FIGURAS

<b>Figura 1.</b> <i>Flujo de trabajo del análisis bioinformático</i> .....	29
<b>Figura 2.</b> <i>Zona de toma de muestras</i> .....	32
<b>Figura 3.</b> <i>Hojas de cucarda con síntomas de virosis</i> .....	33
<b>Figura 4.</b> <i>Concentración viral</i> .....	33
<b>Figura 5.</b> <i>Electroforesis PCR Tombusvirus</i> .....	34
<b>Figura 6.</b> <i>Calidad de secuencias crudas</i> .....	35
<b>Figura 7.</b> <i>Diagrama BoxPlot del ensamblaje</i> .....	39
<b>Figura 8.</b> <i>Visualización de la búsqueda de homólogos</i> .....	42
<b>Figura 9.</b> <i>Número de contigs virales ensamblados</i> .....	43
<b>Figura 10.</b> <i>Porcentaje del genoma alineado</i> .....	44
<b>Figura 11.</b> <i>Estructura del genoma de HLFPV</i> .....	45
<b>Figura 12.</b> <i>Estructura del genoma de RGMoV</i> .....	46
<b>Figura 13.</b> <i>Electroforesis PCR HLFPV</i> .....	47

## RESUMEN

En el Ecuador son pocos los reportes de virus en fuentes de agua y los que existen describen virus humanos mientras que los virus fitopatógenos, que también se transmiten a través del agua, han sido poco estudiados. Dado que no existen métodos de control definitivos para las enfermedades virales en plantas, mejorar la bioseguridad agrícola y la detección temprana de virus es la mejor manera de evitar pérdidas económicas. Diferentes técnicas y métodos han sido empleados para el diagnóstico viral sin embargo, la secuenciación masiva paralela es una poderosa alternativa para el diagnóstico de virus en muestras ambientales. El objetivo del proyecto fue aplicar una tecnología de secuenciación masiva paralela para el descubrimiento de virus fitopatógenos en agua de riego en un reservorio de la Granja Experimental INIAP en el sector de Tumbaco, Pichincha. Utilizando el método de floculación orgánica con leche descremada se obtuvo un concentrado viral a partir de muestras de agua. Se extrajo ARN de los concentrados virales y se realizó secuenciación masiva paralela utilizando la tecnología Illumina. Tras el análisis bioinformático de datos se determinó la presencia de 4 familias de virus fitopatógenos: *Tombusviridae*, *Virgaviridae*, *Solemoviridae* y *Alphaflexiviridae* de un total de 19 familias virales dentro de los dominios Duplodnaviria y Riboviria. Además se realizó una comparación de herramientas bioinformáticas para el ensamblaje de lecturas de secuenciación masiva donde el ensamblador metaSPAdes mostró los mejores resultados. Con dicha herramienta se obtuvo más del 99% de los genomas ensamblados de *Hibiscus latent Fort Pierce virus* (HLFPV) y *Ryegrass mottle virus* (RGMoV).

Palabras clave:

- **FITOPATÓGENOS**
- **HLFPV**
- **RGMoV**
- **ENSAMBLAJE**

## ABSTRACT

In Ecuador there are few reports of viruses in water sources and those that already exist describe human viruses while phytopathogenic viruses, which are also transmitted through water, have been less studied. Since there are no definitive control methods for viral diseases in plants, improving agricultural biosecurity and early virus detection is the best way to avoid economic losses. Different techniques and methods have been used for viral diagnosis, however, parallel massive sequencing is a powerful alternative for the diagnosis of viruses in environmental samples. The objective of the project was to discover phytopathogenic water-borne viruses of in an irrigation water reservoir of the INIAP Experimental Farm in Tumbaco, Pichincha province using high throughput sequencing (HTS). Using the organic flocculation method with skim milk, a viral concentrate was obtained from water samples. RNA was extracted from the viral concentrates and HTS sequencing was performed with the Illumina technology. The sequencing and computational analysis of data showed the presence of 4 families of phytopathogenic viruses: *Tombusviridae*, *Virgaviridae*, *Solemoviridae* and *Alphaflexiviridae* out of a total of 19 viral families within the Duplodnaviria and Riboviria domains. In addition, a comparison of bioinformatics tools for the assembly of massive sequencing reads was performed, as result the assembler metaSPAdes showed the best evaluated parameters. With this tool, the near complete genomes of *Hibiscus latent Fort Pierce virus* (HLFPV) and *Ryegrass mottle virus* (RGMoV) were assembled.

Key words:

- **PHYTOPATHOGENS**
- **HLFPV**
- **RGMoV**
- **ASSEMBLY**

## CAPÍTULO 1: INTRODUCCIÓN

### Tema

Aplicación de secuenciación masiva paralela para el descubrimiento de virus fitopatógenos en agua de riego del ramal “El Pueblo” No. 12.

### Formulación del Problema

Además de los ya conocidos mecanismos de transmisión de virus desde un hospedero hacia otro (vectores, injertos, semillas, polen, transmisión mecánica y suelo), el agua constituye una matriz estudiada por su potencial de transmisión de virus (Mehle & Ravnkar, 2012). La investigación de virus patógenos humanos que se transmiten a través del agua es un tema de importancia para las urbes alrededor del mundo ya que fuentes de agua para consumo contaminadas pueden originar importantes brotes de enfermedades virales (Hellmér et al., 2014; Nieuwenhuijse y Koopmans, 2017). En el área de la agricultura y la horticultura, donde se emplean técnicas de irrigación, cultivos hidropónicos y muchas veces se reutiliza el agua, es de interés evitar el uso de fuentes de agua contaminadas con patógenos. Se han detectado partículas de varios géneros de virus fitopatógenos en diferentes fuentes de agua como ríos, lagos, agua de riego y agua de mar (Bütter y Koenig, 2017); y se conoce que muchos de ellos tienen la capacidad de persistir en el agua y mantenerse infectivos por largos períodos de tiempo, lo que constituye un peligro potencial. No existen métodos de control definitivos para las enfermedades virales, por lo que fuentes de agua contaminadas pueden ocasionar pérdidas en grandes extensiones de cultivos (Mehle & Ravnkar, 2012; Mehle et al., 2014; Jeżewska et al., 2019). Mejorar la bioseguridad agrícola y la detección temprana de virus es la mejor manera de evitar pérdidas económicas.

Los métodos más usados para la detección e identificación de virus implican técnicas serológicas y técnicas de biología molecular con secuenciación Sanger. Esta última ha permitido establecer relaciones entre grupos, mejorando el conocimiento de la diversidad de virus existentes, la especificidad de hospederos y el descubrimiento de virus en nuevos hospederos y nuevas localidades. Sin embargo, las tecnologías de secuenciación de siguiente generación (NGS, por sus siglas en inglés) han tenido un fuerte impacto en la investigación de virus. Desde hace cinco años áreas como la epidemiología, genética de poblaciones, interacción hospedero-patógeno y la etiología han logrado avances gracias a la secuenciación masiva y análisis bioinformático. Este enfoque de secuenciación permite obtener mayor cantidad de información en menor

tiempo y a costos más bajos, identificar virus conocidos y el descubrimiento de nuevos genotipos o incluso virus completamente nuevos, por lo que su aplicación como método de diagnóstico para vigilancia de enfermedades virales es de gran interés (Adriaenssens et al., 2018; Hjelmsø et al., 2019).

El Ecuador es un país con gran potencial en el campo de la agricultura. Favorecido por su ubicación geográfica y diversidad de microclimas tiene la capacidad de producir cultivos de importancia comercial, esto sumado al interés en seguridad alimentaria constituyen un aspecto de importancia económica nacional. En este sentido, el conocimiento de los agentes fitopatógenos que se están transmitiendo a través del agua de riego permitirá elaborar planes de control para reducir las pérdidas registradas en la productividad ocasionadas por virus, mejorar la calidad de los cultivos y promover el uso de herramientas y recursos bioinformáticos como métodos rápidos y efectivos de diagnóstico en programas de cuarentena y de certificación de material vegetal. En trabajo previo (Guevara, et al., 2017), se buscó detectar e identificar virus fitopatógenos en agua de riego utilizando primers genéricos y secuenciación Sanger obteniendo parcialmente un virus, por lo que con esta investigación se pretende profundizar en la detección e identificación de virus a nivel de especie en la fuente de agua de riego analizada.

## **Objetivos de la Investigación**

### ***Objetivo general***

Aplicar la tecnología de secuenciación masiva paralela Illumina para el descubrimiento de virus fitopatógenos en agua de riego del ramal “El Pueblo” No. 12.

### ***Objetivos Específicos***

- Detectar e identificar virus fitopatógenos en agua de riego del ramal “El Pueblo” No. 12 mediante secuenciación masiva.
- Comparar herramientas bioinformáticas para el ensamblaje de genomas de virus.
- Describir y validar un flujo de trabajo para el análisis de datos de secuenciación masiva.

## **Marco Teórico**

### ***Generalidades de los virus***

Los virus no poseen todos los genes necesarios para replicarse, por lo que dependen de la maquinaria biológica de las células de un hospedero vivo, siendo considerados 'parásitos intracelulares'. Son capaces de infectar organismos de los tres dominios: eucariotas, bacterias y arqueas (Hull, 2009; Gergerich & Dolja, 2006).

Las partículas virales (viriones) están conformadas por dos partes principales: su genoma (ácido nucléico) y una cubierta proteica protectora conocida como cápside. Muchos virus pueden encontrarse también envueltos por otra membrana lipoproteica llamada envoltura (Gergerich & Dolja, 2006). Los virus se clasifican tomando en cuenta la forma química de sus genomas, por lo que David Baltimore en 1971, propuso agrupar los virus en siete clases (Cann, 2005; Kumar et al., 2016). La mayoría de virus que infectan plantas presentan un genoma de ARN de cadena simple positiva (Agris, 2005), que puede ser traducido directamente a proteína en los ribosomas.

De las especies de virus que se conocen actualmente, más de 1000 son causantes de importantes enfermedades en las plantas y son responsables de fuertes pérdidas en la producción y calidad de los cultivos. Algunos tienen hospederos específicos, mientras que otros afectan a una amplia gama de especies vegetales alrededor del mundo. Los síntomas representativos de las infecciones virales pueden presentarse en diferentes partes de la planta (desde hojas, hasta frutos), principalmente en forma de manchas o mosaicos, deformaciones, retraso en el crecimiento y, en casos más extremos, la muerte de la planta (Kumar et al., 2016).

El diagnóstico de los virus que afectan los cultivos es una de las principales herramientas para el manejo oportuno de enfermedades. Las prácticas de control incluyen la identificación del agente causal, control de vectores, aplicaciones fitosanitarias y el uso de cultivares resistentes a virus específicos (Chang, 2009).

### ***Virus en fuentes de agua***

Los virus fitopatógenos comúnmente se detectan mediante técnicas serológicas como ELISA ('Enzyme-linked Immunosorbent Assay') y métodos basados en la amplificación de ácidos nucléicos como la Reacción en Cadena de la Polimerasa (PCR, por sus siglas en inglés) (Jooste et al., 2017). Actualmente, la aplicación de

secuenciación masiva paralela es una poderosa alternativa para el diagnóstico de virus, especialmente para elucidar la composición viral de una muestra ambiental.

Se ha observado la presencia de partículas virales en muestras ambientales de suelo, aire y agua (Guerrero-Latorre et al., 2018), por lo que se dice que los virus son los organismos de mayor distribución en el planeta. Varios estudios han demostrado la presencia de virus patógenos humanos, así como virus fitopatógenos en aguas residuales, aguas tratadas, agua de drenaje e incluso agua de ríos y lagos (Mehle & Ravnikar, 2012; Rastrojo & Alcamí, 2017; Bütter y Koenig, 2017). Recientemente, la presencia de virus patógenos humanos en agua ha sido estudiada mediante concentración viral y técnicas de secuenciación de siguiente generación; y se ha observado que la gran diversidad de especies detectadas a través de metagenómica brinda información clave sobre la presencia de importantes agentes infecciosos en fuentes de agua (Guerrero-Latorre et al., 2018).

Los virus fitopatógenos se transmiten planta-planta principalmente a través de vectores, polen, semillas, y mecánicamente al estar en contacto con plantas previamente infectadas (Gergerich & Dolja, 2006). El agua, aunque poco estudiada, constituye otro medio de transmisión de virus fitopatógenos. Se han detectado al menos nueve diferentes géneros de virus fitopatógenos en medios acuáticos: *Cucumovirus*, *Carmovirus*, *Nepovirus*, *Tospovirus*, *Necrovirus*, *Diathovirus*, *Tobamovirus*, *Potexvirus* y *Tombusvirus*, de los cuales, los tres últimos han sido los más reportados. Además, se ha reportado que estos nueve géneros de virus pueden ser transmitidos a través de agua y suelo hacia nuevos hospederos sanos, sin necesidad de vectores. Las partículas virales son liberadas de las raíces de plantas enfermas y reabsorbidas en las partículas del suelo para diseminarse a través del agua e ingresar por pequeñas heridas en las raíces de plantas sanas (Bütter y Koenig, 2017). Ejemplo de ello es el caso de cultivos que se mantienen en invernadero, donde el agua circula por toda el área; o en el caso de cultivos hidropónicos (Mehle & Ravnikar, 2012; Kumar et al., 2016).

### ***Métodos para concentrar virus***

Es difícil detectar e identificar virus fuera de sus hospederos o vectores, en matrices como el agua y el suelo, debido a la baja concentración de partículas virales en estos medios (Haramoto et al., 2018). Por lo que es necesario aplicar protocolos de purificación y/o concentración para la remoción de otros microorganismos que no son

de interés o partículas que pueden interferir con los métodos subsecuentes de extracción de ácidos nucleicos (Hjelmsø et al., 2017). Existen varios métodos para la concentración de virus a partir de muestras de agua.

**Adsorción- elución.** Esta categoría agrupa varios métodos que se basan en la adsorción de las partículas virales a matrices sólidas como membranas de filtros, lana de vidrio, tierra de diatomeas, entre otros; seguido de una elución con un buffer hasta un volumen menor denominado concentrado. Estos métodos muchas veces suelen ser inconsistentes debido a las diferencias de pH de las partículas virales y del agua, lo que lleva a una adsorción selectiva. Además, los buffers usados pueden influenciar en los procesos posteriores como extracción de ARN o la PCR (Cashdollar & Dahling, 2006; Millen et al., 2012).

**Coagulación / Floculación.** Este método descrito por Calgua y colaboradores (2008) consiste en la floculación de una matriz orgánica proteica (leche descremada o skimmed milk) por reducción del pH, a la cual se adhieren las partículas virales mediante interacciones electrostáticas. Seguido de la sedimentación de los flóculos por acción de la gravedad. Las partículas virales se sedimentan junto con la matriz y posteriormente se eluyen con un buffer alcalino que permite separar dichas partículas virales y procesarlas. La aplicabilidad de este método se ha demostrado en agua de mar y agua de río usando volúmenes de agua menores a los 10 L (Calgua et al., 2013; Haramoto et al., 2018).

**Filtración y Ultra filtración.** Para concentrar virus a partir de muestras de agua se puede utilizar la filtración común con los llamados filtros de un solo uso, generalmente con tamaños de poro de 0,2 a 0,45  $\mu\text{m}$  para retener virus y con un paso posterior de elución. Sin embargo, el procesamiento de grandes cantidades de muestra en un tiempo razonable provoca la obstrucción del filtro (Cai et al., 2015).

Por otra parte, la ultrafiltración (UF) es una variedad de filtración a través de una membrana semipermeable definida según un límite de peso molecular (MWCO). Los sólidos suspendidos y los solutos de alto peso molecular se retienen en la membrana (retenido), mientras que el agua y los solutos de bajo peso molecular pasan a través de la membrana en el filtrado (Liu et al., 2012; Kahler et al., 2015). Este método además ha sido probado para concentrar múltiples clases de microbios (virus, bacterias y protozoos) en muestras de agua usando un solo filtro (Haramoto et al., 2018). Dentro de este método se puede encontrar también una variante denominada Ultra filtración de flujo tangencial (TFF, por sus siglas en inglés). Este método consiste en la filtración a

través de flujos de fluido paralelos tangentes a la superficie del filtro. La recirculación de flujo cruzado permite procesar más agua (10–1000 L) dependiendo del sistema, ya que evita la obstrucción del filtro. El equipo de TFF puede equiparse con una membrana de ultrafiltración que posee una amplia gama de tamaños de poros o cortes moleculares (Cai et al., 2015).

**Concentración con polietilenglicol (PEG).** Este método consiste en una combinación del método de adsorción -elución sumada a la adición del reactivo PEG y posterior proceso de centrifugado de la muestra. Las partículas virales se adhieren al PEG que les confiere peso al momento de la centrifugación, de modo que pueden ser recolectadas del pellet que se forma en el proceso (Hjelmsø et al., 2017).

**Ultracentrifugación.** Este método es muy utilizado para concentración de partículas virales. Consiste en la utilización de altas velocidades de centrifugación que van entre los 100 000 a 150 000 rpm. Sin embargo la aplicación de la ultra centrifugación es bastante limitada ya que provoca disrupción de los microorganismos por la alta velocidad de centrifugación, no permite procesar grandes cantidades de muestra debido a las capacidades de volumen y, el elevado costo del equipo (Short & Short, 2008).

### ***Secuenciación de alto rendimiento***

La Secuenciación masiva o de alto rendimiento (High Throughput Sequencing, HTS) hace referencia a tecnologías modernas que tienen la capacidad de generar cientos de millones de lecturas de secuencias por muestra durante la ejecución de una corrida de secuenciación (Carriço et al., 2018).

Se pueden aplicar dos enfoques: amplificación selectiva (targeted) y no selectiva (shotgun); y elegir diferente tamaño de lectura, tasa de error y número de lecturas (Sekse et al., 2017). El enfoque de secuenciación 'targeted' hace referencia a la amplificación de ARN ribosomal codificante de ADN (ADNr), como el 16S ADNr y otros marcadores conservados (Marchesi & Ravel, 2015). Mientras que el enfoque 'shotgun' distingue a la secuenciación de genomas completos.

Las aplicaciones y plataformas existentes para HTS, también denominada secuenciación de siguiente generación (Next Generation Sequencing, NGS) se pueden clasificar en base al tipo de secuenciación y al tamaño de lecturas obtenidas: 1) Secuenciación por ligación (SBL). Este tipo de secuenciación implica la hibridación y ligación de una sonda que está marcada con un fluoróforo y una secuencia ancla al

ADN objetivo. Esta sonda codifica un o dos nucleótidos conocidos y varios nucleótidos universales o degeneradas que se van uniendo complementariamente a un fragmento de ADN mientras que, el ancla codifica una secuencia conocida complementaria al adaptador, originando así el sitio para iniciar la ligadura. Finalmente, el ADN objetivo es visualizado y los nucleótidos conocidos de la sonda son identificadas (Goodwin, McPherson, & McCombie, 2016). 2) Secuenciación por síntesis (SBS). Utiliza una señal como un fluoróforo o un cambio en la concentración iónica para identificar la incorporación de un nucleótido mientras se sintetiza una nueva cadena de ADN (Goodwin, McPherson, & McCombie, 2016).

En la mayoría de los enfoques SBL y SBS, el ADN se amplifica clonalmente en una superficie sólida. Tener miles de copias idénticas de un fragmento de ADN en un área definida garantiza que la señal pueda distinguirse del ruido de fondo. La elección del tamaño de las lecturas en la secuenciación depende de la aplicación del estudio. HTS de lecturas cortas (entre 75 a 400 nt) se utiliza para estudios de metagenómica por su costo accesible mientras que, HTS de lecturas largas (> 400 nt), que resulta ser más costoso, se aplica en el análisis de elementos repetitivos largos, alteraciones en el número de copias y variaciones estructurales de genomas completos que resultan difíciles de resolver con tecnologías de lectura corta. Actualmente, hay dos tipos principales de tecnologías de lectura larga: enfoques de secuenciación en tiempo real de una sola molécula y enfoques sintéticos (Goodwin, McPherson, & McCombie, 2016).

### ***Tecnologías de secuenciación masiva***

Se han desarrollado varias tecnologías (plataformas) de secuenciación masiva con características específicas, sin embargo el esquema de trabajo en todas es bastante similar. Los protocolos incluyen: extracción de ácidos nucleicos, construcción de librerías, secuenciación, generación y análisis de datos (Kukurba & Montgomery, 2015). Actualmente la tecnología de secuenciación de alto rendimiento que prevalece en el mercado es la secuenciación Illumina debido a su alta fidelidad y gran cantidad de datos con lecturas entre 100 y 300 nt. Esta tecnología aplica secuenciación por síntesis con una tasa de error menor al 1% (Goodwin, McPherson, & McCombie, 2016).

La tecnología Roche 454 fue dominante años atrás en cuanto a secuenciación de alto rendimiento, hasta ser eliminada en 2016. Esta tecnología utilizaba

secuenciación por síntesis generando una menor cantidad de datos de lecturas largas ( $\leq 700$  nt) con una tasa de error mayor (Liu et al., 2012).

Lecturas de mayor tamaño, como las obtenidas con la tecnología Pacific Biosciences (1000-10000 nt) permiten un mejor ensamblaje *de novo* (desde el principio) de secuencias genómicas. Recientemente, esta tecnología se ha vuelto popular para armar genomas, ensamblaje de secuencias largas repetitivas, plásmidos y bacteriófagos. Sin embargo la tasa de error de esta tecnología es mayor (~15%) (Rhoads & Au, 2015).

La más reciente de Oxford Nanopore, MinION, presenta un gran potencial competitivo ya que puede proporcionar lecturas largas en tiempo real y el equipo que ofertan es de menor costo que los secuenciadores de otras tecnologías (Goodwin et al., 2015).

En general, se dice que las tecnologías de lecturas cortas proveen una cobertura de secuenciación significativamente mayor que aquellas tecnologías de lecturas largas, a menor costo por base (Sekse et al., 2017).

### ***Aplicaciones de Secuenciación masiva (HTS)***

Cada vez más la disminución de los costos de secuenciación ha provocado una mayor accesibilidad a los mismos, permitiendo a los investigadores desarrollar un amplio catálogo de aplicaciones para la secuenciación masiva paralela. La secuenciación de siguiente generación es utilizada para determinar *de novo* las secuencias del genoma de un organismo. En este contexto se han obtenido los genomas completos de varias especies a la vez que se ha podido re-secuenciar genomas como el Humano, comparando los datos obtenidos con genomas de referencia, pudiendo identificar así nuevas variantes. Desde entonces, los diferentes proyectos de secuenciación del genoma han revelado que los individuos suelen albergar 3.5-4 millones de variantes de nucleótidos individuales (SNV) en total y varios cientos de miles de inserciones-deleciones de nucleótidos cortas en relación con el genoma de referencia (Reuter, Spacek, & Snyder, 2015).

Otra de las áreas de mayor impacto de la secuenciación masiva es el mapeo de elementos reguladores del ADN de todo el genoma. La primera de estas tecnologías fue CHIP-Seq en la que el ADN asociado con un factor de transcripción (TF) o con otro

elemento modificador de la cromatina se inmunoselecciona para luego secuenciar usando HTS. La aplicación a gran escala de estos métodos ha proporcionado una gran cantidad de información invaluable sobre las redes de unión de factores de transcripción, mapas epigenéticos y anotaciones de transcripción (Reuter, Spacek, & Snyder, 2015).

El estudio de los diversos roles celulares del ARN ha mejorado enormemente con el advenimiento de la secuenciación de alto rendimiento. El RNA-seq (secuenciación de ARN) se ha utilizado para caracterizar profundamente los transcriptomas, proporcionando mediciones precisas y completas para la abundancia de mensajes, uso de isoformas, edición de ARN y expresión específica de alelos (Reuter, Spacek, & Snyder, 2015).

Finalmente, los avances en HTS han permitido una extensa catalogación de muestras metagenómicas, proporcionando información sobre la diversidad de especies microbianas de una amplia variedad de fuentes, incluidos el océano, el suelo y el cuerpo humano.

### ***Metagenómica de virus (viroma)***

El término 'metagenomics', metagenómica en español, es usado cuando se habla del análisis de todos los genomas presentes en una muestra usando secuenciación masiva (Nieuwenhuijse & Koopmans, 2017).

Dado que los virus carecen de motivos universalmente conservados, la metagenómica viral hace referencia a la recuperación y análisis de los genomas completos y parciales de todos los virus presentes en una muestra. Los protocolos de análisis de metagenómica de virus implican procedimientos para remover células bacterianas y del hospedero, tratamientos para remover ácidos nucleicos libres mediante nucleasas, amplificación con cebadores (primers) aleatorios utilizando Transcripción inversa y Reacción en Cadena de la Polimerasa (RT-PCR), y finalmente secuenciación masiva paralela (Nieuwenhuijse & Koopmans, 2017). El desarrollo de las técnicas de NGS, especialmente su reducción de costos en los últimos años, ha revolucionado la metagenómica, permitiendo la secuenciación de material genético de comunidades microbianas completas sin las restricciones de las técnicas de cultivo y la limitación del análisis de marcadores genéticos (Rastrojo & Alcamí, 2017).

La metagenómica viral presenta un gran potencial dentro de programas de vigilancia de virus en la cadena alimentaria gracias al amplio rango de detección, sensibilidad e información detallada que puede proveer de los virus detectados. Sin embargo, existen varios inconvenientes para la metagenómica, entre ellos el análisis de datos (bioinformática), la composición de la microbiota de la muestra y la falta de secuencias de referencia en las bases de datos públicas, lo que requiere flujos de trabajo bioinformáticos ('pipelines') complejos y muchas veces específicos para el tipo de datos (Sekse et al., 2017).

### ***Ensamblaje de genomas a partir de datos de secuenciación masiva***

Previa clasificación o asignación taxonómica, es necesario ensamblar las lecturas cortas en secuencias continuas (contigs) de una longitud de nucleótidos mayor y generar secuencias consenso al alinear dichos contigs. Esto permite reducir el error de las lecturas individuales y también la cantidad de datos. El ensamblaje de genomas se puede realizar de dos formas: 1) en base al mapeo de lecturas frente a un genoma de referencia o 2) el llamado ensamblaje *de novo*, que consiste en unir las lecturas en base a solapamientos, frecuencia o lecturas de dos sentidos (Paired-end, PE). Dado que para muchos virus no existen secuencias de referencia completas y curadas, el ensamblaje *de novo* es la mejor opción en análisis de virus (Nooji et al., 2018). Además, el ensamblaje *de novo* de metagenomas facilita el descubrimiento y reconstrucción de genomas nuevos gracias a una predicción de genes más confiable (van der Walt et al., 2017).

En la evaluación de la calidad del ensamblaje se toman en cuenta varios parámetros, entre ellos, el valor N50 que corresponde al tamaño del contig más pequeño tal que el 50% del genoma está contenido en contigs de tamaño N50 o mayor. Por otro lado, el valor L50 hace referencia al número de contigs que son más largos o iguales a la longitud N50 y, por lo tanto, incluyen la mitad de los nucleótidos del genoma o conjunto de datos. Una buena calidad de ensamblaje permite una correcta determinación de la composición de una comunidad microbiana, la comparación entre comunidades de diversos entornos, la extracción de genomas completos a partir de metagenomas y el aislamiento informado de microorganismos. Cada pregunta de investigación requiere enfatizar características específicas del metagenoma. Si el

objetivo son genomas entonces se requieren contigs largos, mientras que si se evalúan genes, los contigs deben ser de alta calidad y confianza (van der Walt et al., 2017).

### **Hipótesis**

El agua de riego del ramal “El Pueblo” No. 12. contiene virus fitopatógenos, convirtiéndolo en una fuente potencial para la transmisión de partículas virales hacia cultivos de importancia económica.

## CAPÍTULO 2: MATERIALES Y MÉTODOS

### **Zona de estudio**

El trabajo de investigación se realizó en el Laboratorio de Microbiología de la carrera de Ingeniería en Biotecnología de la Universidad de las Fuerzas Armadas - ESPE, que se encuentra ubicado en la Av. General Rumiñahui S/N y Sector La Colina (coordenadas: 0°18'53"Sur 78°26'36"Oeste).

La obtención de las muestras se realizó en el reservorio de agua de riego de la Granja Experimental Tumbaco del Instituto Nacional de Investigaciones Agropecuarias (INIAP), ubicado en la Av. Interoceánica km 15 y Eloy Alfaro, sector Tumbaco (coordenadas: 00°12'56"Sur 78°24'44" Oeste), Provincia de Pichincha. Este reservorio es alimentado con agua proveniente del ramal "El Pueblo" No. 12 del Sistema de Riego Tumbaco. Desde este reservorio se distribuye agua a todos los cultivos de frutales que se mantienen en dicho lugar, representando una fuente potencial de transmisión de virus fitopatógenos, más aún con el antecedente de que varios cultivos de la Granja Tumbaco (INIAP) han presentado síntomas de virosis en los últimos 4 años.

### **Duración de la investigación**

La investigación inició el 10 de mayo del 2020 y finalizó el 10 de diciembre del 2020.

### **Procedimientos**

#### ***Toma de muestras***

Se realizaron 2 recolecciones de muestras de agua de riego en el mismo sitio y se tomaron 3 muestras de 10 L cada una utilizando contenedores plásticos estériles. Las 6 muestras fueron tomadas desde la parte central del reservorio a una profundidad aproximada de 20 cm. Con el fin de eliminar residuos vegetales, las muestras fueron filtradas previamente utilizando un filtro casero de 50µm conectado a una bomba de agua periférica de 0,6 HP. Las muestras fueron procesadas en un periodo mínimo de 3 horas posteriores a la recolección para evitar degradación de los ácidos nucleicos.

### **Concentración viral**

Siguiendo la metodología descrita por Calgua *et al.* (2008), se concentraron las partículas virales mediante floculación orgánica con leche descremada. Utilizando un pH metro portátil marca Horiba modelo PH210K, se ajustó el pH de cada muestra hasta el valor de 3,5. Posteriormente, a cada muestra (10 L) se agregaron 100 mL de una solución 1% (w/v) de leche descremada (Difco™) preparada en agua de mar artificial (33,33 g de sal marina por litro de agua destilada) a pH de 3,5. Las muestras fueron agitadas por 8 h a temperatura ambiente en una plancha agitadora Cimarec (Thermo Scientific™), seguido de un período de decantación de los sedimentos de 8 h adicionales. Se removió el sobrenadante y se recolectó el sedimento en tubos plásticos estériles. El sedimento recolectado fue sometido a centrifugación a 4000 rpm por 90 min a 4°C en una centrífuga refrigerada marca Eppendorf modelo 5810r. El pellet fue resuspendido en buffer fosfato 0,2 M (1:2 v/v de Na<sub>2</sub>HPO<sub>4</sub> y NaH<sub>2</sub>PO<sub>4</sub> marca Himedia) a pH 7,5, hasta alcanzar un volumen final de 10 mL. Se almaceno el concentrado viral en un ultra congelador FORMA 7000 marca Thermo Scientific™ a -80°C hasta su posterior uso.

### **Extracción de ARN**

Se obtuvo ARN total a partir de 600µL de concentrado viral y 50 mg de tejido de hojas de Cucarda (*Hibiscus rosa-sinensis* L.) utilizando el kit comercial SV Total RNA Isolation System (Promega Corporation) según las instrucciones del fabricante. Una vez extraído el ARN total, se almacenó a -80°C para uso a largo plazo.

Todo el proceso de extracción se llevó a cabo en frío y se utilizó RNaseZap (Ambion The RNA Company, USA), inhibidor de RNasas sobre las superficies y equipos para evitar la degradación del ARN.

### **Transcripción inversa y Reacción en Cadena de la Polimerasa**

Como un punto de control en el proceso, se sintetizó ADN complementario (ADNc) a partir del ARN previamente extraído, utilizando el kit comercial GoScript™ Reverse Transcription System (Promega Corporation), según instrucciones del fabricante. Adicionalmente, a calidad y cantidad del ARN extraído se obtuvo usando los ratios A260/A280 y A260/A230 mediante un espectrofotómetro Nanodrop 2000 (Thermo Scientific).

Se amplificó el ADNc mediante PCR convencional, usando primers genéricos diseñados por Ochoa Corona, Fletcher, & Olson (2014), para amplificar la ARN polimerasa dependiente de ARN de virus de los géneros: *Potexvirus* (forward 5'-GAYGGIGCIATGCTICAITT-3' y reverse- 5' TCIGTRTTIGCRTCRAAIGT-3') y *Tombusvirus* (forward 5'-TGGATATTIATGGTTGCIGGTT-3' y reverse 5'-GTAGGTTGTGGAGTGCAG-3'). Los dos juegos de primers fueron también utilizados por Guevara (2017) para la detección de virus fitopatógenos en plantas de babaco y agua de riego. La RT-PCR se realizó con la finalidad de observar la presencia de ARN viral en las extracciones previo a su envío para secuenciación.

La reacción de PCR se ensambló utilizando 1X Go Taq Green Master Mix (Invitrogen), 0,6  $\mu$ M de cada primer (Forward y Reverse), 1,5 mM  $MgCl_2$  (Promega) y entre 400 a 600 ng/ $\mu$ L de ADNc. La amplificación se realizó en el termociclador ProFlex PCR System (Applied Biosystems) aplicando un ciclo de 95°C por 3 min, seguido de 35 ciclos de 95°C por 30 s, hibridación de 47 y 50°C por 30 s para *Potexvirus* y *Tombusvirus*, respectivamente; y 72°C por 1 min. Finalmente un ciclo de extensión de 72°C por 2 min. Se realizó la electroforesis de 5  $\mu$ L de producto de PCR en un gel de agarosa al 2% (w/v) por 40 min a 90 V. El gel fue teñido con Bromuro de Etidio (AMRESCO) y visualizado en un transiluminador de luz UV.

### **Secuenciación masiva**

El ARN total extraído de las 6 muestras fue precipitado utilizando 0,1 volúmenes de acetato de sodio 3M a pH= 5.5 y 2 volúmenes de etanol absoluto y enviado a la compañía ©Macrogen Inc. (Seúl, Corea del Sur) para el servicio de secuenciación masiva. La empresa realizó el control de calidad y enriquecimiento de las muestras (ribosomal RNA depletion) para eliminar ARN ribosomal bacteriano contaminante utilizando el kit Ribo-Zero rRNA Removal Kit (Bacteria). Se prepararon librerías de secuenciación usando el kit TruSeq RNA Sample Prep Kit v2 (Illumina, Inc., USA). La secuenciación masiva se realizó mediante la tecnología de secuenciación por síntesis (SBS, por sus siglas en inglés) en la plataforma NovaSeq 6000 de Illumina mediante secuenciación de extremo pareado (paired-end, PE) con lecturas de 101 nt de longitud.

### **Análisis bioinformático**

El análisis de los datos se realizó en el clúster de la Universidad de las Fuerzas Armadas-ESPE. El pipeline o flujo de trabajo (Figura 1) consistió, en su mayoría, en la ejecución de programas de código abierto, como se indica a continuación:

1) Control de calidad y pre-procesamiento. Mediante el programa FastQC se analizó la calidad de las secuencias crudas para evaluar el proceso de secuenciación. Utilizando el programa Trimmomatic-0.39 se removieron los adaptadores, se eliminaron lecturas cortas y con un índice de calidad (Phred Quality) menor a 20. Adicionalmente se eliminaron secuencias duplicadas (deduplicación) utilizando el programa BBDup 38.86

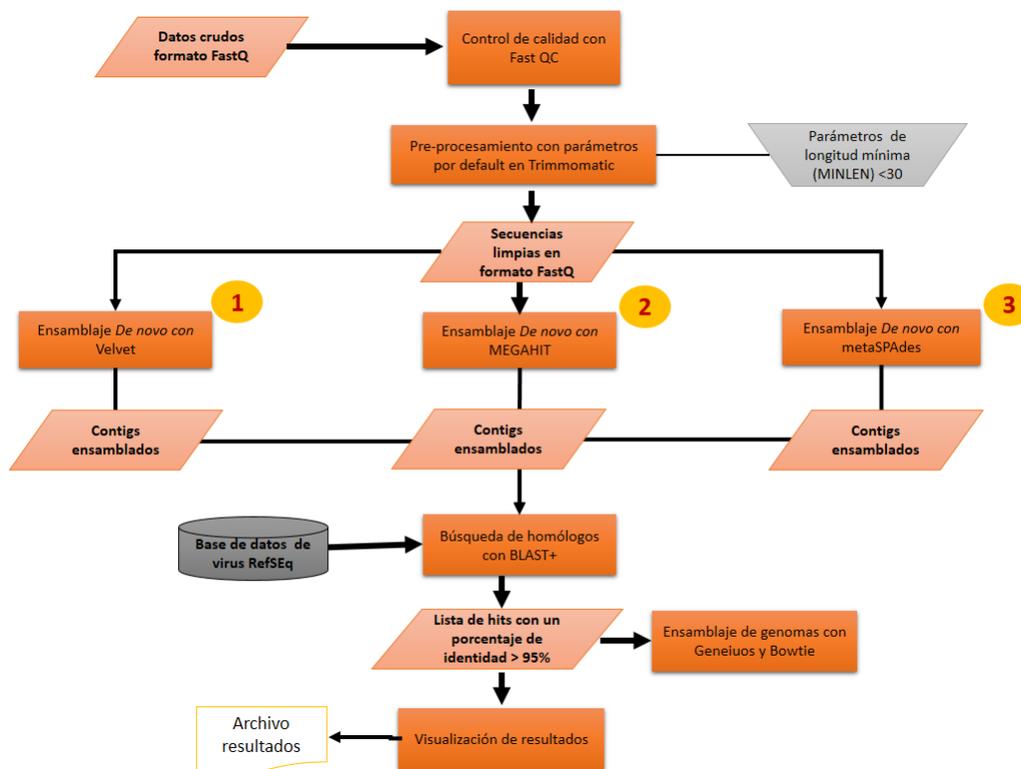
2) Ensamblaje *de novo*. Se utilizaron tres diferentes herramientas de ensamblaje: Velvet v1.2.10, metaSPAdes v.3.14.1 y Megahit v. 1.0.6. Los programas metaSPAdes y Megahit realizan un análisis iterativo del tamaño de k-mer para encontrar el valor óptimo, por lo que se utilizaron k-mers de 21, 33 y 55 para metaSPAdes y 21, 33, 45 y 57 para Megahit. Para el ensamblador Velvet se utilizó un k-mer de tamaño 55.

3) Mapeo o búsqueda de homólogos. Se realizó un análisis de similitud de los contigs ensamblados con genomas virales de referencia depositados en la base de datos RefSeq del NCBI (National Center for Biotechnology Information), mediante el programa Blast-2.10.1+ y comparación con un porcentaje de identidad del 95%.

4) Ensamblaje de genomas completos. Al encontrar genomas parcialmente ensamblados utilizando las herramientas antes mencionadas, se procedió a buscar y ensamblar las lecturas faltantes de dichos genomas. Con ayuda de los programas Bowtie2-2.4.1 (programa de código abierto) y Geneious 2020.2.2 (programa que requiere licencia), se realizó el ensamblaje de lecturas usando como referencia genomas completos curados y revisados disponibles en la base de datos RefSeq del NCBI. El programa Geneious 2020.2.2 también fue utilizado para realizar la anotación de genomas y posterior envío de las secuencias para su registro dentro de la base de datos GenBank del NCBI.

Figura 1

Diagrama de flujo del análisis bioinformático a partir de datos de secuenciación masiva.



Nota: Los círculos amarillos denotan las herramientas de ensamblaje que dividen el proceso en tres flujos de trabajo diferentes a ser comparados.

### Comparación de herramientas

Se aplicó el flujo de trabajo (Figura 1) variando el parámetro MINLEN del programa Trimmomatic. La opción MINLEN hace referencia a la longitud mínima que deben tener las lecturas para ser utilizadas en análisis posteriores. Durante el pre-procesamiento con Trimmomatic las lecturas menores al valor especificado en MINLEN, serán eliminadas. Por una parte se utilizó el valor predeterminado (lecturas <50 nt) y el valor definido de eliminación de lecturas con una longitud <30 nt.

Adicionalmente, se probó el flujo de trabajo con los tres diferentes ensambladores antes mencionados: Velvet, metaSPAdes y Megahit. Utilizando el software MetaQuast se realizó la comparación entre ensambladores con base en el

número de contigs obtenidos, los parámetros estadísticos N50 y L50, y el porcentaje de alineamiento con secuencias de referencia (%GA).

Para evaluar el porcentaje de alineamiento se utilizaron como referencia 7 genomas de virus fitopatógenos presentes en la muestra experimental: *Hibiscus chlorotic ringspot virus* (HCRSV), *Hibiscus latent Fort Pierce virus* (HLFPV), *Pepper mild mottle virus* (PMMoV), *Ryegrass mottle virus* (RGMoV), *Tobacco mosaic virus* (TMV), *Tomato bushy stunt virus* (TBSV) y *Tomato mosaic virus* (ToMV). Adicionalmente, como replicas, se utilizaron lecturas de Illumina generadas *in silico* mediante el programa InSilicoSeq versión 3, a partir de los 7 genomas previamente mencionados. El análisis estadístico se realizó usando el software SAS® Studio. La prueba de normalidad Shapiro-Wilk fue aplicada para determinar la distribución normal de los residuos. Dado que las muestras no presentaron una distribución normal se aplicó la prueba no paramétrica Kruskal-Wallis con un nivel de confianza del 95%. Se realizó una comparación múltiple de los tratamientos mediante el procedimiento Dwass, Steel, Critchlow-Fligner (DSCF) para determinar la mejor combinación MINLEN-Ensamblador. Según el parámetro de secuencias a ser eliminadas y el ensamblador, se obtuvieron las combinaciones presentadas en la Tabla 1.

**Tabla 1**

*Combinaciones de herramientas bioinformáticas evaluadas para análisis de datos de metagenómica de virus.*

<b>Tratamientos (Combinación MINLEN-Ensamblador)</b>	<b>Código*</b>
Trimmomatic con parámetro MINLEN<50 + Ensamblador Velvet.	TDFV
Trimmomatic con parámetro MINLEN<50 + Ensamblador metaSPAdes.	TDFSmeta
Trimmomatic con parámetro MINLEN<50 + Ensamblador MEGAHIT	TDFM
Trimmomatic con parámetro MINLEN<30 + Ensamblador Velvet.	TUSV

<b>Tratamientos (Combinación MINLEN-Ensamblador)</b>	<b>Código*</b>
Trimmomatic con parámetro MINLEN<30+ Ensamblador metaSPAdes.	TUSSmeta
Trimmomatic con parámetro MINLEN<30+ Ensamblador MEGAHIT.	TUSM

Nota: \*TDF=Trimmomatic default; TUS=Trimmomatic user

### ***Validación del pipeline***

Se validó el pipeline usando datos generados *in silico* mediante el programa InSilicoSeq versión 3. Se generaron 6 millones de lecturas Illumina simuladas a partir de genomas completos de 7 virus fitopatógenos identificados previamente en la muestra experimental.

### ***Identificación de Hibiscus latent Fort Pierce virus (HLFPV) en agua de riego y plantas de Cucarda (Hibiscus rosa-sinensis L.).***

Se amplificó el ADNc utilizando primers específicos previamente reportados por Kamenova and Adkins (2004) y diseñados para la cápside de HLFPV: primer forward 5'-GACACAGAGAAATCTGGGGCAG-3' y primer reverse 5'-TTAGTTAGCAGGAGCGGTCCA-3'. Se utilizaron 1X Go Taq Green Master Mix (Invitrogen), 0,6 µM de cada uno de los primers (Forward y Reverse) y entre 400 a 600 ng/µL de ADNc para la reacción de PCR. La amplificación se realizó en el termociclador ProFlex PCR System (Applied Biosystems) aplicando un ciclo de 94°C por 5 min, seguido de 30 ciclos de 94°C por 1 min, hibridación a 59°C por 1 min y 72°C por 1:30 min. Finalmente un ciclo de extensión de 72°C por 10 min. La electroforesis de 5 µL de producto de PCR se realizó en un gel de agarosa al 2% (w/v) por 40 min a 90 V. El gel fue teñido con Bromuro de Etidio (AMRESCO) y visualizado en un transiluminador de luz UV. Uno de los amplicones fue enviado a secuenciación Sanger en la empresa ©MacroGen Inc. (Seúl, Corea del Sur). La limpieza y ensamblaje se realizó utilizando el programa Geneious 2020.2.2 y la comparación de secuencia en la base de datos GenBank del NCBI mediante la herramienta online BLASTn del NCBI.

## CAPÍTULO 3: RESULTADOS

### Lugar de muestreo

Se observó que las condiciones del reservorio (Figura 2A) variaron entre cada toma de muestras, donde la temperatura osciló entre los 18 a 24°C y el pH del agua entre 9 y 10. Además se observó la presencia de un cerco de plantas de Cucarda (*Hibiscus rosa-sinensis* L.) en dos extremos del reservorio (Figura 2B).

### Figura 2

*Zona de toma de muestras en la Granja Experimental INIAP Tumbaco.*

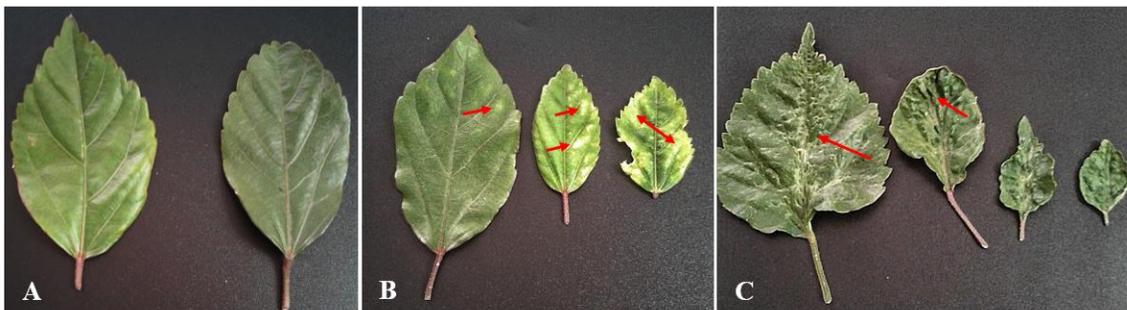


Nota: A) Entrada de agua desde el sistema de riego Tumbaco. B) Reservorio de agua de riego. Las flechas señala la presencia de plantas de Cucarda en la periferia del reservorio.

Las plantas de Cucarda observadas en el cerco presentaban síntomas de virosis como mosaico (Figura 3B) y ampollamiento (Figura 3C), por lo que fueron recolectadas y almacenadas a -80°C con sílica gel para análisis posteriores.

### Figura 3

*Hojas de Cucarda.*



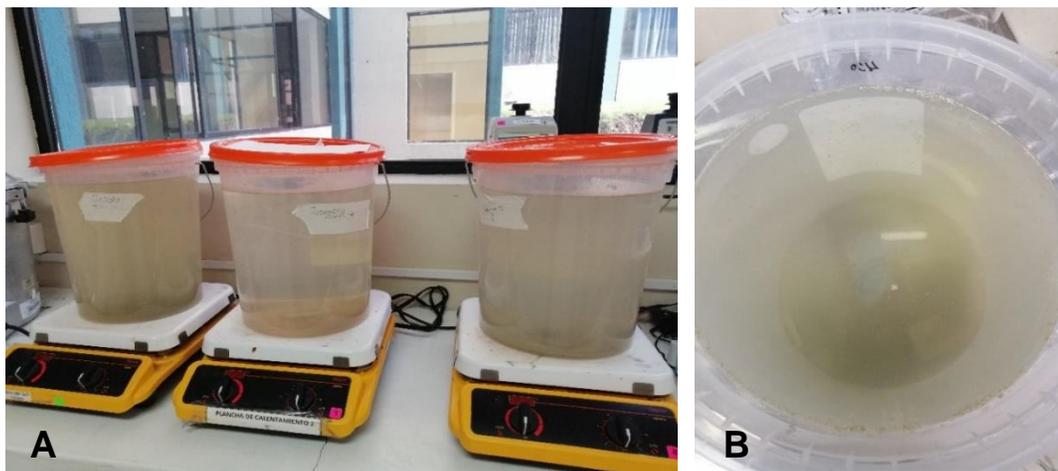
Nota: A) hojas sanas-control. B) Hojas con síntoma de mosaico. C) Hojas con síntoma de ampollamiento.

### Concentración viral

El método para concentración de partículas virales mediante floculación orgánica permitió la obtención de aproximadamente 500 mL de sedimento a partir del cual se realizó la extracción de ARN total (Figura 4).

### Figura 4

*Concentración viral con el método de floculación orgánica.*



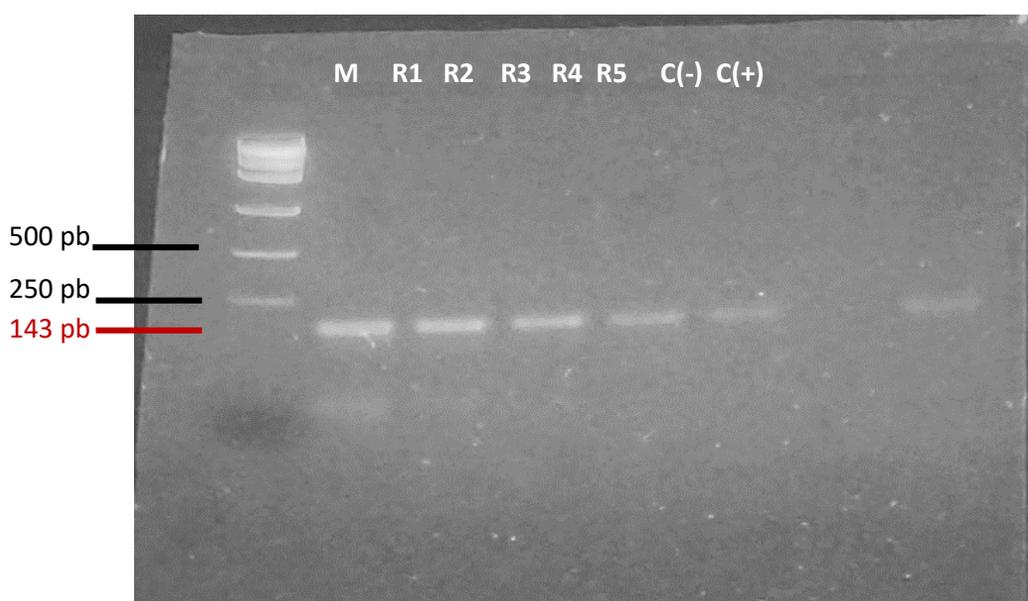
Nota: A) Muestras en agitación. B) Flóculo formado después de decantación.

### Amplificación de muestras mediante PCR convencional

Se evaluó mediante PCR la presencia de partículas virales utilizando ADNc sintetizado a partir de las muestras de ARN total, previo a su envío para secuenciación masiva. Cinco de las seis muestras tomadas resultaron positivas para virus del género *Tombusvirus* con un amplicón de 143 nt (Figura 5), que coincide con el tamaño esperado descrito por Ochoa-Corona, Fletcher, & Olson (2014). Ninguna muestra amplificó para virus del género *Potexvirus*.

### Figura 5

*Electroforesis de los productos de PCR.*



Nota: gel de agarosa al 2% donde se visualizan los productos de PCR de muestras de agua amplificadas con primers para *Tombusvirus*. Pocillo 1: marcador de peso molecular, Pocillo 2 a 6: muestras de agua, Pocillo 7: control negativo y Pocillo 8: control positivo.

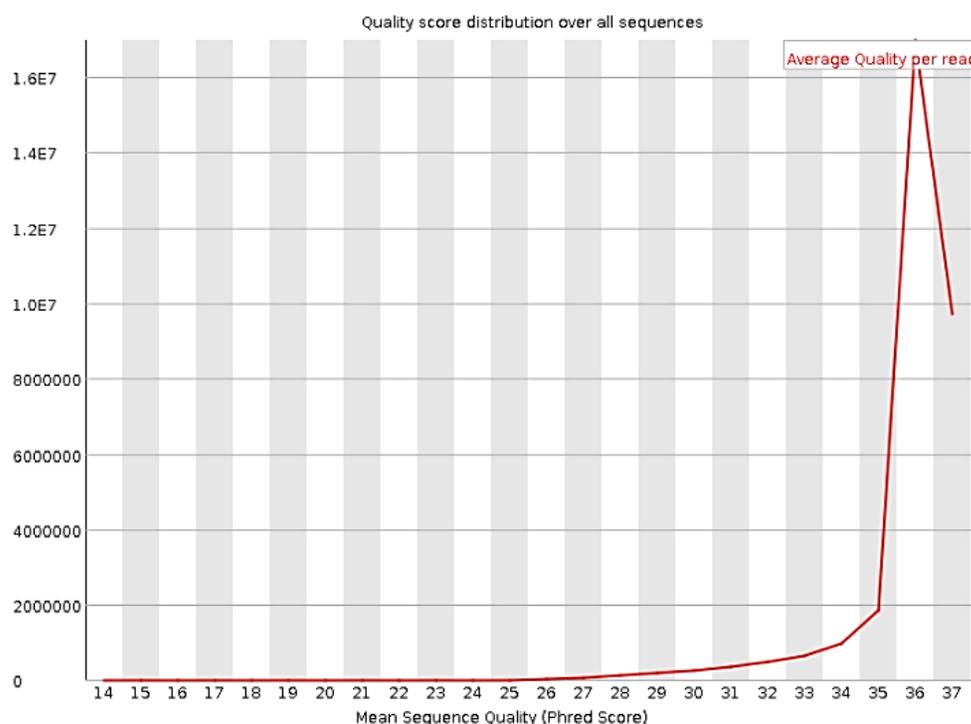
### Secuenciación masiva y análisis bioinformático

Las muestras en las que se evidenció la presencia de partículas virales fueron enviadas para secuenciación en la plataforma Illumina según descripción previa en la sección de Materiales y Métodos. Se obtuvieron archivos en formato .fastq.gz conteniendo las secuencias crudas con un total de 63,909,998 lecturas, un %GC de 50.3, %AT de 49.7 y una calidad Q20 de 98.21%.

Todas las herramientas descritas fueron utilizadas con parámetros predeterminados por cada software con excepción de la remoción de secuencias cortas y los ensambladores. El control de calidad (Figura 6) mostró una calidad (Phred Score) promedio de las lecturas de 36.

## Figura 6

*Gráfico de puntaje de calidad de las secuencias crudas.*



Nota: Gráfico generado utilizando la herramienta FastQC.

## Comparación de pre-procesamiento y ensambladores

Se evaluó el efecto de la depuración de lecturas cortas menores a 30 y 50 nt en el ensamblaje final de genomas virales. En la Tabla 2 se observa el número de lecturas después del proceso de depuración con la herramienta Trimmomatic en las muestras experimental y simulada. Lógicamente se observa que al aplicar el parámetro establecido por el programa (MINLEN) < 50 nt el número de lecturas es menor en comparación con la segunda corrida aplicando un parámetro de longitud mínima de 30 nucleótidos.

**Tabla 2**

*Número de lecturas obtenidas después del pre-procesamiento aplicando dos tamaños de longitud para eliminar secuencias cortas (parámetro MINLEN en el programa Trimmomatic).*

	<b>Muestra experimental</b>		<b>Muestra simulada</b>	
	MINLEN <50 nt	MINLEN <30 nt	MINLEN <50 nt	MINLEN <30 nt
<b>Lecturas crudas</b>	63,909,998		7,999,994	
<b>Lecturas del tamaño especificado</b>	61,379,186	62,223,328	7,753,538	7,814,708
<b>Lecturas después de deduplicación</b>	32,743,168	33,570,524	7,753,536	7,814,706

Después del filtrado y remoción de lecturas de mala calidad y lecturas cortas se realizó el ensamblaje con los diferentes programas donde metaSPAdes mostró el mayor número de contigs generados (Tabla 3) para la muestra experimental utilizando los dos valores de MINLEN (50 y 30 nt). Por el contrario Velvet presentó el menor número de contigs ensamblados con los dos valores de MINLEN. Se evidenció que metaSPAdes y Megahit recuperaron los 7 genomas simulados en un solo contig largo cada uno, mientras que Velvet generó un gran número de contigs de menor tamaño en la muestra simulada.

**Tabla 3**

*Número de contigs totales ( $\geq 1000$  nt) obtenidos con cada ensamblador y parámetro MINLEN.*

<b>Ensamblador</b>	<b>Muestra experimental</b>		<b>Muestra simulada</b>	
	MINLEN <50 nt	MINLEN <30 nt	MINLEN <50 nt	MINLEN <30 nt
<b>Velvet</b>	2,594	2,667	39,433	39,643
<b>metaSPAdes</b>	10,592	10,689	7	7
<b>MEGAHIT</b>	9,059	9,163	7	7

**Tabla 4**

*Parámetro N50 estimado para el ensamblaje con cada programa utilizado.*

<b>Ensamblador</b>	<b>Muestra experimental</b>		<b>Muestra simulada</b>	
	MINLEN <50 nt	MINLEN <30 nt	MINLEN <50 nt	MINLEN <30 nt
<b>Velvet</b>	293	299	109	109
<b>metaSPAdes</b>	374	374	6356	6356
<b>MEGAHIT</b>	495	495	6303	6303

El ensamblador que mostró el parámetro N50 más alto fue Megahit en los dos valores probados para MINLEN (Tabla 4). Por otro lado el ensamblador Velvet mostró el valor más bajo en dicho parámetro. En el caso del parámetro L50 se pudo observar que metaSPAdes y Velvet presentaron los valores más altos y bajos, respectivamente, para los dos valores de MINLEN, según los resultados de MetaQuast (Tabla 5).

**Tabla 5**

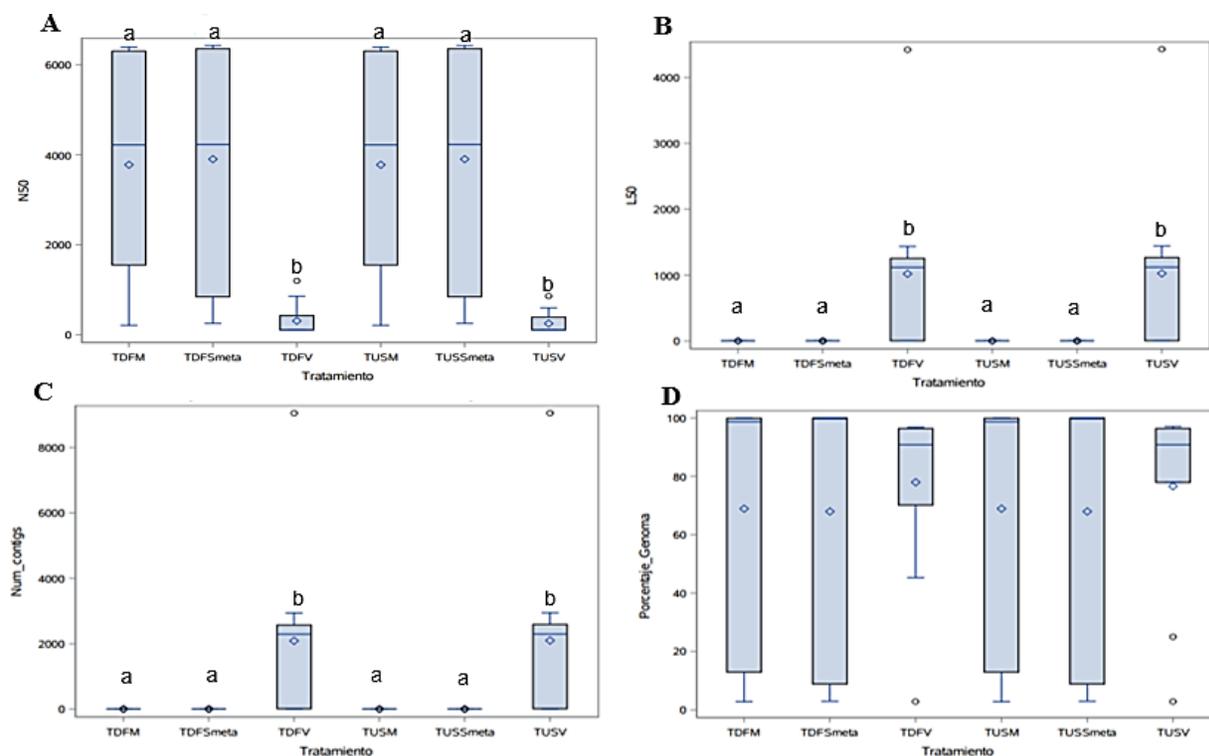
*Parámetro L50 estimado para el ensamblaje con cada programa utilizado.*

Ensamblador	Muestra experimental		Muestra simulada	
	MINLEN <50 nt	MINLEN <30 nt	MINLEN <50 nt	MINLEN <30 nt
<b>Velvet</b>	34,391	33,861	19,300	19,401
<b>metaSPAdes</b>	70,540	70,878	4	4
<b>MEGAHIT</b>	36,092	36,444	4	4

El análisis estadístico mostró diferencia significativa entre las combinaciones MINLEN-ensamblador analizadas (valores de  $p < 0.05$ ) en cuanto al número de contigs ensamblados ( $p < 0.0001$ ), parámetro N50 ( $p < 0.0001$ ) y parámetro L50 ( $p < 0.0001$ ). Sin embargo, el porcentaje de genoma alineado no mostró diferencia significativa entre tratamientos ( $p = 0.2372$ ). Las comparaciones múltiples mostraron que los ensambladores metaSPAdes y Megahit difieren del ensamblador Velvet ( $p < 0.05$ ) pero son similares entre ellos para todas las variables de respuesta analizadas (Figura 7). De igual manera no se observaron diferencias entre aquellas combinaciones que incluyeron MINLEN establecido por el usuario o el parámetro con valor predefinido, solamente diferencia entre tipo de ensamblador.

Figura 7

Diagrama BoxPlot



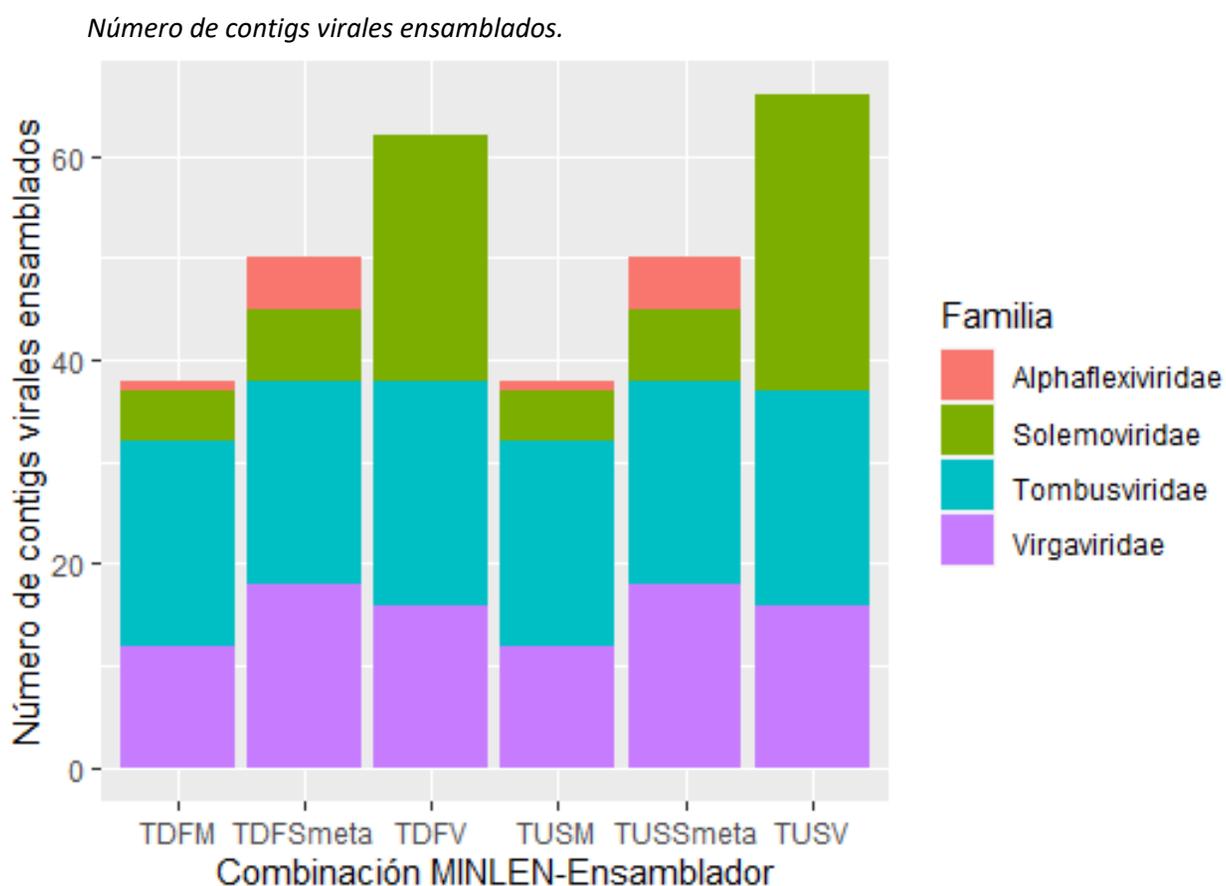
Nota: grafico generado utilizando los datos del ensamblaje de 7 genomas de virus fitopatógenos y 6 combinaciones MINLEN-Ensamblador. A) parámetro N50. B) parámetro L50. C) Número de contigs ensamblados y D) porcentaje de genoma alineado.

En la Figura 8 se muestran los resultados de la búsqueda de homólogos en la base de datos de virus de los contigs ensamblados por el software metaSPAdes. En la asignación taxonómica se observó que el 60% de datos se encuentran asignados dentro del dominio Duplodnaviria (virus de ADN) y el 36% dentro del dominio Riboviria (virus de ARN); la proporción restante lo componen secuencias identificadas como satélites (1%), Virtovirus (1%), Varidnaviria (1%), Baculoviridae (0.7%) y virus no clasificados (0.3%).



Dentro del conjunto de contigs ensamblados se encontraron cuatro familias de virus fitopatógenos (Figura 9) de un total de 19 familias virales. Sin embargo se observó que el ensamblador Velvet, independientemente del parámetro MINLEN, no ensambló contigs de un tamaño adecuado para ser identificados dentro de la familia *Alphaflexiviridae*.

**Figura 9**

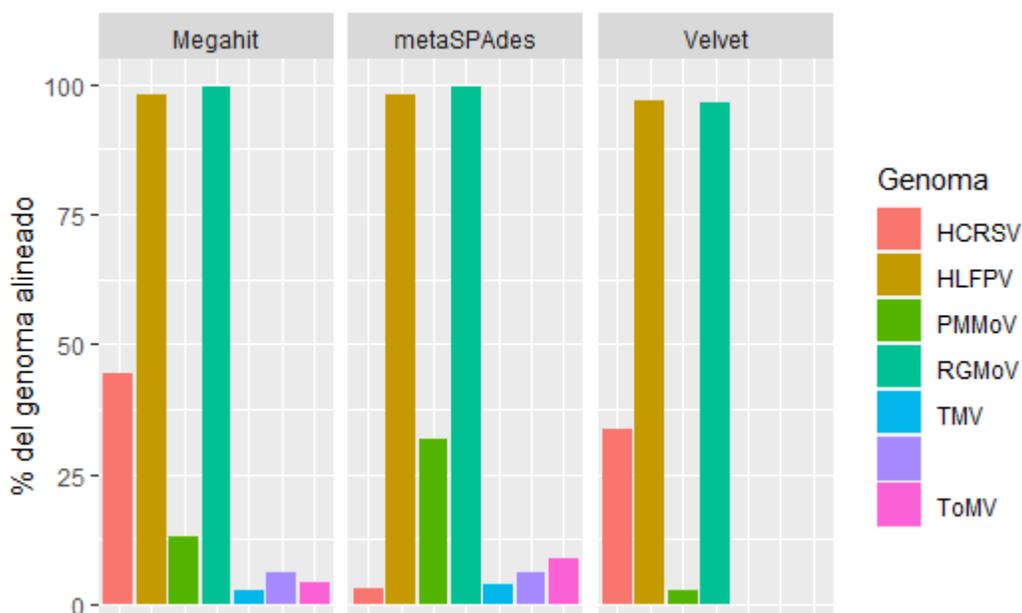


Nota: ensamblajes correspondientes a familias de virus fitopatógenos según las combinaciones MINLEN-Ensamblador probadas dentro del flujo de trabajo.

Al comparar los ensambladores utilizando las secuencias completas de siete genomas de referencia (Figura 10), se observó que los tres ensambladores probados lograron ensamblar más del 90% del genoma de *Hibiscus latent Fort Pierce virus* (HLFPV) y *Ryegrass mottle virus* (RGMoV). Sin embargo, de todos los genomas de referencia usados, Velvet sólo ensambló cuatro de ellos.

**Figura 10**

Porcentaje del genoma alineado con secuencias de referencia.



Nota: alineamiento con genomas de referencia de Hibiscus chlorotic ringspot virus (HCRSV), Hibiscus latent Fort Pierce virus (HLFPV), Pepper mild mottle virus (PMMoV), Ryegrass mottle virus (RGMoV), Tobacco mosaic virus (TMV), Tomato bushy stunt virus (TBSV) y Tomato mosaic virus (ToMV), usando los contigs ensamblados con Megahit, metaSPAdes y Velvet.

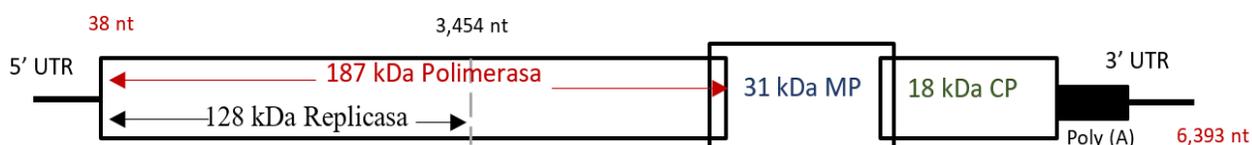
### Ensamblaje y anotación de Genomas

A partir de los datos obtenidos en el ensamblaje y comparación de ensambladores, se buscó el genoma completo de los virus mencionados en la Figura 9. El ensamblador metaSPAdes dio como resultado un contig de 6,313 nt identificado como HLFPV. Utilizando la herramienta Bowtie2-2.4.1 se mapearon las lecturas crudas frente al genoma de referencia (No. accesión NC\_025381.1), que tiene una longitud de 6,431 nt. Se obtuvieron 22,914 lecturas concordantes con una cobertura de 362x que posteriormente fueron ensambladas *de novo* utilizando el software Geneious 2020.2.2. El ensamblado final con una longitud de 6,393 nt mostró un 99% de identidad por pares con las secuencias disponibles en la base de datos del NCBI, indicando que el 99.41% del genoma de HLFPV fue ensamblado. La organización del genoma ensamblado (Figura 11) coincide con la estructura típica reportada para virus de ARN del género *Tobamovirus* (Gao et al., 2016): una región no traducida 5'-UTR de 37 nucleótidos

(nt), 11 nt menos que la región 5'UTR del genoma de referencia; y una región no traducida 3'-UTR de 131 nt incluyendo una cola de adeninas (Poly A) de 59 nt, que es 27 nt más pequeña que la región 3'UTR del genoma de referencia; y 4 marcos abiertos de lectura (ORFs, por sus siglas en inglés) que codifican la polimerasa, la replicasa, la proteína de movimiento (MP) y la cápside (CP) del virus. La secuencia obtenida de HLFPV fue depositada en la base de datos GenBank del NCBI con el número de accesoión MT942636.

## Figura 11

*Estructura del genoma de HLFPV.*



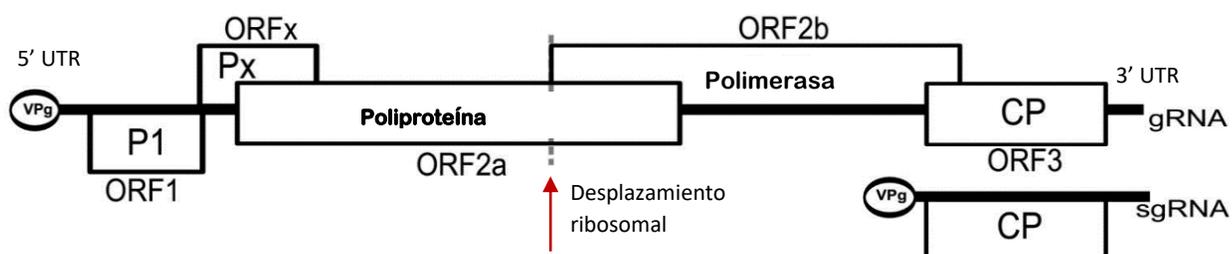
Nota: Los cuadros representan las regiones codificantes de la polimerasa, replicasa, proteína de movimiento (MP) y la cápside (CP) y los números las posiciones de inicio y fin de dichas regiones. Las líneas horizontales representan las regiones no traducidas 5' y 3'-UTR y el cuadro negro representa la cadena interna de Poly(A) en la región 3'-UTR.

Por otra parte, el ensamblador metaSPAdes dio como resultado un único contig de 4,247 nt de longitud que se identificó como *Ryegrass mottle virus* (RGMoV) con un 95% de identidad tras la búsqueda de homólogos con Blast-2.10.1+. El mapeo de las lecturas correspondientes a RGMoV con el genoma de referencia de 4,212 nt (No. de accesoión NC\_003747.2) mostró una profundidad de cobertura de 107.681x utilizando la herramienta Bowtie2-2.4.1. Como se visualiza en la Figura 12, con el software Geneious 2020.2.2, se identificaron cinco marcos abiertos de lectura (ORFs) en el genoma de ARN de cadena simple, una región no traducida 5'-UTR de 87 nt, 12 nt menos que el genoma de referencia y una región 3'-UTR de 247 nt, con 47 nt más que el genoma de referencia. Esta estructura concuerda con la organización típica de virus del género *Sobemovirus* y lo reportado para RGMoV (Ling et al., 2013; Sõmera et al., 2015). La secuencia obtenida RGMoV fue depositada en la base de datos GenBank del NCBI con el número de accesoión MW411579. Los datos crudos están disponibles en la base 'Sequence Read Archive' (SRA) bajo los números de accesoión SRR13980363 y

SRR13070794. Estos resultados fueron publicados en la revista 'Microbiology Resource Announcements' de la Sociedad Americana de Microbiología (Guevara & Flores, 2021).

## Figura 12

*Estructura del genoma de RGMoV.*



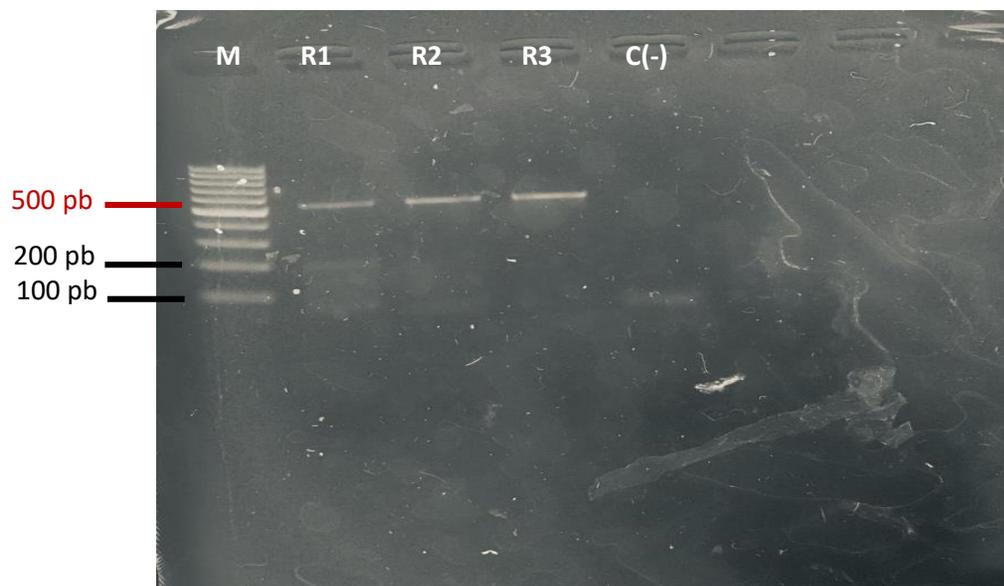
Nota: figura adaptada de Sõmera et al. (2015). P1: primera proteína, Px: proteína X, CP: cápside, VPg: proteína ligada al genoma viral.

### Identificación molecular de HLFPV en agua de riego y plantas de Cucarda

Mediante el pipeline bioinformático se identificó la presencia de HLFPV en agua de riego, por lo que dichos resultados fueron corroborados en las muestras de agua y plantas de Cucarda cercanas al reservorio (Figura 2) utilizando primers específicos para amplificar la cápside de HLFPV. Se obtuvieron amplicones con un tamaño de 535 nt (Figura 13), que coincidieron con lo reportado por Kamenova y Adkins (2004). Los amplicones fueron secuenciados y una secuencia representativa fue subida a la base de datos GenBank (No. de acceso MW601340). La comparación de la secuencia con las secuencias de referencia de la base de datos de GenBank del NCBI mostró un 99.79% de identidad con la secuencia de nucleótidos de la cápside de HLFPV (Secuencia de referencia AY250831.1).

**Figura 13**

*Electroforesis de productos de PCR para HLFPV.*



Nota: gel de agarosa al 2% donde se visualizan productos de PCR de muestras de agua amplificadas con primers para HLFPV. Pocillo 1: marcador 100 nt (PROMEGA), Pocillo 2 a 4: muestras de agua concentrada por el método de Leche Descremada (SKM), Pocillo 5: control negativo.

## CAPÍTULO 4: DISCUSIÓN

En el presente trabajo la concentración viral, secuenciación masiva con la tecnología Illumina y el análisis bioinformático permitieron detectar e identificar 4 familias de virus fitopatógenos en agua de riego del ramal “El Pueblo” No. 12. La concentración de partículas virales en medios acuáticos resulta ser muy baja para que puedan ser detectadas por métodos moleculares; por tanto, se necesita una concentración previa a la extracción de ácidos nucleicos. A pesar de que se conocen varios métodos para concentrar virus a partir de muestras de agua, el método de floculación orgánica usando leche descremada (SKM), propuesto por Calgua et al. (2008) para la concentración de Adenovirus y Enterovirus, es una técnica fácil y de costo reducido en comparación con técnicas que requieren equipos y conocimientos especiales. Además, se puede utilizar para otro tipo de virus, como se concluye en este trabajo y en estudios previos (Guevara, 2017).

Desde que surgieron las tecnologías de secuenciación masiva paralela (HTS), se han desarrollado una gran cantidad de flujos de trabajo para analizar datos, especialmente datos de metagenómica de bacterias y virus. La elección del mejor flujo de trabajo es algo que puede ser modificado acorde al tipo de muestra, tecnología de secuenciación y enfoque de la investigación. Dentro de los pasos establecidos para el análisis bioinformático de datos de secuenciación masiva, el pre-procesamiento es fundamental, dado que los parámetros de calidad y tamaño de las secuencias influyen en el número de lecturas a ser analizadas posteriormente. No se encontró diferencia estadísticamente significativa en el uso de MINLEN < 30 nt, como tamaño de las secuencias a ser eliminadas en el pre-procesamiento al utilizar la herramienta Trimmomatic, en comparación con el valor estándar de 50 nt; sin embargo, como mencionan Nooji et al. (2018), valores muy bajos de longitud mínima o de calidad de las lecturas pueden reducir la sensibilidad del análisis y dejar por fuera organismos no muy comunes dentro del metagenoma, por lo que, se recomienda utilizar el valor pre-establecido o valores más altos.

Por otra parte, el tipo de ensamblador si mostró diferencias significativas en las variables N50, L50 y número de contigs. En la comparación entre ensambladores

metaSPAdes y Megahit no presentaron diferencia y además, identificaron cuatro familias de virus fitopatógenos mientras que el ensamblaje con Velvet es significativamente diferente pudiendo identificarse con dicho ensamblador solo tres familias de virus. Lo mismo sucedió al utilizar genomas de referencia, donde de siete genomas usados, Velvet solo ensambló contigs de cuatro especies mientras que metaSPAdes y Megahit mostraron resultados para todos los genomas probados. En cuanto a los valores N50 y L50, se observaron resultados consistentes con el ensamblador metaSPAdes, comparables con lo reportado por Van de Walt et al. (2017) donde, de siete herramientas de ensamblaje evaluadas, metaSPAdes mostró una de las cantidades de contigs más alta y valores N50 y L50 también altos. Por otra parte utilizando los datos simulados se pudo observar la precisión de los ensambladores y así validar el flujo de trabajo. metaSPAdes y Megahit ensamblaron los 7 genomas probados en un solo contig largo lo que demuestra mayor precisión en el resultado final.

El ensamblaje es otro paso crítico en el análisis de datos de secuenciación masiva puesto que, la calidad del ensamblaje mejora la determinación de la composición de una comunidad viral o microbiana. El flujo de trabajo propuesto, junto con el ensamblador metaSPAdes, permitieron ensamblar más del 99% del genoma de los virus HLFPV y RGMoV en muestras de agua de riego. Varios flujos de trabajo han sido desarrollados para la detección y descubrimiento de virus en plantas a partir de datos de secuenciación masiva, entre ellos VirFind y VirusDetect (Rott et al., 2017). Sin embargo, resulta complejo comparar herramientas bioinformáticas y flujos de trabajo, e incluso utilizar un único flujo de trabajo ya que todo depende del tipo de muestra y el objetivo de la investigación (Van de Walt et al., 2017). Otras variables como el tiempo de ejecución de un programa, la capacidad computacional (memoria) o la disponibilidad (licencia o código abierto), también influyen al momento de decidir sobre la mejor herramienta bioinformática para un determinado análisis o pregunta de investigación y deberían ser incluidas en el mismo (Van de Walt et al., 2017).

Aunque la presencia de varios virus fitopatógenos ha sido reportada en medios acuáticos y su capacidad de transmisión a través del agua ha sido evaluada mediante ensayos de infectividad (Bütter y Koenig, 2017), no existe información disponible acerca de los virus HLFPV y RGMoV en fuentes de agua. En general, se conoce que los viriones de los géneros *Potexvirus* y *Tobamovirus*, y la familia *Tombusviridae* se replican en altas concentraciones en sus hospederos, tienen un amplio rango de

hospederos y sus partículas virales son muy estables en diferentes ambientes y temperaturas (Rosario et al. 2009). Otros virus que no pertenecen a dichos grupos también han sido encontrados en fuentes de agua y se ha determinado que las partículas de arcilla o materia orgánica presente en el agua pueden proteger a los virus frente a la degradación una vez que son adsorbidos a dichas partículas. Aunque fuera de sus huéspedes celulares y en el medio ambiente, los virus tienen el potencial de sobrevivir, persistir y ser transportados por varias rutas a otros huéspedes susceptibles, no todos los virus liberados al medio ambiente logran mantenerse infectivos (Fermin, 2018). Se requieren ensayos de infectividad para determinar la estabilidad y capacidad de infección de HLFPV y RGMoV al utilizar agua de riego del ramal “El Pueblo” No. 12.

Además de los virus ensamblados se detectó la presencia de *Hibiscus chlorotic ringspot virus* (HCRSV), *Pepper mild mottle virus* (PMMoV), *Tobacco mosaic virus* (TMV), *Tomato bushy stunt virus* (TBSV) y *Tomato mosaic virus* (ToMV), de los cuales PMMoV, TMV y TMoV han sido previamente reportados en soluciones nutritivas de sistemas hidropónicos, mientras que TBSV en agua de riego y ríos (Bütter y Koenig, 2017). La presencia de virus del dominio Duplodnaviria, que corresponde a virus de ADN, puede explicarse como contaminación al momento de la extracción de ARN total ya que el protocolo empleado puede extraer simultáneamente los dos ácidos nucleicos. Aunque, dentro del kit de extracción utilizado, se realizó un tratamiento enzimático para eliminar ADN, pudo suceder que no se eliminó por completo y que luego al momento de secuenciación la ADN polimerasa tuviera preferencia por dichas secuencias (Adewale 2020). Otra posible explicación es la presencia de virus que infectan bacterias (bacteriófagos) que al encontrarse en presencia de su hospedero dentro del agua de riego analizada, estuvieran en proceso de replicación generando ARNm que fue extraído al mismo tiempo y luego sintetizado a ADNc para secuenciación.

## CAPÍTULO 5: CONCLUSIONES

- Mediante secuenciación masiva y análisis bioinformático se identificaron 4 familias de virus fitopatógenos: *Tombusviridae*, *Virgaviridae*, *Solemoviridae* y *Alphaflexiviridae* de un total de 19 familias virales dentro de los dominios Duplodnaviria y Riboviria.
- La variación del parámetro de longitud mínima de secuencias (MINLEN-Trimomatic), no presentó diferencia significativa en las variables analizadas (número de contigs ensamblados, parámetros N50 y L50 o porcentaje de genoma ensamblado) dentro del flujo de trabajo propuesto.
- El tipo de ensamblador (Velvet, Megahit, metaSPAdes), si presentó diferencia significativa en las variables analizadas.
- Se observó que los ensambladores metaSPAdes y Megahit no son significativamente diferentes sin embargo, el ensamblador metaSPAdes presentó mayor número de contigs ( $\geq 1000$  nt) ensamblados, pudiendo considerarse la mejor herramienta de ensamblaje de las 3 comparadas.
- Mediante análisis bioinformático con el flujo de trabajo propuesto y el ensamblador metaSPAdes, se obtuvo  $>99\%$  de los genomas de *Hibiscus latent Fort Pierce virus* (HLFPV) y *Ryegrass mottle virus* (RGMoV) con  $>95\%$  de identidad por pares.

## CAPÍTULO 6: RECOMENDACIONES

- Realizar ensayos de transmisión de virus utilizando el agua de riego para determinar si las partículas virales de RGMoV y HLFPV se mantienen infectivas.
- Utilizar una cantidad mayor de muestras para una comparación más precisa entre ensambladores.
- Ampliar la zona de investigación y tomar muestras de diferentes provincias alrededor del país para determinar la incidencia de RGMoV y HLFPV en agua de riego.

## CAPÍTULO 7: BIBLIOGRAFÍA

- Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years?. *African journal of laboratory medicine*, 9(1), 1340. <https://doi.org/10.4102/ajlm.v9i1.1340>
- Adriaenssens, E. M., Farkas, K., Harrison, C., Jones, D. L., Allison, H. E., & McCarthy, A. J. (2018). Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems*, 3(3), e00025-18. <https://doi.org/10.1128/mSystems.0>
- Agrios, G. N. (2005). Plant diseases caused by viruses. En *Plant Pathology* (5 ed., págs. 723-824). San Diego, California: Elsevier Academic Press.
- Büttner, C., & Koenig, R. (2017). CHAPTER 10: Plant Viruses in Irrigation Water. In *Biology, Detection, and Management of Plant Pathogens in Irrigation Water* (pp. 97–110). The American Phytopathological Society. <https://doi.org/10.1094/9780890544914.012>
- Cai, L., Yang, Y., Jiao, N., & Zhang, R. (2015). Evaluation of Tangential Flow Filtration for the Concentration and Separation of Bacteria and Viruses in Contrasting Marine Environments. *PLoS*, 10(8), 12. <https://doi.org/10.1371/journal.pone.0136741>
- Calgua, B., Fumian, T., Rusiñol, M., Rodriguez-Manzano, J., Mbayed, V. A., Bofill-Mas, S., . . . Girones, R. (2013). Detection and quantification of classic and emerging viruses by skimmed-milk flocculation and PCR in river water from two geographical areas. *Water Research*, 47, 2797-2810. <https://doi.org/10.1016/j.watres.2013.02.043>
- Calgua, B., Mengewein, A., Grunert, A., Bofill-Mas, S., Clemente-Casares, P., Hundesa, A., Wyn-Jones, A. P., López-Pila, J. M., & Girones, R. (2008). Development and application of a one-step low cost procedure to concentrate viruses from seawater samples. *Journal of Virological Methods*, 153(2), 79–83. <https://doi.org/10.1016/j.jviromet.2008.08.003>
- Cann, A. J. (2005). *Principios de Virología Molecular*. Zaragoza, España: Elsevier Inc.
- Carrico, J. A., Rossi, M., Moran-Gilad, J., Van Domselaar, G., & Ramirez, M. (2018). A primer on microbial bioinformatics for nonbioinformaticians. *Clinical Microbiology and Infection*, 24, 342-349. <https://doi.org/10.1016/j.cmi.2017.12.015>
- Cashdollar, J., & Dahling, D. R. (2006). Evaluation of a method to re-use electropositive cartridge filters for. *Journal of Virological Methods*, 132, 13-17. <https://doi.org/10.1016/j.jviromet.2005.08.016>
- Chang, P.-G. S. (2009). Plant Virus Diagnostics: Comparison of classical and membrane-based techniques for immunoassay and coat protein sequence characterization for Cucumber mosaic virus and three potyviruses. Faculty of the Virginia Polytechnic Institute and State University. <https://vtechworks.lib.vt.edu/handle/10919/28017>
- Fermin, G. (2018). Host Range, Host–Virus Interactions, and Virus Transmission. In Tennant P, Fermin G and Foster JE (eds), *Viruses*. Academic Press 101-134. <https://doi.org/10.1016/B978-0-12-811257-1.00005-X>.

- Gao, R., Niu, S., Dai, W., Kitajima, E., & Wong, S. M. (2016). Hibiscus latent Fort Pierce virus in Brazil and synthesis of its biologically active full-length cDNA clone. *Virus genes*, 52(5), 754–757. <https://doi.org/10.1007/s11262-016-1344-8>
- Gergerich, R., & Dolja, V. (2006). *Introduction to Plant Viruses, the Invisible Foe*. (T. A. Society, Ed.). <https://doi.org/10.1094/PHI-I-2006-0414-01>
- Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpan, P., Schatz, M. C., & McCombie, R. W. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25(11), 1750–1756. <https://doi.org/10.1101/gr.191395.115>
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of nextgeneration sequencing technologies. *Nature Reviews*(17), 333-351. <https://doi.org/10.1038/nrg.2016.49>
- Guerrero-Latorre, L., Romero, B., Bonifaz, E., Timoneda, N., Rusiñol, M., Girones, R., & Rios-Touma, B. (2018). Quito's virome: Metagenomic analysis of viral diversity in urban streams. *Science of the Total Environment*, 645, 1334-1343. <https://doi.org/10.1016/j.scitotenv.2018.07.213>
- Guevara, F. (2017). Detección e identificación de Potexvirus, Tombusvirus y Tobamovirus en aguas de riego de babaco (*Vasconcellea x heilbornii*) en la provincia de Pichincha. Tesis de pregrado. Universidad de las Fuerzas Armadas ESPE.
- Guevara, F. & Flores, F. (2021). Near-complete genome sequence of ryegrass mottle virus from irrigation water in Ecuador. *Microbiol Resour Announc* 10:e00037-21. <https://doi.org/10.1128/MRA.00037-21>
- Haramoto, E., Kitajima, M., Hata, A., Torreyd, J. R., Masagoe, Y., Sano, D., & Katayama, H. (2018). A review on recent progress in the detection methods and prevalence of human enteric viruses in water. *Water Research*, 135, 168-186. <https://doi.org/10.1016/j.watres.2018.02.004>
- Hellmér, M., Paxéus, N., Magnius, L., Enache, L., Arnholm, B., Johansson, A., & Bergstrom, T. &. (2014). Detection of Pathogenic Viruses in Sewage Provided Early Warnings of Hepatitis A Virus and Norovirus Outbreaks. *Applied and Environmental Microbiology*, 80(21), 6771-6781. <https://doi.org/10.1128/AEM.01981-14>
- Hjelmsø, M. H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenka, O., Seidel, M., . . . Schultz, A. C. (2017). Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. *PLoS ONE*, 12(1), e0170199. <https://doi.org/10.1371/journal.pone.0170199>
- Hjelmsø, M. H., Mollerup, S., Jensen, R. H., Pietroni, C., Lukjancenka, O., Schultz, A. C., . . . Hansen, A. J. (2019). Metagenomic analysis of viruses in toilet waste from long distance flights—A new procedure for global infectious disease surveillance. *PLoS ONE*, 14(1). <https://doi.org/10.1371/journal.pone.0210368>
- Hull, R. (2009). *Comparative plant virology* (2 ed.). China: Elsevier Academic Press.
- INIAP. (2014). Instituto Nacional de Investigaciones Agropecuarias-INIAP. Granja Experimental Tumbaco. <http://tecnologia.iniap.gob.ec/index.php/estaciones/meecat/mgraecat>

- Jeżewska , M., Trzmiel , K., & Zarzyńska-Nowak, A. (2019). Detection of infectious Brome mosaic virus in irrigation ditches and draining strands in Poland. *European Journal of Plant Pathology*, 153, 285–292. <https://doi.org/10.1007/s10658-018-1531-7>
- Jones, R. (2018). Plant and Insect Viruses in Managed and Natural Environments: Novel and Neglected Transmission Pathways. In Malmstrom CM (ed), *Environmental Virology and Virus Ecology. Advances in Virus Research*, Academic Press 101:149-187. <https://doi.org/10.1016/bs.aivir.2018.02.006>.
- Jooste, T. L., Visser, M., Cook, G., Burger, J. T., & Maree, H. J. (2017). In Silico Probe-Based Detection of Citrus Viruses in NGS Data. *Phytopathology®*, 107(8), 988–993. <https://doi.org/10.1094/PHYTO-10-16-0379-R>
- Kahler, A. M., Johnson, T. B., Hahn, D., Narayanan, J., Derado, G., & Hill, V. R. (2015). Evaluation of an Ultrafiltration-Based Procedure for Simultaneous Recovery of Diverse Microbes in Source Waters. *Water*, 7(3), 1202-1216. <https://doi.org/10.3390/w7031202>
- Kamenova, I. & Adkins, S. (2004). Comparison of Detection Methods for a Novel Tobamovirus Isolated from Florida Hibiscus. *Plant Disease* 88:34-40
- King, A. M., Adams, M. J., Carstens, E. B., & Lefkowitz, E. J. (2012). Virus Taxonomy, Ninth report of the International Committee on Taxonomy of Viruses. En Virus. Londres: Elsevier.
- Kukurba , K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 11, 951–969. <https://doi.org/10.1101/pdb.top084970>
- Kumar, R., Manchev, N., Patil, B., & Ivanona, M. (2016). *Plant Viruses: Evolution and Management*. Singapore: Springer Nature.
- Kuroda, K., Nakada, N., Hanamoto, S., Inaba, M., Katayama, H., Do, A. T., . . . Takizawa, S. (2015). Pepper mild mottle virus as an indicator and a tracer of fecal pollution in water environments: Comparative evaluation with wastewater-tracer pharmaceuticals in Hanoi, Vietnam. *Science of The Total Environment*, 506-507(15), 287-298. <https://doi.org/10.1016/j.scitotenv.2014.11.021>
- Li, J. W. (1998). A new and simple method for concentration of enteric viruses from water. *Journal of Virological Methods*, 74, 99-108.
- Ling, R., Pate, A.E., Carr, J.P. and Firth, A.E. (2013). An essential fifth coding ORF in the sobemoviruses. *Virology*, 446(1-2):397-408. <https://doi.org/10.1016/j.virol.2013.05.033>
- Liu, L., Li, S., Hu, N., He, Y., Pong, R., Lin, D., . . . Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012. <https://doi.org/10.1155/2012/251364>
- Liu, P., Hill, V. R., Hahm, D., Johnson, T. B., Pan, Y., Jothikumar, N., & Moe, C. L. (2012). Hollow-fiber ultrafiltration for simultaneous recovery of viruses, bacteria and parasites from reclaimed water. *Journal of Microbiological Methods*, 88, 155-161. <https://doi.org/10.1016/j.mimet.2011.11.007>

- Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, 3, 31. <https://doi.org/10.1186/s40168-015-0094-5>
- Mehle, N., & Ravnikar, M. (2012). Plant Viruses in aqueous environment - Survival, mediated transmission and detection. *ELSEVIER Review*, 4903-4907.
- Mehle, N., Gutiérrez-Aguirre, I., Prezelj, N., Delic, D., Vidic, U., & Ravnikar, M. (2014). Survival and Transmission of Potato Virus Y, Pepino Mosaic Virus and Potato Spindle Tuber Viroid in Water. *Applied and Environmental Microbiology*, 80(4), 1455-1462. <http://dx.doi.org/10.1128>
- Millen, H. T., Gonnering, J. C., Berg, R. K., Spencer, S. K., Jokela, W. E., Pearce, J. M., . . . Borchardt, M. A. (2012). Glass Wool Filters for Concentrating Waterborne Viruses and Agricultural Zoonotic Pathogens. *Journal of Visualized Experiments*, 61. <https://doi.org/10.3791/3930>
- Nieuwenhuijse, D. F., & Koopmans, M. P. (2017). Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases. *Frontiers in Microbiology*, 8, 230. <https://doi.org/10.3389/fmicb.2017.00230>
- Ochoa Corona, F. M., Fletcher, J., & Olson, J. (2014). Final Report Research Suggestions for USDA-APHIS FY12 Farm Bill section 10201. Unpublished, 18.
- Rastrojo, A., & Alcamí, A. (2017). Aquatic viral metagenomics: Lights and shadows. *Virus Research*, 239, 87–96. <https://doi.org/10.1016/j.virusres.2016.11.021>
- Reuter, J. A., Spacek, D., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular cell*, 58(4), 586-597. <https://doi.org/10.1016/j.molcel.2015.05.004>
- Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics*, 13(5), 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>
- Rosario, K., Nilsson, C., Lim, Y.W., Ruan, Y. and Breitbart, M. (2009). Metagenomic analysis of viruses in reclaimed water. *Environmental Microbiology*, 11:2806-2820. <https://doi.org/10.1111/j.1462-2920.2009.01964.x>
- Rott, M., Xiang, Y., Boyes, I., Belton, M., Saeed, H., Kesanakurti, P., Hayes, S., Lawrence, T., Birch, C., Bhagwat, B., & Rast, H. (2017). Application of Next Generation Sequencing for Diagnostic Testing of Tree Fruit Viruses and Viroids. *Plant Disease*, 101(8), 1489–1499. <https://doi.org/10.1094/PDIS-03-17-0306-RE>
- Ryu, K. H., & Hong, J. S. (2008). Potexvirus. En B. M. Mahy, & M. H. van Regenmortel (Edits.), *Encyclopedia of Virology* (3 ed., Vol. 3, págs. 310-313). Elsevier.
- Sekse, C., Holst-Jensen, A., Dobrindt, U., Johannessen, G. S., Li, W., Spilsberg, B., & Shi, J. (2017). High Throughput Sequencing for Detection of Foodborne Pathogens. *Frontiers in Microbiology*, 8, 2029. <https://doi.org/10.3389/fmicb.2017.02029>
- Sevik, M. A. (2011). Water Pollution: Water-Borne Plant Viruses. *Cilt*, 27.
- Short, S. M., & Short, C. M. (2008). Diversity of algal viruses in various North American freshwater environments. *Aquatic microbial ecology*, 51(1), 13-21. <https://doi.org/10.3354/Ame01183>

- Sõmera, M.; Sarmiento, C.; Truve, E. (2015). Overview on Sobemoviruses and a Proposal for the Creation of the Family Sobemoviridae. *Viruses*, 7:3076-3115. <https://doi.org/10.3390/v7062761>
- van der Walt, A., van Goethem, M., Ramond, JB. et al. (2017). Assembling metagenomes, one community at a time. *BMC Genomics*, 18:521. <https://doi.org/10.1186/s12864-017-3918-9>