Diagnóstico de *Grapevine virus* A y *Grapevine virus* B mediante el uso del dispositivo MinION - Oxford Nanopore Sequencing

Ramos López, Álvaro Daniel

Departamento de Ciencias de la Vida y de la Agricultura

Carrera de Ingeniería en Biotecnología

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

Flores Flor, Francisco Javier, Ph.D.

4 de marzo de 2021

URKUND

Document Information

Analyzed document	TesisARamos_paraUrkund.txt (D97145689)
Submitted	3/3/2021 11:28:00 PM
Submitted by	
Submitter email	fjflores2@espe.edu.ec
Similarity	0%
Analysis address	fjflores2.espe@analysis.urkund.com

Sources included in the report

Firma

Firmado electrónicamente por:
FRANCISCO JAVIER FLORES FLOR

Flores Flor, Francisco Javier

DIRECTOR

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCIAS DE LA VIDA Y DE LA AGRICULTURA

CARRERA DE INGENIERÍA EN BIOTECNOLOGÍA

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"Diagnóstico de Grapevine virus A y Grapevine virus B mediante el uso del dispositivo MinION - Oxford Nanopore Sequencing"** fue realizado por el señor **Ramos López, Álvaro Daniel** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 4 de marzo de 2021

Firma:

Firmado electrónicamente por:
FRANCISCO JAVIER FLORES FLOR

Flores Flor, Francisco Javier, Ph.D.

C.C. 1713443479

DEPARTAMENTO DE CIENCIAS DE LA VIDA Y DE LA AGRICULTURA

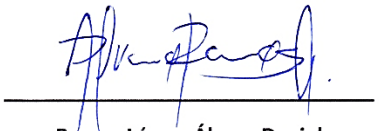
CARRERA DE INGENIERÍA EN BIOTECNOLOGÍA

RESPONSABILIDAD DE AUTORÍA

Yo, **Ramos López, Álvaro Daniel**, con cédula de ciudadanía n◦ 1725039828, declaro que el contenido, ideas y criterios del trabajo de titulación: "**Diagnóstico de Grapevine virus A y Grapevine virus B mediante el uso del dispositivo MinION - Oxford Nanopore Sequencing"** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 4 de marzo de 2021

Firma:



Ramos López, Álvaro Daniel

C.C.: 1725039828

DEPARTAMENTO DE CIENCIAS DE LA VIDA Y DE LA AGRICULTURA

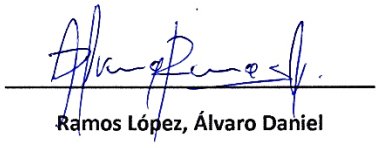
CARRERA DE INGENIERÍA EN BIOTECNOLOGÍA

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Ramos López, Álvaro Daniel**, con cédula de ciudadanía n° 1725039828, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Diagnóstico de *Grapevine virus* A y *Grapevine virus* B mediante el uso del dispositivo MinION - Oxford Nanopore Sequencing"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 4 de marzo de 2021

Firma:



Ramos López, Álvaro Daniel

C.C.: 1725039828

Dedicatoria

Dedicado a mi madre Diyanira Coralia López Sengés, mi padre Allan Arturo Ramos Pilco y mi hermano Allan David Ramos López, sin su apoyo en mi formación académica y personal nada de lo que he hecho o lo que soy podría ser posible.

Daniel

Agradecimiento

A mi tutor, profesor y guía, el Doctor Francisco Flores por toda la instrucción, las oportunidades y el apoyo que me ha dado a lo largo de mis estudios universitarios. Gracias por su actitud como docente y por la confianza que siempre ha depositado en mí, que ha hecho posible llegar a donde estoy en mi vida profesional.

A mi tutor en Oklahoma State University (OSU), institución colaboradora en esta investigación, el Doctor Andrés Espíndola, por toda la asistencia y guía que me proporcionó a lo largo del desarrollo de este trabajo y sobretodo la oportunidad de poder colaborar con él y la OSU en nombre de la Universidad de las Fuerzas Armadas –ESPE y del país.

A la Universidad de las Fuerzas Armadas – ESPE y a la Carrera de Ingeniería en Biotecnología, por acogerme durante toda mi vida universitaria, proporcionándome todos los medios posibles para mi progreso como profesional.

A mi madre por siempre estar conmigo y darme esa fuerza para siempre seguir adelante, a mi padre por el apoyo que me ha brindado en mi formación y a mi hermano por todo lo que he aprendido gracias a él y por la motivación que tengo gracias a su determinación.

Índice de contenidos

Índice de tablas

Índice de figuras

Resumen

Los viñedos son uno de los cultivos de mayor producción en el mundo debido a sus propiedades alimenticias e industriales. Estos cultivos pueden hospedar patógenos virales cuya infección produce una reducción de la calidad de los frutos, causando pérdidas de hasta $ 47 000 por hectárea anualmente. La mejor forma de controlar estos agentes es la producción y distribución de plantas libres de virus mediante la selección de ejemplares sanos para su reproducción. Los métodos tradicionales de detección no permiten diagnosticar infecciones en estados tempranos, por lo que se ha propuesto la identificación por medio de secuenciación con MinION por su rendimiento y sensibilidad en condiciones de campo. El presente estudio propone un flujo de procesos *in silico*, basado en las herramientas E-probe Diagnostic Nucleic acid Analysis y NanoSim*,* orientado a la validación hasta la fase de reconocimiento provisional en la utilización de e-probes como prueba para la detección de varias cepas de *Grapevine virus* A y de *Grapevine virus* B en diversas zonas geográficas, en el que la simulación de muestras metagenómicas y posterior análisis con e-probes permitió determinar los parámetros de desempeño de la prueba para su posterior implementación en la producción de plantas libres de estos virus.

PALABRAS CLAVE:

- **METAGENÓMICA**
- **SIMULACIÓN DE NANOPORE**
- **VALIDACIÓN DE PRUEBA DIAGNÓSTICA**

Abstract

Vineyards are one of the highest production crops in the world due to their alimentary and industrial properties. These crops can host viral pathogens whose infection causes a reduction in fruit quality, reflecting losses of up to $ 47,000 per hectare annually. The best way to control these agents is the production and distribution of virus-free plants by selecting healthy specimens for their reproduction. Traditional detection methods do not allow diagnosing infections in early stages, therefore identification through MinION sequencing has been proposed due to its performance and sensitivity under field conditions. The present study proposes an *in silico* pipeline, based on the E-probe Diagnostic Nucleic acid Analysis and NanoSim tools, aimed at validation up to the provisional recognition phase in the use of e-probes as a test for the detection of various strains of *Grapevine virus* A and *Grapevine virus* B in various geographical areas, in which the simulation of metagenomic samples and subsequent analysis with e-probes allowed to determine the performance parameters of the test for its subsequent implementation in the production and distribution of plants free of these viruses.

KEY WORDS:

- **METAGENOMICS**
- **NANOPORE SIMULATIONS**
- **VALIDATION OF DIAGNOSTIC TEST**

Introducción

Formulación del problema

Los viñedos son de los cultivos frutales con mayor importancia económica debido a la cantidad de ingresos generados por la comercialización de uvas y la producción de vinos desde hace miles de años (This et al., 2006). Estos cultivos a lo largo de los años han mostrado que pueden hospedar muchas entidades patogénicas que incluyen 65 virus, cinco viroides y ocho fitoplasmas, cuyas interacciones con las plantas generan enfermedades que alteran sus reacciones al ambiente y viceversa (Martelli, 2014).

En términos de relevancia económica, de los 65 virus que pueden hospedar los viñedos cerca de la mitad son reconocidos como importantes agentes causantes de enfermedades como: infección degenerativa, *Grapevine fanleaf virus*, *Arabis mosaic virus* y otros miembros del género *Nepovirus*; enfermedad de enrollamiento de hojas, virus que pertenecen a los géneros *Ampelovirus*, *Closterovirus*, *Velavirus* y *Vitivirus* (Martelli, 2014). La forma de propagación convencional de la vid es mediante reproducción vegetativa, por lo que el impacto de la presencia de una enfermedad viral es amplio si se generan plantas clones a partir de plantas madres ya infectadas (Naidu et al., 2015).

Las metodologías que se han empleado durante los últimos 20 años para controlar las infecciones virales han permitido obtener plantas certificadas para la producción. Se ha utilizado termoterapias *in vitro* o *in vivo*, cultivo de meristemos, quimioterapia y crioterapia, estrategias cuyos resultados varían dependiendo de la especie (Maliogka et al., 2015). Aunque la embriogénesis somática para la producción de plantas libres de agentes virales ha sido efectiva

para un gran número de virus, no es una técnica utilizada ampliamente debido al riesgo de variaciones somaclonales (Gambino et al., 2011).

Los impactos principales de las enfermedades causadas por infecciones virales en viñedos son la reducción de la calidad y producción de los frutos, disminución del vigor de la planta, pérdida de pigmentos, proteínas solubles y actividad fotosintética (FAO, 2020; Fuller et al., 2019; Raman & Muthukathan, 2015). Las pérdidas globales financieras causadas por estos virus llegan a ser de millones de dólares anualmente, tal es el caso que se han estimado pérdidas que van desde los $25 000 hasta los $40 000 por hectárea durante una vida útil de 25 años de un viñedo "Cabernet franc" (Atallah et al., 2012) y hasta $47 000 por hectárea a lo largo de 20 años de vida útil de cultivos de “Sauvignon blanc” y “Merlot” (Ricketts et al., 2015).

Debido al desarrollo de tecnologías de secuenciación de alto rendimiento, se ha dado enfoque al diagnóstico de las enfermedades de los viñedos mediante la identificación de varios virus a la vez, lo que permite el desarrollo de trabajos de mejoramiento genético orientado a la detección y eliminación de todo agente patógeno presente en los fitobiomas a los que pertenecen estas plantas (Dicke, 2016).

Justificación del problema

La sintomatología de las enfermedades causadas por patógenos virales suele deberse a la interacción de varios virus a la vez, por lo que no es posible identificar a ninguna variante de *Grapevine virus* por inspección simple, además que los síntomas suelen aparecer luego de 2 años de haberse comenzado la infección. Los métodos convencionales para la detección son pruebas serológicas viables solo en infecciones tardías en plantas maduras, mientras que la detección temprana de la infección por *Grapevine virus* es posible mediante el uso de técnicas moleculares (Meng et al., 1999).

Para evitar la propagación de los virus, la mejor estrategia es la producción y distribución de plantas libres de virus, dado que no se han reportado ensayos a nivel de campo para el control químico o biológico de vectores asociados específicamente a estos patógenos (Arora et al., 2020; CABI, 2020; Crnogorac et al., 2021), ni plantas con resistencia natural (Martinelli et al., 2002). Existen esquemas de certificación para el comercio de plantas sanas, evitando así la propagación de los virus asegurando cultivos que puedan dar productividades óptimas para reducir pérdidas, esto requiere que los métodos de detección puedan identificar posibles infecciones virales lo más temprano posible además de ser ensayos rápidos y de bajo costo (Maliogka et al., 2015).

El dispositivo MinION ha sido utilizado para la secuenciación en tiempo real de cadenas largas de ARN de sentido positivo, permitiendo la detección de los virus: *Wheat streak mosaic virus* (Fellers et al., 2019), *Yam mild mosaic virus*, *Yam chlorotic necrosis virus* (Filloux et al., 2018) y *Cucumber green mottle mosaic virus*, *Tomato brown rugose fruit virus* y *Zucchini yellow mosaic virus* (Chalupowicz et al., 2019).

Las ventajas del uso del dispositivo MinION para el diagnóstico de infecciones virales son la rápida implementabilidad y confiabilidad bajo condiciones de campo, la capacidad de lecturas largas para el ensamblaje de genomas *de novo*, alta fidelidad en regiones repetitivas, menos costos operacionales para la secuenciación respecto a tecnologías de segunda generación como Illumina (Zhang et al., 2020) y la obtención de resultados en menos de 24 horas una vez obtenida la muestra (Lu et al., 2016).

Los datos producto de secuenciación con MinION han generado desafíos informáticos como almacenamiento, transmisión, manipulación y análisis de la información, por lo que se han estado desarrollando métodos computacionales para aumentar el desempeño de los equipos, sin embargo, es esencial que estas nuevas implementaciones puedan ser evaluadas referencialmente con herramientas existentes para probar su rendimiento para lo cual se utiliza por lo general datos empíricos o simulados debido a su facilidad de manejo y reproducibilidad (Escalona et al., 2016).

Los datos empíricos, si bien representan escenarios reales, no pueden emular los procedimientos reales que se requerirían para obtener dichos escenarios, de manera alternativa, las simulaciones *in silico* permiten la generación de datos de manera indefinida, bajo parámetros controlados y procesos conocidos que pueden, en conjunto con datos empíricos, dar buenas aproximaciones para valores reales (Angly et al., 2012).

Objetivos de investigación

Objetivo General

Proponer una metodología de detección de *Grapevine virus* A y *Grapevine virus* B mediante el uso del dispositivo MinION - Oxford Nanopore Sequencing por medio de e-probes obtenidas por simulaciones *in silico* de lecturas metagenómicas.

Objetivos Específicos

- Generar datos de entrenamiento para la simulación *in silico* de lecturas de Oxford Nanopore Sequencing para *Grapevine virus* A y *Grapevine virus* B mediante la caracterización de modelos de perfiles de lectura y ajuste de errores utilizando el paquete NanoSim.
- Efectuar simulaciones de muestras metagenómicas que puedan contener copias de *Grapevine virus* A y *Grapevine virus* B como patógenos en el huésped *Vitis vinifera* a ser detectadas mediante la plataforma MinION - Oxford Nanopore Sequencing.
- Diseñar e-probes como sondas para la detección de *Grapevine virus* A y *Grapevine virus* B mediante el uso de la plataforma MiFi: Microbe Finder bajo parámetros de sensibilidad ajustados a muestras metagenómicas.

Marco teórico

Género *Vitis*

Los miembros del género *Vitis* son de los cultivos frutales más valiosos del mundo, son ampliamente utilizados para vino, uvas de mesa, pasas, jugos y licores, recientemente la producción también se ha centrado en antioxidantes y productos saludables derivados de las uvas. *Vitis vinifera* L. subsp. *vinifera* es la especie de uva más cultivada, pero su productividad es limitada debido a su susceptibilidad a plagas, enfermedades y estrés abiótico como el frío (Reisch & Pratt, 1996).

Los centros de diversidad de la vid se encuentran en el sureste de los Estados Unidos y Asia Oriental. Hasta 30 especies son nativas de una vasta área en el este de Asia, China, Japón y Java, dos especies en el centro de Asia y Europa, y hasta 28 especies en el este y suroeste de los Estados Unidos y México (Comeaux, 2013). El género *Vitis* se divide en dos subgéneros: *Muscadinia* planch. (2n = 40, una o dos especies) y *Vitis* planch. (2n = 38, las especies restantes). Las divisiones adicionales dentro de *Vitis* son "series" que son agrupaciones subgenéricas que se han usado históricamente en la sistemática de *Vitis* (Moore, 1991).

El vino como bebida mundialmente reconocida, se incorporó profundamente en la tradición de muchas naciones y la forma en que se consume a menudo refleja la cultura y estilo de vida de ciertas comunidades. Además de los beneficios para la salud, el vino a menudo presenta un signo de prestigio e influencia en religiones, hábitos, y el nivel de vida de los consumidores (Vlahovic et al., 2012).

La producción de uva está muy extendida en el mundo, especialmente en Europa, en el mercado mundial del vino está dominado por tres países: Francia, Italia y España, cuya producción en conjunto representa la mitad de la producción mundial total. Además de estos tres países europeos, los productores mundiales de uva son China, Estados Unidos y Chile. Los mayores exportadores de vino del mundo son Francia, Italia, España, Australia y Chile (FAO, 2020).

Los ingresos generados a nivel mundial debido al mercado de la uva ascendieron a $136.6 billones en 2018 con una producción de 76 millones de toneladas. Los mercados de exportación más grandes, en términos de exportación, fueron Chile con $1.2 billones, Estados Unidos con $925 millones y Perú con $820 millones que en conjunto suman el 36% de exportaciones mundiales, mientras que en importación se tiene a Estados Unidos con $1 billón, Alemania con $750 millones y China con $743 millones que conforman el 30% de importaciones totales. Se prevé que para el final de 2025, siguiendo una tasa de crecimiento anual compuesta de +1.3%, haya una producción de 83 millones de toneladas (IndexBox, 2020).

Vitivirus

Es un grupo de variantes de virus pertenecientes a al grupo IV – virus monocatenario de ARN en sentido positivo (+ssARN), no poseen envoltura, están construidos helicoidalmente a manera de partículas filamentosas flexibles, que miden aproximadamente 800 x 12 nm, con un diámetro de 3.3-3.5 nm y aproximadamente 10 subunidades por cada vuelta de la hélice (Milne et al., 1984). Las partículas están formadas por 95% de proteína y 5% de ácido nucleico en peso y como sedimento está conformada por un único componente de 92 S de coeficiente de sedimentación (Monette, PL; James, 1990).

La proteína de la cubierta es un polipéptido de 21.5 kDa, cada partícula contiene una molécula de ARN monocatenario en sentido positivo de 7.6 kb aproximadamente. El ARN genómico contiene cinco marcos de lectura abierta que codifican proteínas relacionadas con la replicación (ORF 1), un producto de 19 kDa con funciones no conocidas (ORF 2), una proteína de movimiento (ORF 3), una proteína de la cubierta (ORF 4) y un producto de 10 kDa (ORF 5) con propiedades de unión a nucleótidos que se asocia a funciones de supresión de silenciamiento de genes (Galiakparov et al., 2003). La mayoría de especies están relacionadas serológicamente de forma distante, se han producido anticuerpos monoclonales para las especies *Grapevine virus* A (GVA), *Grapevine virus* B (GVB) y *Grapevine virus* D (GVD) además de proteínas recombinantes como anticuerpos dirigidos a la proteína de movimiento putativa de GVA (ICTV, 2020).

Criterio de Demarcación

Se refiere al criterio con el que se puede determinar que una especie pertenece a cierto género al ser comparada con otras dentro del mismo, este es valorado bajo parámetros cualitativos y cuantitativos. Para el género *Vitivirus* en general se tiene:

- El rango de huéspedes naturales
- Especificidad serológica determinada por el uso de anticuerpos monoclonales o policlonales discriminatorios.
- Epidemiología: se toma en cuenta el tipo y especies de vectores que pueden transmitir una sola especie o un grupo de especies dentro del grupo.
- Diferencias en el patrón de ARN de doble cadena

- 72% de identidad en la cadena de nucleótidos (80% de identidad en cadena de aminoácidos traducida) entre su proteína de la cubierta o genes asociados a la polimerasa (ICTV, 2020).

Distribución

La amplia distribución geográfica posiblemente se deba a la diseminación internacional involuntaria del virus en el germoplasma de los cultivos de viñedos infectados no identificados a tiempo para su descarte (Martelli, 2014).

Se lo ha registrado en África (Algeria, Egipto, Marruecos, Sudáfrica y Túnez), Asia (Afganistán, Armenia, China, Irán, Israel, Jordania, Kazajistán, Líbano, Siria, Turquía y Yemen), Europa (Albania, Croacia, Chipre, República Checa, Francia, Alemania, Grecia, Hungría, Italia, Malta, Macedonia del Norte, Portugal, Rusia, Eslovaquia, Eslovenia, España, Suiza, Ucrania y el Reino Unido), Australia y América (Brasil, Chile, Canadá y Estados Unidos) (CABI, 2020).

Transmisión

La única fuente conocida de los virus para su transmisión es la presente en plantas ya infectadas, los patógenos se transmiten a plantas sanas por medio de vectores como cochinillas o por injerto. El virus está restringido al floema en el que se multiplica en las células del parénquima y los tubos de criba diferenciadores en los que se acumulan partículas (Rosciglione et al., 1983) y son adquiridas por los vectores (La Notte et al., 1997). El virus se transmite de una manera no persistente semi-circulatoria por las cochinillas pseudococcídeas *Pseudococcus longispinus*, *Pseudococcus affinis* [P. viburni], *Planococcus citri* y *Planococcus ficus* (Garau et al., 1995; Tsai et al., 2008).

Síntomas de Infección

Inducen en cultivares de *Vitis vinifera* infectados síntomas descritos como el complejo de madera rugosa que comprende síndromes de: picaduras de tallo, ranurado del tallo de Kober (KSG por sus siglas en inglés), taponamiento en la corteza (CB) y ranurado del tallo del tipo LN33 (Bonavia et al., 1996).

Los síndromes del complejo de madera rugosa se causan por agentes de especies pertenecientes a los géneros *Vitivirus* y *Foveavirus*, en particular las infecciones con GVA producen KSG y las causadas con GVB provocan CB (Meng et al., 2017). Informes de Sudáfrica y Australia indican que también pueden estar involucrados en la enfermedad de Shiraz de otros cultivos de viñedos como *Vitis californica* y *Vitis rupestris* (Klaassen et al., 2011). Las vides individuales pueden infectarse simultáneamente por más de una variante (Goszczynski & Jooste, 2003).

Diagnóstico

Los métodos serológicos incluyen microscopía electrónica inmunosorbente de barrido (ISEM) (Agran et al., 1990), ensayo inmunosorbente asociado a enzimas en sándwich de doble anticuerpo (DAS-ELSA) con antisuero policlonal y anticuerpos monoclonales (M'hirsi et al., 2001) y Western blotting (Borgo et al., 2006).

Las técnicas moleculares comprenden la hibridación del ácido nucleico (NASH) (Saldarelli et al., 1994) y amplificación cuantitativa del ácido nucleico por medio de retrotranscripción seguida de reacción en cadena de la polimerasa (RT-qPCR) que permite la detección del virus en savia diluida en una proporción de 1:10000 con un paso previo de inmunocaptura (Chevalier et al., 1995). Se conocen varias cepas de GVA que difieren biológica y

molecularmente (Monette & James, 1990), estas variantes moleculares se condensan en tres grupos que pueden identificarse selectivamente por medio de RT-qPCR utilizando primers específicos para las variantes.

Oxford Nanopore Technologies – Secuenciador MinION

La demanda de tecnologías que puedan operar a mayor velocidad y producir lecturas más largas ha dado como resultado la llegada de nuevos enfoques de secuenciación: la llamada secuenciación de tercera generación (TGS). Las plataformas principales de secuenciación de segunda generación (SGS) adaptan las tecnologías de secuenciación por síntesis (SBS) que dependen de la PCR para ampliar los grupos de una plantilla de ADN dada (Lander et al., 2001).

Las tecnologías TGS, por el contrario, se dirigen directamente a moléculas de ADN individuales, lo que permite la secuenciación en tiempo real, donde las lecturas están disponibles para el análisis tan pronto como hayan pasado por el secuenciador (Ashton et al., 2015). Hay tres mejoras importantes en las plataformas TGS: (1) aumento en la longitud de lectura de decenas de bases a decenas de miles de bases por lectura; (2) reducción del tiempo de secuencia de días a horas (o minutos para aplicaciones en tiempo real); y (3) reducción o eliminación de sesgos de secuenciación introducidos por amplificación por PCR (Schadt et al., 2010).

En 2014, Oxford Nanopore Technologies (ONT) lanzó una nueva plataforma TGS, el dispositivo MinION, a través de un programa de acceso temprano (The MinION Access Program, MAP). Cada celda de flujo consumible puede generar hasta 30 Gb de datos de secuencia de ADN o 7-12 millones de lecturas si se analiza ARN. Son posibles longitudes de lectura largas (cientos de kb) ya que se puede elegir la longitud del fragmento para realizar la lectura. El MinION

transmite datos en tiempo real para que el análisis se pueda realizar durante el experimento y los flujos de trabajo sean totalmente versátiles. Debido a su pequeño tamaño y bajo costo de equipo, el secuenciador MinION está atrayendo un considerable interés en la comunidad genómica, particularmente para la vigilancia de patógenos y las aplicaciones de diagnóstico clínico, ya que estas áreas se beneficiarían de la naturaleza en tiempo real de esta plataforma de secuenciación (Oxford Nanopore Technologies, 2020).

MinION identifica las bases de ADN midiendo los cambios en la conductividad eléctrica generados a medida que las cadenas de ADN pasan a través de un nanoporo biológico. Su portabilidad, accesibilidad y velocidad en la producción de datos lo hace adecuado para aplicaciones en tiempo real. Mientras que los ensamblajes de genomas *de novo* se pueden producir a bajo costo a partir de datos de SGS, la continuidad del ensamblaje a menudo es relativamente pobre, debido a la capacidad limitada de lecturas cortas para manejar repeticiones largas, MinION permite lecturas más largas que facilitan el procesamiento de los algoritmos de ensamblaje que a su vez mejora el análisis de regiones con repeticiones (Lu et al., 2016).

Basecalling

La forma en la que MinION y otros equipos (como GridION y PromethION) con la tecnología de nanoporos procesan las señales registradas por los cambios de conductividad a lecturas de nucleótidos se conoce como Basecalling, para este fin se diseñan algoritmos capaces de dicho proceso de traducción de señales crudas a secuencias llamados Basecallers, cuyo desarrollo se puede dividir de manera general en dos etapas, debido a los avances tanto computacionales como de diseño de los nanoporos (Zeng et al., 2020).

La primera que adopta Modelos Ocultos de Markov (Hidden Markov Models o HMM por sus siglas en inglés) seguidas por un algoritmo de decodificación de Viterbi para el modelamiento de las secuencias de nucleótidos, en esta fase se tiene a Metrichor y Nanocall que son compatibles con medidas de señales de poros bajo la versión R7.3 (David et al., 2017).

La segunda etapa comprende al desarrollo de aproximaciones basadas en aprendizaje profundo que usualmente recurre a redes neurales recurrentes (Recurrent Neural Networks o RNN por sus siglas en inglés) cuyo rendimiento es superior a los Basecallers de la primera etapa, dentro de los cuales hay algoritmos como Deepnano (Boža et al., 2017), Nanonet, Albacore (antes de v2.0.1) y BasecRAWller (Stoiber & Brown, 2017).

Los Basecallers oficiales y otros de código libre son actualizados rápidamente para obtener mayor rendimiento en función de la capacidad computacional como: Guppy, un basecaller que puede beneficiarse de la aceleración de unidades de procesamiento de gráficos (GPU); Flappie, que utiliza un algoritmo capaz de distinguir regiones con bases repetitivas para evitar pérdida de información en homopolímeros y Albacore (a partir de v2.0.1) que utiliza estrategias libres de eventos para poder ser ejecutado de manera más fluida en unidades centrales de procesamiento (CPU), una forma de estimar el rendimiento de los Basecallers es mediante el mapeo de los resultados del procesamiento con un genoma de referencia con algoritmos de alineamiento como LAST (Frith et al., 2010) y minimap2 (Li, 2018) para evaluar la tasa de identidad por medio de la detección de inserciones, deleciones y alineamiento erróneo de bases para cada mapeo (Zeng et al., 2020).

Metagenómica

Se refiere a un análisis independiente de cultivo de la información genética presente en los genomas colectivos de los microorganismos presentes en un entorno delimitado por como este se toma como muestra a partir de un ambiente generalmente para detectar posibles interacciones entre los organismos y el medio. La recolección de datos se puede efectuar tomando como objetivo: material genético, lípidos, proteínas y otras moléculas pequeñas como metabolitos, posteriormente los análisis se pueden enfocar a: conservación de secuencias, filogenética, filogenómica, funcionalidad o diversidad genética (Izard, 2015).

Uso en virus

La metagenómica puede aplicarse para identificar la diversidad viral en un entorno por medio de la purificación de partículas virales y posterior secuenciación de los ácidos nucleicos a partir de una muestra ambiental tanto como de matrices vegetales y animales (Reske et al., 2007). La ventaja de la metagenómica viral por sobre otros métodos como la reacción en cadena de la polimerasa (PCR por sus siglas en inglés) y sus variantes es que permite caracterizar un espectro amplio de virus de diversas configuraciones y grupos que podrían no ser identificados debido a la naturaleza específica de otros métodos serológicos y moleculares y a la divergente conformación de las comunidades de virus que pueden estar presentes en una muestra (Ng et al., 2011).

Varios estudios de metagenómica enfocados en detección de virus realizados en MinION han sido capaces de secuenciar e identificar genomas completos de varios virus como el chikungunya, Zaire ebolavirus, hepatitis C, virus de encefalitis equina de Venezuela y Zika por medio de secuenciación directa dirigida al ARN o por un método de pre-amplificación mediada por extensión con cebadores (Greninger et al., 2015). Si bien la detección de virus bajo este

enfoque se puede realizar por medio de focalización de secuenciación o agotamiento de ARN ribosomal previos, se los puede combinar con un cultivo viral clásico para el enriquecimiento de las secuencias, método normalmente utilizado cuando se dispone de una cantidad limitada de muestra o ensayos netamente analíticos (Young et al., 2019).

NanoSim

Es un flujo de procesos diseñado para el análisis y simulación de lecturas de secuenciación obtenidas por medio de equipos que cuenten con Oxford Nanopore Technologies. Esta herramienta analiza lecturas de ONT para obtener características a manera de modelos como perfiles de error y distribuciones de tamaño a partir de datos experimentales para luego poder utilizarlos en la generación de lecturas *in silico* para una secuencia de referencia. Está implementado utilizando R para el ajuste de modelos de error y Python para el análisis de tamaño de lecturas y simulación (Yang et al., 2017).

El primer paso de NanoSim es la caracterización la cual requiere datos reales de secuenciación obtenida en ONT en un archivo fasta o fastq que efectúa un análisis comprensivo basado en alineamiento frente a un archivo de referencia que genera un conjunto de perfiles de lecturas a modo de datos de entrenamiento que sirven como entrada a la fase de simulación para producir lecturas *in silico* dado un genoma de referencia, además proporciona una lista de los errores introducidos en la simulación, especificando la posición y el tipo de error junto con la o las bases correspondientes a la referencia utilizada (BCGSC, 2020).

E-probe Diagnostic Nucleic acid Analysis (EDNA) – Microbe Finder (MiFi)

Herramienta bioinformática diseñada a manera de un flujo de procesos que utiliza sondas denominadas e-probes, para el análisis de detección de ácidos nucleicos cuyo objetivo es

minimizar e ignorar datos relacionados a secuencias enfocándose en secuencias específicamente asociadas a patógenos. Utiliza la herramienta de búsqueda de alineamientos locales básica (BLAST por su acrónimo en inglés) como algoritmo para el análisis solo que, en lugar de hacer búsquedas exhaustivas de cada secuencia frente a una base de datos curada, como el GenBank utilizado como base por el servidor de BLAST del National Center for Biotechnology Information (NCBI), la búsqueda se realiza tomando secuencias únicas denominadas e-probes frente a lecturas crudas de secuenciación como base de datos en sistemas computacionales locales o en el servidor oficial proporcionado por el Departamento de Entomología y Fitopatología de Oklahoma State University, Microbe Finder (MiFi), asegurando una detección rápida sin necesidad de análisis extensivo del metagenoma (Anthony H. Stobbe et al., 2013).

Validación de ensayos diagnósticos para enfermedades infecciosas

Proceso utilizado para determinar la aptitud de un ensayo el cual ha sido desarrollado, optimizado y estandarizado apropiadamente para un propósito deseado para una especie específica en la que va a ser utilizado. La validación incluye principalmente, registros de estimaciones de las características analíticas y diagnósticas del rendimiento del ensayo, parámetros de utilización y propósito del ensayo. Un ensayo debe mantener su estado de validado para poder seguir siendo utilizado de forma segura por lo que es necesario monitorear su comportamiento y eficacia en la población objetivo (Cardwell et al., 2018).

Los ensayos validados permiten: documentar la dinámica de una enfermedad y prevalencia de la infección en un país o región para facilitar un análisis de riesgo, prevenir su propagación debido a actividades de comercio, confirmar diagnósticos de casos posibles e identificar casos positivos para tomar medidas de control. Un solo ensayo puede ser validado

para varios propósitos mediante la optimización de sus características de desempeño como pruebas de cribado (alta sensibilidad y baja especificidad) o pruebas confirmatorias (baja sensibilidad y alta especificidad) (OIE, 2018a). Debido al creciente número de ensayos orientados a la detección de ácidos nucleicos, estos están siendo utilizados para el diagnóstico de enfermedades infecciosas. Para asegurar la validez de un ensayo se requieren cinco fases en las que cada una debe presentar un resultado favorable para continuar a la siguiente, estas fases son:

Criterio de desempeño analítico

En la primera fase se debe verificar que el ensayo debe mantener constancia de resultados al ser realizado varias veces en la misma muestra bajo los mismos parámetros en diferentes ocasiones y diferentes operarios de ser posible. Se debe valorar la especificidad analítica determinada por la capacidad de detectar un solo o un espectro de agentes patogénicos de forma selectiva frente otras cepas que no sean de interés para el ensayo, pudiéndose incluir aislados de una variedad de áreas geográficas que causen síndromes similares para verificar el poder discriminatorio o inclusivo de la prueba. También se debe estimar la sensibilidad analítica que comúnmente se refiere a evaluar el límite de detección (LOD por sus siglas en inglés), que consiste en determinar cuál es la concentración o título mínimo en el que la prueba es capaz de determinar si el patógeno está presente en una muestra, suele realizarse por medio de diluciones seriales de una muestra altamente positiva del patógeno en una matriz altamente negativa. Tanto para el ensayo de sensibilidad como de especificidad se debe tener un método estándar como referencia para poder contrastar los resultados obtenidos por el ensayo a validar, de preferencia que la naturaleza del estándar sea lo más similar posible al ensayo de prueba (OIE, 2018b).

Criterio de desempeño diagnóstico

Para la segunda fase se estiman tanto la sensibilidad como la especificidad diagnóstica que son considerados como los indicadores principales de desempeño de un ensayo diagnóstico. Para esto se necesita un método estándar para poder clasificar como verdaderos o falsos, los resultados positivos o negativos que el ensayo a validar pueda emitir tomando como referencia un estándar de detección, teniendo resultados categóricos en función de los puntos de corte o límites de decisión como tolerancia para el propósito del ensayo (Cardwell et al., 2018).

Consta de dos sub-fases, una para determinar una estimación previa en paneles reducidos de muestras (Fase 2a) para luego definir en la segunda sub-fase, cuántas muestras de resultado conocido se requieren para establecer el desempeño diagnóstico de campo (Fase 2b) (OIE, 2018d). En ambas fases se estima la sensibilidad y la especificidad mediante las ecuaciones:

$$Se = \frac{VP}{FN+VP} \times 100\% \quad (1)$$

$$Sp = \frac{VN}{FP+VN} \times 100\% \quad (2)$$

En las que: Se, representa la sensibilidad; Sp, la especificidad; VP, verdaderos positivos; FN, falsos negativos; VN, verdaderos negativos y FP, falsos positivos.

Reproducibilidad y estimaciones de repetitividad aumentada

Es la capacidad de llegar a un acuerdo entre los resultados obtenidos utilizando el mismo protocolo, similar equipamiento y el mismo panel de muestras en diferentes laboratorios o centros de análisis (OIE, 2018b).

Implementación del programa

Se refiere a la interpretación y posterior toma de decisiones una vez se han obtenido los resultados del ensayo en una comunidad real. En programas de vigilancia, dependiendo de los resultados obtenidos, se puede concluir la implementación del ensayo si los resultados concuerdan con proyecciones previas o si estos son consistentes con la naturaleza del ensayo y datos de la población, en caso de no ser así, es posible utilizar un ensayo estándar (como el efectuado para las fases previas) para realizar pruebas confirmatorias (OIE, 2018b).

Monitoreo del desempeño del ensayo luego de la validación inicial

La repetitividad del ensayo, incluidas modificaciones técnicas o inherentes a los parámetros de funcionamiento del ensayo se controlan en esta fase utilizando muestras control para poder confirmar el desempeño. Cambios mayores como descontinuación del ensayo o proposición de modificaciones para aumentar o modificar el alcance del mismo requieren de una revalidación del ensayo bajo las nuevas condiciones que produjeron cualquiera de los dos escenarios mencionados (OIE, 2018b).

Hipótesis

El diseño de e-probes para *Grapevine virus* A y *Grapevine virus* B y su posterior evaluación de desempeño por medio de la herramienta EDNA como método de detección en lecturas crudas de comunidades metagenómicas simuladas utilizando como base el flujo de procesos de NanoSim, permite el avance hasta la fase 2a de su validación como ensayo diagnóstico para los agentes patógenos virales definidos.

Materiales y métodos

MetaSpore

Una de las principales limitaciones de NanoSim como flujo de procesos para simular lecturas obtenidas mediante ONT es la necesidad de datos reales de secuenciación para establecer modelos de errores en la etapa de caracterización para luego ser utilizados como parte de la entrada en la etapa de simulación, además de la capacidad de sólo realizar un flujo de caracterización o simulación a la vez (BCGSC, 2020).

Es por esto que, previo al desarrollo de la investigación de validación, se trabajó en la propuesta de MetaSpore como una modificación y expansión de los algoritmos diseñados en el flujo de NanoSim para poder simular lecturas crudas de comunidades metagenómicas obtenidas por ONT. Las principales adiciones y ligeras modificaciones realizadas fueron:

- Posibilidad de lectura para caracterización de más de una cepa para una variedad de especies de organismos que pueden ser definidos como patógenos o como huéspedes en el mismo flujo.
- Algoritmo para la inserción de errores *de novo* por secuenciación con equipos que cuenten con ONT debido al error producido por varios Basecallers en un determinado rango de organismos basados en perfiles de error (Zeng et al., 2020).
- Generación de directorios con lecturas simuladas con errores *de novo* por secuenciación, métricas de las simulaciones (tamaño de lectura, errores, precisión y tasa de error) y datos de entrenamiento.
- Posibilidad de lectura para posterior simulación de uno o varios diseños de comunidades metagenómicas a manera de abundancias relativas fijas, un número exacto de lecturas

para los patógenos o abundancias relativas pseudoaleatorias entre un rango definido por el usuario.

- Algoritmo para la interpretación de los diseños de comunidades para la generación de lecturas crudas de ONT correspondientes a las comunidades metagenómicas.
- Generación de archivos fasta correspondientes a comunidades metagenómicas, métricas para cada simulación para cada miembro de la comunidad y la composición global de cada comunidad.
- Rediseño del algoritmo de lectura de archivos de entrada y generación de archivos de salida para permitir al usuario tener la opción de elegir como administrarlos según sus propósitos de investigación o capacidad computacional disponible.

MetaSpore funciona de una manera similar a la de NanoSim, comenzando con una fase de caracterización, que puede o no incluir la generación de errores *de novo* según lo que se requiera, seguido de la fase de simulación.

Dado un conjunto de secuencias de ADN de patógenos y huéspedes más archivos con instrucciones de diseño bajo un formato específico, se generan lecturas crudas de ONT correspondientes a comunidades metagenómicas.

MetaSpore está basado en la versión 2.6.0 de NanoSim actualizada hasta el día 7 de julio de 2020, la primera versión (v0.1) de MetaSpore fue pre-publicada el día 12 de septiembre de 2020, la versión más reciente (v0.5.3) fue pre-publicada el día 16 de enero de 2021 en github.com/adramoslp/MetaSpore a modo de un repositorio privado hasta su publicación posterior para libre acceso bajo la licencia general pública GNU en su versión 3 del 29 de junio de 2007, la misma bajo la cual la versión mencionada de NanoSim ha sido publicado.

Métricas de simulación

Una de las adiciones realizadas al flujo base de NanoSim es un algoritmo iterativo que permite recuperar los datos de los perfiles de error para cada lectura, para registrar el perfil de errores totales por lectura y posteriormente realizar el cálculo de la precisión y tasa de error de cada una las entradas o secuencias en las simulaciones generadas en el proceso.

Los datos obtenidos de esta parte del proceso conforman las métricas de simulación, que permiten definir la eficiencia de la secuenciación, sea esta en simulaciones o en ensayos reales, las ecuaciones utilizadas para estimar estos parámetros son:

$$ACC = \frac{matches * 100\%}{matches + mismatches + \sum(length(insertions \in read)) + \sum(length(deletions \in read))} \quad (3)$$

$$ERR = \frac{(mismatches + \sum(length(insertions \in read)) + \sum(length(deletions \in read))) * 100\%}{matches + mismatches + \sum(length(insertions \in read)) + \sum(length(deletions \in read))} \quad (4)$$

Ecuaciones en las que: ACC, se refiere a la precisión; ERR, a la tasa de error; $matches$, al número de alineamientos correctos entre una base de la lectura y la referencia mientras que $mismatches$ al número de alineamientos erróneos. Los términos con sumatoria de la longitud total para $insertions$ (inserciones) y $deletions$ (deleciones) se refieren a la suma total de los tamaños, en número de bases, de las regiones de la lectura en las cuales se extiende el error correspondiente (Rang et al., 2018).

Perfiles de error de diferentes Basecallers

Debido a la capacidad de procesamiento de los equipos que cuentan con ONT y a ciertas limitaciones de los diversos algoritmos que permiten el base calling de las secuencias, ciertos errores son introducidos al momento de interpretar las señales crudas captadas por los equipos

(Rang et al., 2018). Es por esto que para simular estos errores *de novo* debidos a la secuenciación, se propuso un algoritmo que pueda modelar estos errores para obtener lecturas lo más reales posibles de acuerdo a los perfiles de error de funcionamiento de cada Basecaller, los cuales han sido estimados para diferentes organismos modelo a modo de porcentaje, cuyos valores se presentan a continuación.

Tabla 1

Rendimiento de cinco Basecallers en cuatro conjuntos de prueba

Especie	Basecaller	Deleción (%)	Inserción (%)	Mismatch (%)
Fago Lambda	Causalcall	6.48	1.84	4.3
	Chiron (DNAde)	8.2	2.27	5.77
	Chiron (DNAre)	6.86	2.22	4.71
	Guppy	4.6	2.02	3
	Flappie	5.01	2.28	3.5
E. coli	Causalcall	5.95	2.07	4.57
	Chiron (DNAde)	7.07	2.47	6.04
	Chiron (DNAre)	5.91	2.34	4.65
	Guppy	4.06	1.97	3.02
	Flappie	4.6	2.28	3.6
Humano	Causalcall	8.06	2.27	5.06
	Chiron (DNAde)	8.49	2.92	5.41
	Chiron (DNAre)	7.76	2.98	5.56
	Guppy	4.78	2.46	2.86
	Flappie	5.33	2.67	3.35
K. pneumoniae	Causalcall	5.58	4.82	6.29
	Chiron (DNAde)	5.7	6.41	7.92
	Chiron (DNAre)	5.13	6.26	6.9
	Guppy	4.1	4.16	4.49
	Flappie	5.06	4.26	5.42

Nota: Recuperado y Modificado de Causalcall: Nanopore Basecalling Using a Temporal Convolutional Network por Zeng et al., 2020. Chiron (DNAde) y Chiron (DNAre) representan las

versiones de Chiron por defecto y la re-entrenada respectivamente. Mientras menor sea el error porcentual presentado, mejor es la calidad de secuenciación.

MetaSpore puede utilizar un consenso (estimado como la media), a preferencia del usuario, de los cuatro conjuntos de prueba para especies que puedan estar no relacionadas con ninguna de las presentadas en la Tabla 1.

Selección de Secuencias

Secuencias correspondientes a varias cepas de *Grapevine virus* A (GVA) (secuencia de referencia en el NCBI: NC_003604.2) y *Grapevine virus* B (GVB) (secuencia de referencia en el NCBI: NC_003602.1) fueron seleccionadas de la base de datos GenBank del NCBI por medio de una búsqueda por BLASTN de las secuencias de referencia de ambos virus, tomando cepas que correspondan a genomas completos que tengan al menos un 72% de identidad y 90% de cobertura respecto a las secuencias de referencia.

Para GVA se recuperaron diez secuencias correspondientes a cuatro zonas geográficas: Croacia (MF979533.1), Francia (MG925333.1), Israel (AF007415.2 y AY244516.1) y Sudáfrica (DQ855081.2, DQ855083.2, DQ855084.2, DQ855086.2, DQ855087.2 y KC962564.1).

Para GVB se recuperaron nueve secuencias correspondientes a cinco zonas geográficas: Brasil (KX790785.1), Canadá (JX513897.1 y KY426923.1), Chile (KF700375.1), Croacia (MF991949.1) y Sudáfrica (EF583906.1, GU733707.1, KJ524452.1 y KX522545.2).

Se clasificaron a las secuencias en cuatro grupos para GVA y cinco para GVB, nueve grupos en total, basado en las zonas geográficas mencionadas para el posterior flujo de ensayos para la validación en cada una.

Como secuencias para el huésped se recuperaron 19 secuencias correspondientes a todos los cromosomas correspondientes a cada uno de los cromosomas de *Vitis vinifera* al ser el huésped de interés para la investigación.

Todas las secuencias, incluidas las de los virus, fueron almacenadas en formato fasta y transformadas a formato Unix (LF) para su posterior procesamiento.

Diseño de e-probes mediante la herramienta EDNA

Se establecieron nueve grupos de e-probes correspondientes a cada zona geográfica, tomando como referencia taxonómica a la secuencia de referencia correspondiente a cada especie (NC_003604.2 para GVA y NC_003602.1 para GVB) y como secuencia objetivo un archivo en formato fasta con todas las cepas presentes en cada zona.

En caso de obtener e-probes que presenten poca relevancia biológica (cantidades bajas de e-probes o alineamientos ambiguos), la cepa correspondiente se descartó como objetivo para el análisis de sensibilidad.

Las e-probes fueron diseñadas utilizando la herramienta MiProbe presente en la plataforma MiFi con los siguientes parámetros comunes para todos los grupos de e-probes: longitud fija de 40 nucleótidos para el tamaño de las e-probes y un mínimo de 10 bases de emparejamiento entre las e-probes y una secuencia taxonómicamente cercana al objetivo.

Simulación de lecturas crudas de comunidades metagenómicas

Para la presente investigación todo el flujo de procesos relacionados a la simulación, fueron llevados a cabo en una computadora personal con Intel® Core™ i5-8300H CPU @ 2.30 GHz + 8GB RAM utilizando Microsoft Windows [Versión 10.0.18363.1379] como sistema operativo base, además de Unix - Ubuntu 20.04 64bit montado como sistema operativo virtual

en Oracle VM VirtualBox exclusivamente para la etapa de caracterización debido a la compatibilidad del alineador minimap2 (solo disponible en UNIX y en MacOS). La etapa de simulación, recolección, procesamiento de datos y análisis estadístico fueron llevados a cabo en el sistema operativo base.

Para obtener las comunidades necesarias para la validación del ensayo en cada zona geográfica para GVA y GVB se utilizó MetaSpore para obtener las simulaciones correspondientes en dos etapas.

Caracterización

En esta etapa se utilizó como base las secuencias provenientes del GenBank previamente recuperadas y se las procesó mediante la etapa de análisis de MetaSpore en dos pasos: introducción de errores *de novo* y establecimiento de los modelos de perfiles de error a utilizar en la etapa de simulación, esto debido a la falta de lecturas reales de ONT para su caracterización por no disponer de muestras reales para secuenciación.

Los parámetros específicos para la caracterización de las cepas de GVA y GVB como patógenos fueron: 1000 lecturas en total con errores *de novo* y “virus” como especie para el perfil de errores, mientras que los utilizados para cada uno de los 19 cromosomas de *Vitis vinifera* como huésped fueron: 50000 lecturas en total con errores *de novo* y “consensus” como especie para el perfil de errores.

Los parámetros comunes de caracterización para todos los miembros de las comunidades fueron: 2000 bases como tamaño de cada lectura, 100 bases como desviación estándar para el tamaño de cada lectura, “guppy” como perfil de rendimiento para simular los errores *de novo* del Basecaller, “minimap2” como algoritmo de mapeo genómico, 4 threads para

el procesamiento de los datos y Unix en Ubuntu 20.04 64bit como sistema operativo montado en Oracle VM VirtualBox sobre Microsoft Windows [Versión 10.0.18363.1379].

El tamaño de las secuencias correspondientes a cada organismo es de aproximadamente ~7.5KB para los patógenos y ~21MB para cada cromosoma del huésped, debido a esto, con el fin de asegurar la cobertura de toda la secuencia de referencia, el número total de lecturas con errores *de novo* simuladas es mayor para las secuencias correspondientes al huésped, para así poder mantener el mismo tamaño de lectura para todos los miembros de las comunidades.

Simulación

El diseño de la composición de las comunidades metagenómicas simuladas (detallado en la siguiente sección) fue distinto en función de la etapa del ensayo y del grupo de análisis al que estas pertenecieron, sin embargo, todas compartieron los siguientes parámetros generales de simulación: 10000 lecturas en total para la comunidad metagenómica, número de lecturas correspondientes a secuencias del huésped distribuidas de forma uniforme mientras que se mantuvo el valor por defecto para el resto de parámetros opcionales. El sistema operativo utilizado fue Microsoft Windows [Versión 10.0.18363.1379].

Validación de e-probes como método diagnóstico

Debido a la naturaleza del ensayo, fue posible ejecutar el proceso hasta la fase 2a de validación según el Manual de Pruebas diagnósticas y vacunas para animales terrestres estandarizado por la Organización Mundial para la salud animal (OIE), tomado como referencia al tener el manual más completo para validación de ensayos nuevos para diagnóstico de

enfermedades infecciosas y ser referido como base en trabajos de validación de pruebas en plantas (Cardwell et al., 2018).

Para la detección se utilizó las e-probes previamente diseñadas en la plataforma MiFi mediante la herramienta MiDetect bajo los parámetros de $1x10^{-9}$ como e-value a manera de límite para valores esperados y 250 como el número mínimo de hits para la detección para cada grupo de e-probes.

La validación para cada zona geográfica se realizó en las siguientes etapas con su respectivo conjunto de e-probes diseñado previamente. En caso que en una zona geográfica existiera más de una cepa, para el diseño de la comunidad, el número de lecturas correspondientes a la zona se distribuyó de forma aleatoria entre dichas cepas, esta consideración se mantuvo para todas las simulaciones de todas las zonas geográficas en los distintos grupos de análisis para cada etapa del ensayo.

Sensibilidad analítica (ASe)

Se utilizó el LOD al 95% como enfoque para estimar la ASe mediante el establecimiento de tres concentraciones decrecientes de patógenos correspondientes a la zona geográfica, en un estilo similar a una dilución serial reduciendo la concentración a la mitad en cada nueva dilución. Las tres concentraciones utilizadas se definieron como abundancias relativas porcentuales para todas las cepas de la localización geográfica teniendo: 0.25% o 25 lecturas, 0.125% o 13 lecturas y 0.0625% o 6 lecturas por cada comunidad, en las que el resto de lecturas correspondieron a los 19 cromosomas del huésped.

Para cada abundancia relativa porcentual se realizaron 20 simulaciones de comunidades a modo de réplicas del ensayo para la estimación del LOD al 95%, teniendo en total 60

simulaciones para esta etapa. En caso que una sola réplica haya tenido un resultado negativo, se tomó dicha concentración como el LOD, mientras que, si hubo más de un resultado negativo, se tomó la concentración previa como el LOD (OIE, 2018c).

Especificidad analítica (ASp)

Se determinaron tres grupos de ensayo para establecer el panel de especificidad en cada zona geográfica para las que se hicieron simulaciones con diferentes composiciones: selectividad, únicamente presente el huésped; exclusión, presentes el huésped con lecturas correspondientes a la referencia del virus (NC_003604.2 para GVA y NC_003602.1 para GVB) y por último inclusión, presentes el huésped y lecturas correspondientes a cepas del mismo virus de todas las localizaciones geográficas que no fuesen la objetivo. Estos tres grupos funcionaron como controles negativos en las etapas analíticas, por lo que la detección se hizo en paralelo.

Tanto para el grupo de inclusión como el de exclusión, la composición de los patógenos en abundancia relativa porcentual fue del 0.25% o 25 lecturas para cada zona geográfica (grupo de inclusión) o para la referencia (grupo de exclusión). Se efectuaron 60 simulaciones como réplicas para el grupo de inclusión y se reportó cualquier posible reacción cruzada generada en estos grupos (OIE, 2018c).

De manera general para todo el ensayo se realizaron únicamente 60 réplicas del grupo de selectividad ya que estas comunidades fueron utilizadas como controles negativos para todas las zonas geográficas, de la misma forma, para los grupos de exclusión se realizaron solo 60 réplicas con la referencia de GVA y 60 con la de GVB ya que cada grupo fue utilizado para el panel de las zonas geográficas que los tengan como referencia, es decir, para las cuatro de GVA y las cinco de GVB respectivamente.

En total, entre la etapa de ASe y ASp se simuló un total de 1260 comunidades metagenómicas de cada una de las cuales se recuperó la información correspondiente a las métricas de simulación y de composición de la comunidad. Las detecciones fueron llevadas a cabo utilizando un solo conjunto de e-probes correspondiente a la zona geográfica respectiva a la del presente análisis.

Repetitividad y reproducibilidad preliminar

Debido a la naturaleza del procesamiento de secuencias y estadística del método de detección, tanto la utilización de la herramienta EDNA por medio de la plataforma o en su versión descargable, siempre que se utilicen los mismos datos de entrada (comunidad metagenómica, e-probes y parámetros de detección), los resultados emitidos serán independientes del centro de investigación o análisis en el que se aplique el método o de las veces que este sea efectuado bajo la misma muestra, por lo que el ensayo es totalmente reproducible y susceptible a cuantas repeticiones sean determinadas a procesar por el usuario, razones por las cuales esta etapa no requirió mayor investigación.

Sensibilidad diagnóstica de reconocimiento provisional (pDSe)

En esta etapa se estableció un grupo de ensayo de 30 comunidades metagenómicas simuladas con abundancias generadas de formas pseudoaleatorias con MetaSpore, estas comunidades fueron compuestas por secuencias correspondientes al huésped y a la cepa o cepas de patógenos correspondientes a la zona geográfica de análisis, esto con el fin de estimar la capacidad preliminar de sensibilidad de detección de los patógenos en diversas composiciones de comunidades confirmadas como infectadas por el agente objetivo (OIE, 2018d).

En caso de obtener un resultado de detección positivo, se registró como verdadero positivo, mientras que en caso de negativo se lo registró como falso negativo. Se recuperó la información correspondiente a la composición de las comunidades simuladas en términos de número de lecturas para cada miembro.

Especificidad diagnóstica de reconocimiento provisional (pDSp)

Se realizó la simulación de 30 comunidades metagenómicas cuya composición constó de secuencias correspondientes al huésped, cepas de todas las zonas geográficas que no fuesen la objetivo y a la de referencia de la especie, esto con el fin de poder estimar la especificidad del ensayo de forma preliminar en comunidades confirmadas como no infectadas con el agente objetivo (OIE, 2018d).

En caso de obtener un resultado de detección positivo, se registró como falso positivo, mientras que en caso de negativo se lo registró como verdadero negativo. Se recuperó la información correspondiente a la composición de las comunidades simuladas como número de lecturas de patógenos y del huésped.

Determinación del cut-off

Se estimó la cantidad de muestras necesarias para la fase 2b de validación del ensayo como prueba diagnóstica, en la que se requeriría archivos de secuenciación con lecturas crudas reales de comunidades metagenómicas en caso de querer continuar la validación, para esto se tomó como referencia la siguiente tabla.

Tabla 2

Número de muestras para establecer los parámetros diagnósticos de una prueba diagnóstica

pDSe o pDSp	2% de error en la estimación			5% de error en la estimación		
	90%	95%	99%	90%	95%	99%
90%	610	864	1493	98	138	239
92%	466	707	1221	75	113	195
94%	382	542	935	61	87	150
95%	372	456	788	60	73	126
96%	260	369	637	42	59	102
97%	197	279	483	32	45	77
98%	133	188	325	21	30	52
99%	67	95	164	11	15	26

Nota: Recuperado de Chapter 1.1.6. Principles and methods of validation of diagnostic assays for infectious diseases por OIE, 2018a. Se presenta el número teórico de muestras necesarias para estimar la Sensibilidad y Especificidad diagnóstica, considerando un 2% o 5% de margen de error en la estimación previa y tres niveles de confianza (90%, 95% y 99%) para la estimación final (Fase 2b)

Para todos los ensayos de detección, la estimación de numero de muestras y los análisis estadísticos se estableció un nivel de confianza del 95% con un margen de error de 5% donde correspondiera. Para el procesamiento de datos y determinación de parámetros estadísticos se utilizó el software GraphPad Prism 9.

Resultados

Diseño de e-probes

Mediante el uso de la herramienta MiProbe en la plataforma MiFi se obtuvo las siguientes e-probes para las distintas cepas de cada zona geográfica establecida:

Tabla 3

Número de e-probes para las distintas zonas geográficas para GVA y GVB

Especie	Zona Geográfica	No. de accesión en el GenBank	Número de e-probes obtenidas por cepa	Total por Zona
GVA	Croacia	MF979533.1	15	15
	Francia	MG925333.1	13	13
	Israel	AF007415.2	9	18
		AY244516.1	9	
	Sudáfrica	DQ855081.2	15	82
		DQ855083.2	19	
		DQ855084.2	13	
		DQ855086.2	8	
		DQ855087.2	15	
	Brasil	KC962564.1	12	14
		KX790785.1	14	
	Canadá	JX513897.1	12	13
		KY426923.1	1	
GVB	China	KF700375.1	15	15
	Croacia	MF991949.1	15	15
	Sudáfrica	EF583906.1	11	58
		GU733707.1	11	
		KJ524452.1	18	
		KX522545.2	18	

Figura 1

Regiones a las que se alinean las e-probes diseñadas para cada cepa de GVA

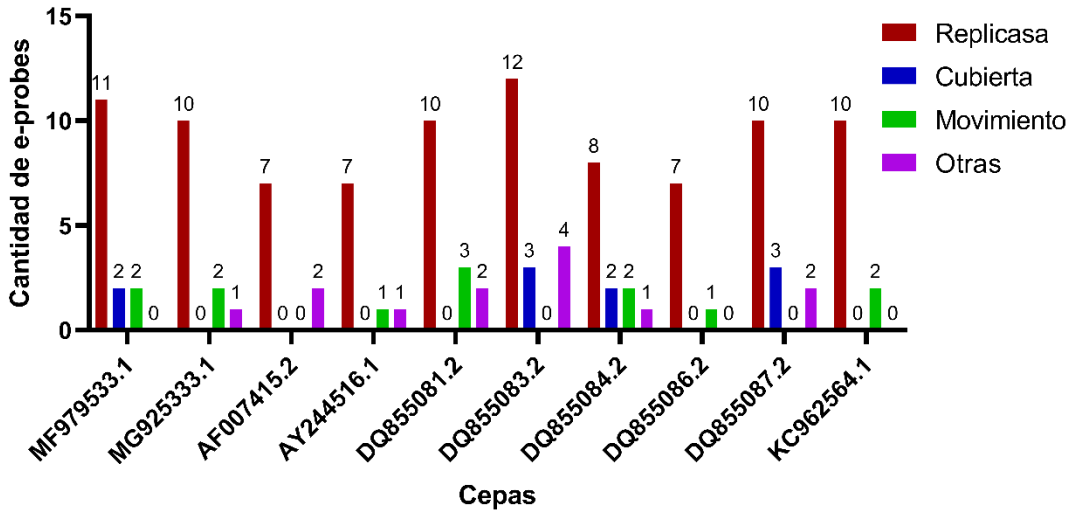
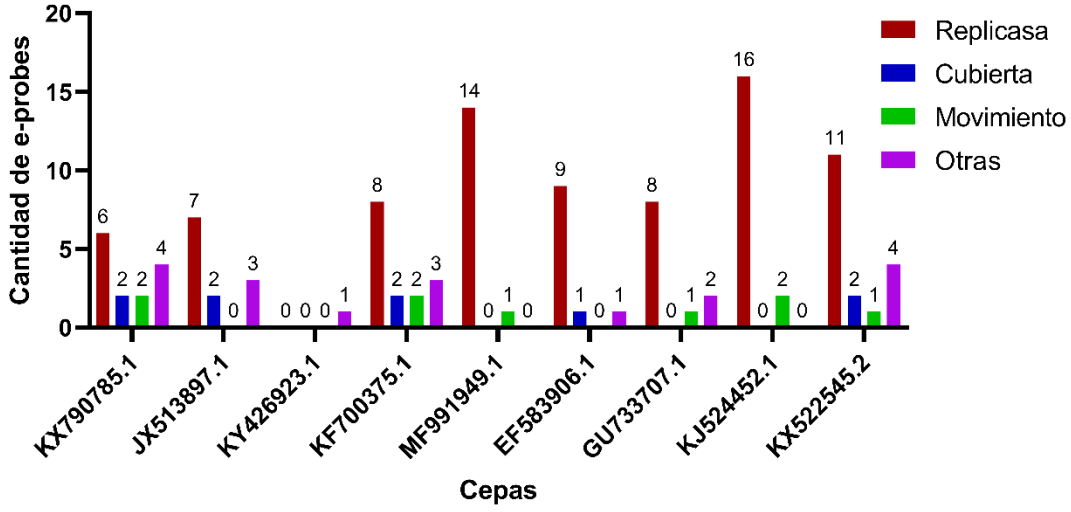


Figura 2

Regiones a las que se alinean las e-probes diseñadas para cada cepa de GVB



En las Figuras 1 y 2, se presenta la cantidad de e-probes diseñadas para cada cepa que se alinean a distintos marcos de lectura abiertos (ORFs). La clasificación “otras” se refiere a ORFs correspondientes proteínas hipotéticas, de función no descrita o de unión a ácidos nucleicos.

Simulación de comunidades metagenómicas

Composición de las comunidades

Para las simulaciones de los grupos de exclusión en el análisis de especificidad analítica se tiene 25 lecturas correspondientes a la referencia respectiva para cada una de las 60 comunidades simuladas, teniendo 1500 lecturas en total para NC_003604.2 así como 1500 para NC_003602.1. La composición del resto de comunidades se detalla a continuación.

Tabla 4

Número de lecturas para cada cepa de patógenos de GVA en las etapas analíticas

Zona Geográfica	No. de accesión en el GenBank	ASe (abundancia)			ASp (Grupos de Inclusión)			
		0.2500%	0.1250%	0.0625%	Croacia	Francia	Israel	Sudáfrica
Croacia	MF979533.1	500	260	120	-	1500	1500	1500
Francia	MG925333.1	500	260	120	1500	-	1500	1500
Israel	AF007415.2	264	147	69	799	795	NA	721
	AY244516.1	236	113	51	701	705	-	779
Sudáfrica	DQ855081.2	95	39	10	241	222	364	-
	DQ855083.2	104	31	23	271	322	196	-
	DQ855084.2	79	48	23	248	352	185	-
	DQ855086.2	89	27	29	261	230	264	-
	DQ855087.2	36	55	16	245	157	219	-
	KC962564.1	97	60	19	234	217	272	-

Tabla 5

Número de lecturas correspondientes a cada cepa de patógenos de GVB en las etapas analíticas

Zona Geográfica	No. de accesión en el GenBank	ASe (abundancia)			ASp (Grupos de Inclusión)				
		0.2500%	0.1250%	0.0625%	BRA	CAN	CHI	CRO	SA
Brasil (BRA)	KX790785.1	500	260	120	-	1500	1500	1500	1500
Canadá (CAN)	JX513897.1	500	260	120	779	-	800	825	793
	KY426923.1	-	-	-	721	-	700	675	707
China (CHI)	KF700375.1	500	260	120	1500	1500	-	1500	1500
Croacia (CRO)	MF991949.1	500	260	120	1500	1500	1500	-	1500
Sudáfrica (SA)	EF583906.1	141	67	38	332	376	287	348	-
	GU733707.1	75	55	24	457	404	430	380	-
	KJ524452.1	211	55	33	374	451	451	442	-
	KX522545.2	73	83	25	337	269	332	330	-

Para las Tablas 4 y 5, cada entrada representa el número de lecturas de la cepa un total de 20 comunidades en las columnas correspondientes a sensibilidad y en un total de 60 comunidades para las correspondientes a especificidad, debido a que la recuperación de datos se efectuó para cada conjunto de réplicas.

Para las lecturas correspondientes a los 19 cromosomas del huésped *Vitis vinifera* los datos fueron recolectados de las 1260 comunidades debido a su presencia en todas las simulaciones. A diferencia de las secuencias de patógenos, para el huésped también se obtuvo secuencias no alineadas (más del 90% en la tasa de errores según NanoSim) a la referencia respectiva en la simulación que representan el 1.487% de las lecturas correspondientes al huésped y el 1.480% de todas las lecturas de las 1260 comunidades.

Tabla 6

Número de lecturas correspondientes al huésped para las comunidades en las etapas analíticas

Cromosoma	No. de accesión en el GenBank	Lecturas alineadas	Lecturas no alineadas
Chr1	NC_012007.3	651231	8820
Chr2	NC_012008.3	651238	8820
Chr3	NC_012009.3	653752	6300
Chr4	NC_012010.3	651256	8820
Chr5	NC_012011.3	651233	8820
Chr6	NC_012012.3	647435	12600
Chr7	NC_012013.3	651230	8820
Chr8	NC_012014.3	653751	6300
Chr9	NC_012015.3	647451	12600
Chr10	NC_012016.3	647464	12600
Chr11	NC_012017.3	651240	8820
Chr12	NC_012018.3	653752	6300
Chr13	NC_012019.3	651247	8820
Chr14	NC_012020.3	652498	7560
Chr15	NC_012021.3	649985	10080
Chr16	NC_012022.3	644902	15136
Chr17	NC_012023.3	648729	11340
Chr18	NC_012024.3	647458	12600
Chr19	NC_012025.3	648711	11361

Los datos en detalle de la composición en términos de secuencias de los patógenos de todas las comunidades en las etapas diagnósticas se presentan en anexos.

Precisión de las secuencias

Los datos se recuperaron como el promedio de la precisión de todas las lecturas presentes en una sola comunidad para cada miembro de la misma y se clasificaron según la cepa para cada patógeno o para cada cromosoma del huésped.

Figura 3

Precisión de la simulación de lecturas correspondientes a cepas de cuatro zonas geográficas en las que se ha registrado presencia de GVA

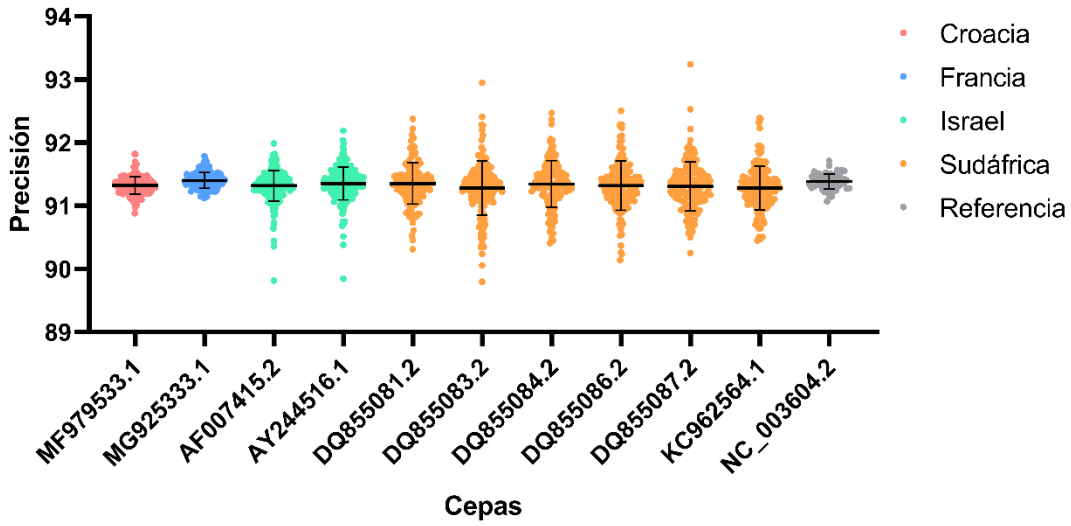


Tabla 7

Presencia de las cepas de GVA y precisión de su simulación

Zona Geográfica	No. de accesión en el GenBank	Comunidades en las que está presente	Precisión (%)
Croacia	MF979533.1	240	91.342 ± 0.135
Francia	MG925333.1	240	91.355 ± 0.124
Israel	AF007415.2	236	91.336 ± 0.240
	AY244516.1	230	91.338 ± 0.259
Sudáfrica	DQ855081.2	143	91.389 ± 0.328
	DQ855083.2	157	91.367 ± 0.429
	DQ855084.2	141	91.328 ± 0.370
	DQ855086.2	147	91.383 ± 0.391
	DQ855087.2	142	91.365 ± 0.391
	KC962564.1	155	91.378 ± 0.347
Referencia	NC_003604.2	60	91.384 ± 0.123

Figura 4

Precisión de la simulación de lecturas correspondientes a cepas de cinco zonas geográficas en las que se ha registrado presencia de GVB

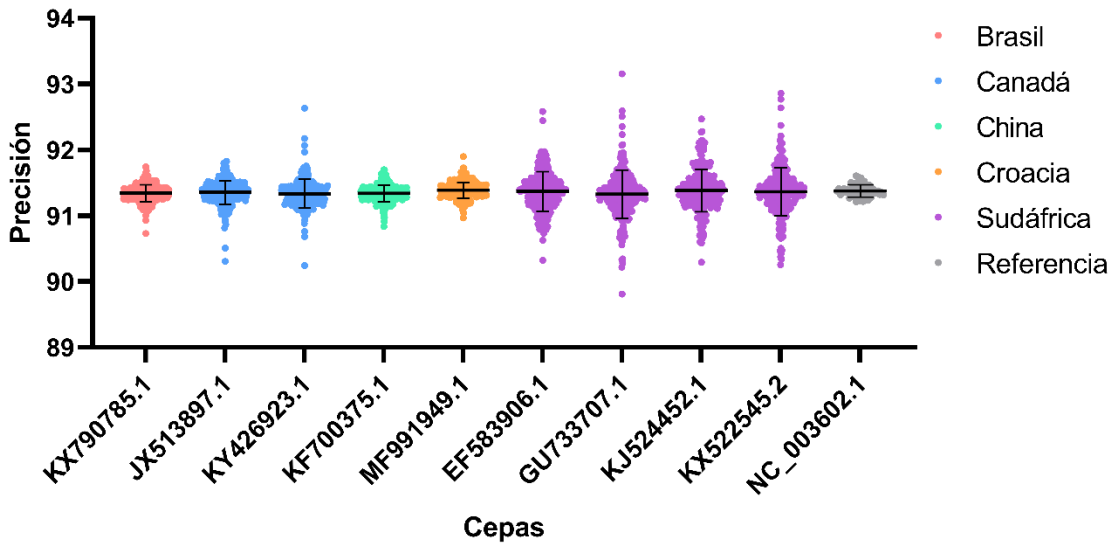


Tabla 8

Presencia de las cepas de GVB y precisión de su simulación

Zona Geográfica	No. de accesión en el GenBank	Comunidades en las que está presente	Precisión (%)
Brasil	KX790785.1	300	91.342 ± 0.127
Canadá	JX513897.1	298	91.355 ± 0.180
	KY426923.1	237	91.336 ± 0.222
China	KF700375.1	300	91.338 ± 0.125
Croacia	MF991949.1	300	91.389 ± 0.122
Sudáfrica	EF583906.1	239	91.367 ± 0.304
	GU733707.1	244	91.328 ± 0.366
	KJ524452.1	250	91.383 ± 0.324
	KX522545.2	240	91.365 ± 0.366
Referencia	NC_003602.1	60	91.378 ± 0.093

Las Figuras 3 y 4 muestran la precisión en la simulación de las secuencias de los patógenos estimados utilizando la Ecuación 3 agrupados por zona geográfica, en las simulaciones correspondientes a la Fase 1 de validación. Cada valor graficado representa al promedio de los valores de la precisión de las secuencias correspondientes a cada una de las cepas en una sola comunidad.

Las Tablas 7 y 8 presentan el número de comunidades en las que existe al menos una lectura de la cepa y la precisión de simulación de las lecturas como el promedio y desviación estándar de los valores graficados en las Figuras 3 y 4 respectivamente. Los datos presentados corresponden a las comunidades de los grupos de análisis para la estimación de ASe y ASp.

Figura 5

Precisión de la simulación de lecturas alineadas del huésped Vitis vinifera

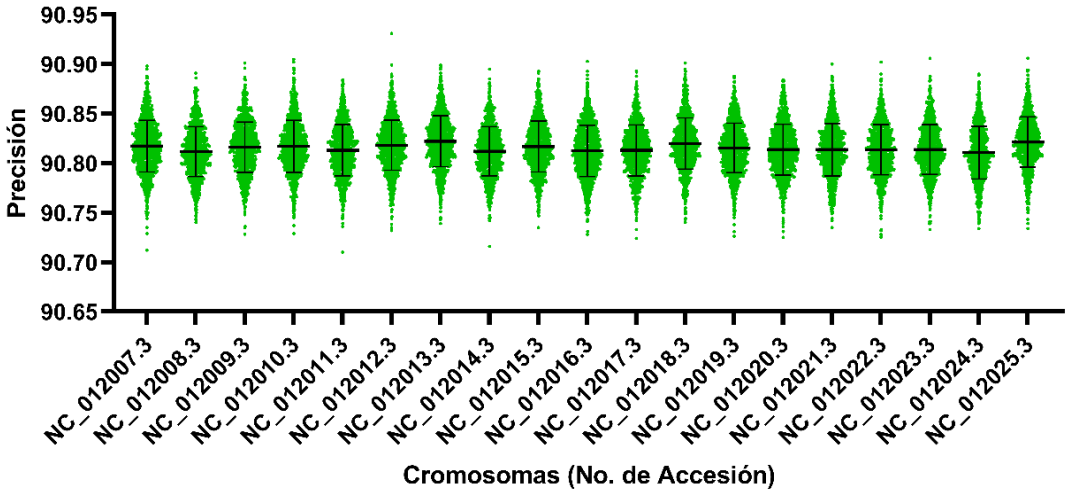
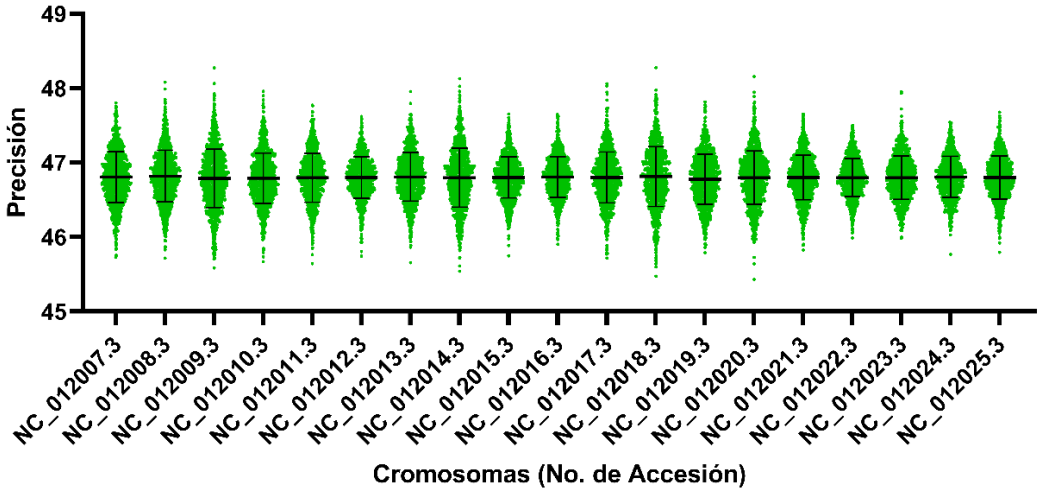


Figura 6

Precisión de la simulación de lecturas no alineadas del huésped Vitis vinifera



En las Figuras 5 y 6 se presentan los valores de la precisión en la simulación de las lecturas correspondientes a cada cromosoma del huésped en las comunidades utilizadas en las detecciones de la Fase 1 de validación. Cada valor graficado representa el promedio de los valores de la precisión en la simulación de todas las lecturas correspondientes a cada cromosoma en una sola comunidad, luego de haberlas clasificado como alineadas o no alineadas.

Las lecturas correspondientes a los cromosomas del huésped para cada comunidad fueron divididas en alineadas y no alineadas en base a la tasa de error de cada una de las secuencias simuladas, se clasificó una lectura como no alineada si esta presentó una tasa de error del 90% estimada con la Ecuación 4. Esta clasificación es realizada por defecto en el flujo de procesos de NanoSim.

Tabla 9

Precisión en la simulación de las secuencias para el huésped Vitis vinifera

Cromosoma	No. de Accesión en el GenBank	Precisión de secuencias alineadas (%)	Precisión de secuencias no alineadas (%)
Chr1	NC_012007.3	90.817 ± 0.026	46.806 ± 0.345
Chr2	NC_012008.3	90.812 ± 0.025	46.819 ± 0.348
Chr3	NC_012009.3	90.816 ± 0.025	46.788 ± 0.393
Chr4	NC_012010.3	90.817 ± 0.026	46.790 ± 0.339
Chr5	NC_012011.3	90.813 ± 0.026	46.798 ± 0.328
Chr6	NC_012012.3	90.818 ± 0.025	46.799 ± 0.278
Chr7	NC_012013.3	90.822 ± 0.026	46.808 ± 0.326
Chr8	NC_012014.3	90.812 ± 0.025	46.797 ± 0.396
Chr9	NC_012015.3	90.817 ± 0.026	46.801 ± 0.275
Chr10	NC_012016.3	90.812 ± 0.026	46.807 ± 0.276
Chr11	NC_012017.3	90.813 ± 0.026	46.800 ± 0.342
Chr12	NC_012018.3	90.820 ± 0.026	46.812 ± 0.400
Chr13	NC_012019.3	90.815 ± 0.025	46.777 ± 0.339
Chr14	NC_012020.3	90.814 ± 0.026	46.798 ± 0.357
Chr15	NC_012021.3	90.814 ± 0.027	46.800 ± 0.298
Chr16	NC_012022.3	90.814 ± 0.025	46.798 ± 0.257
Chr17	NC_012023.3	90.814 ± 0.025	46.797 ± 0.290
Chr18	NC_012024.3	90.811 ± 0.027	46.808 ± 0.276
Chr19	NC_012025.3	90.821 ± 0.025	46.800 ± 0.288

Nota: La precisión se representa como el promedio y la desviación estándar de los valores graficados en las Figuras 5 y 6 para las secuencias alineadas y no alineadas respectivamente. Los datos presentados corresponden a las lecturas correspondientes a los cromosomas del huésped en la fase 1 de validación.

Validación de la detección de GVA y GVB con e-probes

Sensibilidad y Especificidad

Tabla 10

Validación de la detección con e-probes para GVA y GVB hasta la Fase 2a

Especie	Zona Geográfica	ASe (LOD)	ASp (%)			pDSe (%)	pDSp (%)
			Selectividad	Exclusión	Inclusión		
GVA	Croacia	13	0	0	0	100	100
	Francia	13	0	0	0	100	100
	Israel	13	0	0	0	100	100
	Sudáfrica	13	0	0	0	100	91.667
GVB	Brasil	13	0	0	0	100	100
	Canadá	13	0	0	0	100	98.333
	China	13	0	0	0	100	100
	Croacia	6	0	0	0	100	98.333
	Sudáfrica	13	0	0	0	100	85

Nota: LOD en número de lecturas del patógeno. ASp presentado como porcentaje de detección de muestras como positivo para los paneles de: Sel. = Selectividad, Exc. = Exclusión, Inc. = Inclusión. Los valores de pDSe y pDSp fueron estimados utilizando las Ecuaciones 1 y 2 respectivamente.

Para la detección en las comunidades de todas las zonas geográficas en la etapa de ASp, todas las réplicas para los tres grupos definidos se tuvo un diagnóstico negativo con valores-p que no fueron calculados debido a la ausencia de alineamientos significativos de las e-probes. Para la etapa de pDSe y ASe en las abundancias relativas porcentuales de 0.25% y 0.125%, se tuvo un diagnóstico positivo con valores-p < 0.05 para la detección en todas las comunidades para todas las zonas geográficas. Para la detección en las comunidades correspondientes las comunidades con abundancias relativas porcentuales de 0.0625% en la etapa de ASe y todas las

comunidades para la etapa de pDSp, los valores-p correspondientes a los resultados del diagnóstico se presentan como anexos.

Reacción Cruzada

Tabla 11

Reacción cruzada presentada en la fase de estimación de especificidad diagnóstica de reconocimiento provisional

Especie	Zona Geográfica	Objetivo de las e-probes	Regiones de alineamiento	NAS	Cepas en las que se presentó reacción	Zonas Geográficas
GVA	Sudáfrica	DQ855081.2	R	11	AF007415.2, AY244516.1	Israel
		DQ855083.2	R	9	AF007415.2, AY244516.1	Israel
		DQ855084.2	M, C	12	MF979533.1	Croacia
		DQ855087.2	R	10	AF007415.2, AY244516.1	Israel
		KC962564.1	R	10	AF007415.2, AY244516.1	Israel
GVB	Canadá	JX513897.1	R, U	12	MF991949.1, GU733707.1	Croacia, Sudáfrica
	Croacia	MF991949.1	R, M	7	JX513897.1, KJ524452.1	Canadá, Sudáfrica
	Sudáfrica	EF583906.1	R, C, 3'UTR	74	MF991949.1, JX513897.1	Canadá, Croacia
		GU733707.1	R, 3'UTR	46	KX790785.1, MF991949.1, JX513897.1	Brasil, Croacia, Canadá
		KJ524452.1	R	30	MF991949.1	Croacia
		KX522545.2	U, C	3	MF991949.1, KF700375.1, JX513897.1	Croacia, China, Canadá

Nota: R, Proteína de la replicasa; M, Proteína de movimiento; C, proteína de la cubierta; U, Proteína de unión a ácido nucleico; 3'UTR, región no transcrita en el extremo 3'; NAS, Número de alineamientos significativos con un e-value igual o inferior a $1x10^{-9}$. La reacción cruzada se registró únicamente para cepas de la misma especie en diferentes zonas geográficas.

Estimación de muestras para la Fase 2b

Tabla 12

Estimación de muestras necesarias para la Fase 2b de validación de la detección con e-probes para GVA y GVB

Especie	Zona Geográfica	Muestras para establecer DSe	Muestras para establecer DSp
GVA	Croacia	15	15
	Francia	15	15
	Israel	15	15
	Sudáfrica	15	138
GVB	Brasil	15	15
	Canadá	15	30
	China	15	15
	Croacia	15	30
	Sudáfrica	15	-

Nota: El Manual de Pruebas diagnósticas y vacunas para animales terrestres, tomado como referencia para la validación, no especifica el número de muestras necesarias para el pDSp de 85% de la zona geográfica de Sudáfrica para GVB.

Discusión

Diseño de e-probes

En estudios previos se ha analizado el potencial de la herramienta EDNA para la detección de fitopatógenos, logrando detectar a GVA y GVB utilizando e-probes diseñadas con EDNA a nivel de especie en muestras pertenecientes a secuenciación con NGS (Visser et al., 2016) además que se ha señalado el potencial de detección de estos agentes a nivel de cepas y variedades debido a la capacidad de modular la especificidad del método (A. H. Stobbe et al., 2014). Para el diseño de e-probes se debe proveer una secuencia con una estrecha relación filogenética denominada “near neighbor” por EDNA, en una metodología previa se utilizaron secuencias de especies dentro del mismo género para los *Vitivirus* como near neighbors (Visser et al., 2016), por lo que en la presente investigación se utilizó, de forma análoga, la secuencia de referencia de cada especie en el GenBank como near neighbor en el diseño de e-probes para las diversas cepas en cada zona geográfica.

El tamaño de las e-probes permite modular los parámetros diagnósticos de la detección, teniendo de 40 a 60 nucleótidos como los tamaños más óptimos, puesto que si son más pequeñas se pierde especificidad y si son más grandes pierden sensibilidad (Melcher et al., 2014). El tamaño elegido fue fijo de 40 nucleótidos debido a que se ha reportado valores de rendimiento altos para este tamaño, además de obtener mayor cantidad de e-probes que diseños con 60 nucleótidos (Espindola et al., 2015).

La mayor cantidad de e-probes obtenidas se alinean a regiones correspondientes a ORFs que codifican la replicasa de ARN dependiente de ARN siendo 171 en total, esto se debe a que estos ORFs son los más grandes en los virus de interés por lo que hay mayor probabilidad de que

las e-probes sean diseñadas en base a estas zonas ya que, junto a las proteínas de la cubierta, son las que definen el criterio de demarcación para la especie (ICTV, 2020), por lo que en MiProbe al tener como near neighbor la referencia de la especie, se tendió a diseñar e-probes para estas regiones al ser las más similares debido a su relevancia para la delimitación del género *Vitivirus*.

El descarte de la única e-probe diseñada para la cepa KY426923.1 de la zona geográfica de Canadá para el GVB, se debe a que esta se alineó a un ORF correspondiente a una proteína hipotética que no dispone de mayores detalles sobre su función o relevancia y que además está presente en otras especies no relacionadas, por esto la detección que podría realizarse con dicha e-probe carecería de la relevancia biológica que presentan las demás e-probes diseñadas para las otras cepas. Al descartar esta e-probe como material para la detección, se descartó también su cepa correspondiente para análisis de sensibilidad al no ser posible una detección relevante, sin embargo, dicha cepa se conservó en caso de poder ser objetivo para otras e-probes como reacción cruzada para los análisis de especificidad.

Simulación de comunidades

Nanopore como tecnología de secuenciación permite el desarrollo de ensayos orientados a la secuenciación de genomas completos de plantas debido a su capacidad para procesar lecturas largas, para lograr este objetivo se requiere varios tamaños de lecturas para diferentes regiones de diferentes cromosomas, especialmente las repetitivas, con el fin de tener buen material crudo para posteriores ensamblajes (Jiao & Schneeberger, 2017). En la presente investigación, se optó por diseñar comunidades en las que la planta huésped esté representada con lecturas simuladas de forma homogénea respecto al tamaño y abundancia en cromosomas,

esto con el fin de asegurar que cualquier posible reacción cruzada con el huésped, tenga una probabilidad similar de ocurrir para cada región.

Otra razón para el diseño de lecturas de tamaño homogéneo se debe a que NanoSim, herramienta en la que está basada MetaSpore, requiere de una capacidad computacional mayor a la disponible para poder simular tamaños variados en un rango fijo determinado por el usuario (BCGSC, 2020). Todas las comunidades simuladas con MetaSpore presentaron composiciones, tamaño de lecturas, archivos de salida y sus correspondientes formatos consistentes con los parámetros exactos de diseño establecidos para las fases analíticas, para las comunidades de las fases diagnósticas la única diferencia fue la composición variable de cada una, esto debido a la generación de abundancias pseudoaleatorias en un rango definido con el fin de reproducir un mayor espectro de muestras para los propósitos de dichas etapas.

Para la simulación de lecturas de patógenos en caso de varias cepas para una sola zona geográfica, al no disponer previamente de abundancias relativas conocidas, el diseño planteado distribuyó de forma pseudoaleatoria el número fijo de lecturas por zona, con el fin de cubrir una variedad más amplia de composiciones como lo sería si la distribución fuese uniforme, es por esto que en comunidades para estas zonas geográficas existe la posibilidad de que una o varias cepas no estén presentes como se muestra en los datos de en las Tablas 7 y 8.

Precisión de las simulaciones

El basecaller utilizado como base para la introducción de errores de secuenciación en las simulaciones fue Guppy ya que es uno de los más utilizados debido a la posibilidad de optimización por medio de uso de GPUs (Nobile et al., 2017), para las secuencias respectivas a la planta huésped se utilizó un consenso de los perfiles presentados para los demás modelos

presentes, esto se debe a que de momento no hay perfiles de errores definidos para plantas ya que usualmente los Basecallers están entrenados para la secuenciación de muestras humanas o de bacterias (Dumschott et al., 2020), mientras que se utilizó el perfil de virus disponible para las secuencias de los patógenos.

La precisión de las lecturas simuladas se puede clasificar de manera general en alineadas, correspondientes al 98.520% del total de lecturas y no alineadas, correspondientes al 1.480%, las lecturas no alineadas se generan mediante una distribución arbitraria de perfiles de errores para regiones no caracterizadas previamente, esto puede deberse a que la cobertura planteada inicialmente para el huésped no fue la suficiente para poder representar el genoma, sin embargo, NanoSim tiende a generar este tipo de lecturas también como una forma de representar errores aleatorios de secuenciación en equipos reales (Yang et al., 2017). Los valores de precisión para lecturas alineadas, detallados en las Tablas 7, 8 y 9, concuerdan con la precisión de 86.536 ± 5.858% reportada para el rendimiento del basecaller Guppy para lecturas crudas de secuenciación para bases de datos modelo (Wick, R; Judd, L; Holt, 2018) y de hasta el 91% para *Klebsiella pneumoniae* como punto de referencia para secuenciación (Wick et al., 2019).

De forma particular, las valores obtenidos son consistentes con precisiones de hasta el 92% en lecturas crudas reportadas para secuencias de ARN mensajero y ribosomal (Jain et al., 2016) como referencia para las lecturas de patógenos virales y de hasta el 90% reportado para lecturas crudas de genomas de plantas previas al ensamblaje en el que se tiene precisiones de hasta el 99.4% (Jiao & Schneeberger, 2017) como referencia para lecturas de la planta huésped.

Si bien la precisión de secuencias consenso, obtenidas luego del ensamblaje, pueden llegar hasta un 99.81% con Casualcall (Zeng et al., 2020) y >99.9% con Guppy (Wick et al., 2019), en la investigación presente se tomó como referencia la precisión de las lecturas crudas, dado que el objetivo del método de diagnóstico propuesto es evitar el procesamiento de lecturas y ensamblaje posterior a la secuenciación.

Rendimiento de las e-probes para la detección de patógenos

Investigaciones previas han utilizado e-values desde $1x10^{-3}$ como punto de corte para los alineamientos de las e-probes (Visser et al., 2016), sin embargo, valores de $1x10^{-9}$ son los más adecuados para una detección confiable de patógenos con alto puntaje como soporte (A. H. Stobbe et al., 2014), por lo que se utilizó este valor como punto de corte en todos los análisis mientras que el número mínimo de hits por e-probe fue fijado en 250 (valor más alto disponible en MiDetect), ya que al diseñar un método de detección capaz de discriminar patógenos a nivel de cepas se requiere de mayor especificidad para reducir la oportunidad de que alineamientos de baja calidad cumulativos representen detecciones positivas falsas.

Sensibilidad analítica y diagnóstica

La sensibilidad analítica, estimada como el LOD, fue de 13 lecturas de patógenos para cada zona geográfica a excepción de la zona de Croacia para el GVB en la que el LOD fue de 6 lecturas. La disminución de detección de patógenos en muestras conocidas como positivas se debe a la rigurosidad de los parámetros con los que se configuró MiDetect para los análisis, mientras menos lecturas estén presentes, menor es la posibilidad de tener aciertos o hits con un e-value igual o menor a $1x10^{-9}$. Existen dos formas con las que se podría reducir el LOD, la primera es reduciendo la rigurosidad en MiDetect, no obstante, esto implicaría la pérdida de especificidad con la que se diseñó la prueba diagnóstica de forma particular para la detección a

nivel de cepas, también debido a que se ha reportado que a mayor sensibilidad la detección en muestras con una cantidad muy reducida de patógenos se vuelve poco confiable (Espindola et al., 2015).

La segunda forma es mediante la simulación de lecturas de mayor tamaño, lo que aumentaría la posibilidad de que las e-probes se alineen a más regiones por tanto incrementando el número de hits, sin embargo, esto supondría la necesidad de mayor capacidad computacional ya que los requerimientos escalan conforme más lecturas o mayor tamaño de estas se procese (BCGSC, 2020), considerando además que el propósito de la investigación es efectuar análisis en comunidades relativamente pequeñas, puesto que al implementar la prueba con muestras reales en MinION no se dispone de la misma libertad de obtención y procesamiento de datos ofrecida por simulaciones.

El LOD más bajo presentado en el presente estudio fue el registrado para la detección con el conjunto de e-probes diseñadas para la zona de Croacia para el GVB, siendo el que contó con una de las cantidades más altas de e-probes alineadas al ORF de la proteína de la replicasa como se muestra en la Figura 2, lo cual permitió que la detección el ORF de mayor tamaño del virus se efectúe con más e-probes que en las demás zonas geográficas, aumentando la cantidad de regiones objetivo para la detección y la posibilidad de obtener alineamientos significativos resultando en una disminución del LOD.

La composición patogénica de las comunidades de Sudáfrica para el GVB está distribuida entre varias cepas, esto produjo que en las comunidades con abundancia de 6 lecturas para la estimación de la ASe existieran réplicas en las que no hubo lecturas correspondientes a la cepa KJ524452.1, al no contar con estas lecturas como objetivo de las e-probes para efectuar una

detección significativa total que aproveche las e-probes disponibles para esta cepa, no se repitió la disminución de LOD presenciada anteriormente en la zona de Croacia para el GVB a pesar de que se disponía de una mayor cantidad de e-probes, acertando en no detectar a la cepa KJ524452.1 pero fallando al dar un diagnóstico negativo dado que sí existían las 6 lecturas pertenecientes a otras cepas de la zona geográfica en la comunidad.

La sensibilidad diagnóstica obtenida para todas las zonas geográficas es del 100% en las que se detectó a los patógenos objetivos en todas las comunidades, consistente con los resultados de LOD obtenidos en la fase analítica, ya que todas las comunidades presentaron abundancias mayores a los LOD para cada zona geográfica. La mayoría de ensayos orientados a la detección de estos virus se enfocan a determinar la presencia o ausencia del virus, ya sea por medio de PCR (Goszczynski, 2015), detección con microarrays (Engel et al., 2010) y PCR seguida de subclonaje y posterior secuenciación con NGS (Sabella et al., 2018), estos métodos permiten la detección pero no proporcionan información acerca de la abundancia inicial de los patógenos, a diferencia de los resultados obtenidos por medio de la secuenciación de muestras metagenómicas, debido a que el diagnóstico con e-probes y análisis estructurales o de composición pueden efectuarse con la misma muestra, pudiendo llegar a establecer correlaciones entre abundancia y estado de infección.

Especificidad analítica y diagnostica de reconocimiento provisional

A diferencia de los resultados obtenidos en las fases de sensibilidad, en la etapa analítica no hubo casos de reacción cruzada significativa, a diferencia de los casos para la etapa diagnóstica donde incluso existieron falsos positivos. Para la etapa analítica no hubo la suficiente representación de lecturas de los patógenos, por lo que la detección de regiones en las que se reportó reacciones cruzadas en la parte diagnóstica, solo fue significativa cuando

estas aparecían en mayor cantidad en la comunidad, escenario posible en comunidades con abundancias altas de patógenos. Al ser la etapa analítica, orientada a la estimación de LOD, los ensayos fueron diseñados para operar con una abundancia reducida de patógenos por lo que reacciones cruzadas, como las que se presentaron en la etapa diagnóstica, son menos frecuentes ya que se reduce la posibilidad de alineamientos fortuitos, además como se mencionó anteriormente, la sensibilidad disminuye cuando se cuenta con un número reducido de lecturas para patógenos (Espindola et al., 2015). Las e-probes diseñadas presentan alta sensibilidad para los objetivos de cada diseño, mientras que la especificidad puede variar para las e-probes de ciertas zonas, decreciendo mientras existan más agentes interferentes que tengan características estructurales similares a los objetivos.

La mayoría de reacciones cruzadas se reportaron para los grupos de e-probes correspondientes a la zona geográfica de Sudáfrica que se alinean a regiones de los ORF que codifican la proteína de la replicasa, seguido de regiones pertenecientes a la proteína de la cubierta. Estudios previos en los que se analiza la distribución filogeográfica del GVA, clasifica a secuencias de cepas pertenecientes a Sudáfrica en tres grupos distintos basándose en la replicasa y la proteína de la cubierta además de reportar una extensa heterogeneidad entre secuencias de la replicasa (Goszczynski et al., 2008; Goszczynski & Habili, 2012). La propagación clonal y la naturaleza perenne de los viñedos además de la tendencia a la inserción de errores en la secuencia correspondiente a la replicasa, llevan a una alta frecuencia de errores en la replicación del genoma dado que es el ORF más extenso (Drake & Holland, 1999). Las e-probes que presentan la mayor cantidad de reacciones cruzadas son las correspondientes a la zona geográfica de Sudáfrica, en las que la mayoría se alinean a regiones de los ORF de las replicasas cuya variabilidad ha probado ser amplia y con estructura similar a las de otras zonas, es por esto

que la especificidad en la detección se ve disminuida al poder tener como objetivo regiones similares e incluso idénticas en ciertos casos, a la de diseño en cepas de zonas distintas. Los casos de reacciones cruzadas ocurren menos frecuentemente en la detección con conjuntos de e-probes diseñados para zonas geográficas con menor cantidad de cepas debido a que la variedad de e-probes es más restringida, de la misma forma que ocurrió en la zona de Sudáfrica, los casos reportados en las zonas de Canadá y Croacia para el GVB también se deben a alineamientos en regiones de ORFs de las replicasas.

En contraste, la separación filogenética basada en las secuencias correspondientes a la proteína de la cubierta es más consistente con la separación basada en genomas completos, sugiriendo que son mejores indicadores de variabilidad significativa para la clasificación de considerando su origen geográfico (Alabi et al., 2014), por lo que para tener una detección más específica se podría emplear exclusivamente e-probes que se alineen a regiones de ORFs correspondientes a proteínas de la cubierta, sin embargo, la cantidad total de e-probes se ve muy reducida además que se pierde relevancia biológica debido a que en algunos casos, las e-probes se alinean a regiones que codifican proteínas hipotéticas, de función desconocida o a regiones no transcritas.

La utilización de e-probes para las zonas geográficas de Croacia y Canadá para el GVB que presentan reacciones cruzadas no se ve afectada ya que la especificidad sigue siendo alta y en ambos casos solo ocurrió en una muestra de las 30 correspondientes a cada zona, para las zonas de Sudáfrica, tanto para el GVA como para el GVB, el rendimiento se ve disminuido al presentar una baja especificidad para una prueba orientada a la discriminación a nivel de cepas, por lo que el método pierde su propósito inicial ya que una forma de confirmar si existe reacción cruzada o no, es mediante el análisis particular de secuencias. La alta sensibilidad de las e-

probes pertenecientes a Sudáfrica puede ser evidencia de recombinación, ya que es común que tanto el GVA como el GVB se presenten en infecciones mixtas, incluso con virus de otros géneros, causando que estos eventos sean comunes dentro del huésped (Alabi et al., 2010), por lo que estas e-probes pueden ser indicadores de puntos calientes de recombinación que permitirían el análisis del historial filogenético de estos virus.

Estimación de muestras para la Fase 2b

El número de muestras necesarias para evaluar la sensibilidad diagnóstica es el menor posible para el nivel de confianza seleccionado, mientras que para la especificidad diagnóstica aumenta para las e-probes de zonas que hayan presentado reacción cruzada. De manera particular, las e-probes para la zona de Sudáfrica presentan una baja especificidad de reconocimiento provisional, lo cual aumenta el número de muestras para la estimación confiable de la especificidad diagnóstica de campo requiriendo 138 para el GVA y siendo no determinado para el GVB.

La disponibilidad de muestras puede representar una limitación para la estimación de los parámetros diagnósticos, por lo que se sugiere el trabajo con la mayor cantidad de muestras posibles o realizar el trabajo progresivamente mientras estas puedan adquirirse, mientras que en el caso de las e-probes correspondientes a la zona de Sudáfrica es recomendable optimizar el método (OIE, 2018a), esto puede lograrse mediante el descarte de las e-probes alineadas a regiones del ORF de la replicasa que presentaron reacción cruzada, sin embargo, de momento no pueden utilizarse e-probes personalizadas en la plataforma MiFi y la versión descargable, que permitiría la personalización, no está en la última actualización. Otra alternativa es realizar una detección adicional con las e-probes diseñadas específicamente para las zonas geográficas de

posible reacción cruzada, con el fin de confirmar que una detección positiva previa es resultado de alineamiento al objetivo apropiado y no por reacción cruzada.

La utilización de e-probes puede ser individual o en grupos en el mismo análisis ya que tanto la detección como el diagnóstico, se procesan de forma independiente para cada conjunto de e-probes, por lo que no habría interferencia ni necesidad de validar un método de detección de un grupo de conjuntos de e-probes que ya hayan sido validados individualmente (Anthony H. Stobbe et al., 2013).

Este flujo de procesos, desde la fase de simulación hasta la de detección, permitiría la estimación de varios parámetros para métodos diagnósticos similares, orientados a varios fitopatógenos e incluso patógenos animales y humanos con la información de secuencias de entrada y capacidad computacional necesarias, debido a la alta reproducibilidad, repetitividad y accesibilidad. Particularmente, la etapa de simulación permitiría el desarrollo de investigaciones en el campo de la metagenómica, debido a que el uso de equipos con ONT tiene un alto potencial para estas aplicaciones y la generación de datos por medio de simulación, permitiría establecer puntos de referencia para el desarrollo y optimización de análisis metagenómico de comunidades.

Conclusiones

Los datos de entrenamiento necesarios para la simulación *in silico* de lecturas de Oxford Nanopore Sequencing correspondientes a *Grapevine virus* A y *Grapevine virus* B pueden ser generados con el algoritmo de inserción de errores *de novo* del flujo de procesos de MetaSpore en conjunto con la fase de caracterización de perfiles de lectura y ajuste de errores presentada en el paquete NanoSim.

La simulación de lecturas crudas de Nanopore correspondientes a comunidades metagenómicas compuestas por *Grapevine virus* A y *Grapevine virus* B como patógenos en el huésped *Vitis vinifera,* efectuada mediante MetaSpore con secuencias curadas provenientes del GenBank como información de entrada presenta una precisión consistente con la ofrecida por el rendimiento actual de la tecnología de secuenciación.

La posibilidad de diseño de comunidades metagenómicas propuesta por el flujo de procesos de MetaSpore permite la simulación de las muestras de referencia necesarias en la estimación de parámetros para la validación del desempeño de una prueba diagnóstica, hasta la fase de reconocimiento provisional, orientada a la detección de varias cepas en distintas zonas geográficas de los virus *Grapevine virus* A y *Grapevine virus* B.

El diseño de e-probes como sondas para la detección de varias cepas de *Grapevine virus* A y *Grapevine virus* B y la estimación de sus parámetros de desempeño fue posible mediante la utilización de MiFi: Microbe Finder bajo parámetros de sensibilidad ajustados para el diagnóstico en muestras de comunidades metagenómicas obtenidas por medio de simulaciones con MetaSpore.

Los valores de reconocimiento provisional de la detección por medio de e-probes de cepas de los virus *Grapevine virus* A en las zonas geográficas de Croacia, Francia e Israel y *Grapevine virus* B en las zonas de Brasil, Canadá, China y Croacia sirven de respaldo para su posible utilización en el diagnóstico de infecciones con las diversas cepas de los agentes patogénicos mencionados.

Recomendaciones

Las simulaciones realizadas en la presente investigación toman como base perfiles de error de secuencias para organismos modelo y secuencias tomadas del GenBank, una forma de obtener simulaciones lo más aproximadas a muestras reales es disponiendo de al menos una muestra de lecturas reales de secuenciación para la etapa de caracterización, esto podría aumentar la fidelidad de las simulaciones respecto a la composición de las comunidades y precisión de las lecturas, dado que sea una muestra confiable y de buena calidad.

MetaSpore es un flujo de procesos que puede ejecutarse en computadoras personales para simulaciones de comunidades relativamente pequeñas, considerando la capacidad de ONT para secuenciación. Capacidades computacionales como las de clústeres o supercomputadoras, además de poder realizar las mismas simulaciones que una computadora personal en mucho menor tiempo, permiten una mayor cantidad de simulaciones con composiciones más complejas en cuanto a número de lecturas y el tamaño de estas, aumentando la cantidad de análisis posibles con MetaSpore.

La estimación de los parámetros de desempeño diagnóstico es la fase que requiere la mayor variedad de muestras, para el panel de reconocimiento inicial y para la determinación final de sensibilidad y especificidad. La metodología propuesta puede resultar efectiva para el panel inicial, ya que los valores de desempeño preliminares no siempre están disponibles y suelen utilizarse muestras que podrían ser analizadas en la etapa de estimación final. Las simulaciones *in silico* de estas muestras para el panel preliminar, podrían emplearse para obtener estos valores iniciales y aprovechar al máximo las muestras disponibles para la validación en la Fase 2b.

Debido a la cantidad de cepas que formaron parte del análisis y el proceso de validación, el nivel de confianza fue fijado como del 95%, la ventaja de la metodología presentada es la facilidad de repetitividad y reproducibilidad, por lo que es posible fijar un nivel de hasta 99%, sin embargo, requeriría un mínimo de 100 muestras como base en lugar de 20. Análisis que requieran niveles de confianza más elevados pueden ser realizados bajo la misma estrategia.

Los resultados que se obtienen de la metodología propuesta pueden actuar como soporte para el inicio o descarte de trabajos en laboratorio para la validación de pruebas de diagnóstico de varios fitopatógenos, considerando que si se obtienen e-probes de buen desempeño, es muy probable que estas se comporten de forma similar en muestras reales, mientras que, si son deficientes en la detección, es mucho menos probable que su desempeño sea bueno en comunidades experimentales reales.

Bibliografía

Agran, M. ., Di Terlizzi, B., Boscia, D., Minafra, A., Savino, V., Martelli, G. ., & Askri, F. (1990). Occurrence of grapevine virus A (GVA) and other clostroviruses in Tunisian grapevines affected by leafroll disease. *Vitis*, *29*, 43–48. https://doi.org/https://doi.org/10.5073/vitis.1990.29.43-48

Alabi, O. J., Al Rwahnih, M., Mekuria, T. A., & Naidu, R. A. (2014). Genetic diversity of Grapevine virus a in Washington and California vineyards. *Phytopathology*, *104*(5), 548–560. https://doi.org/10.1094/PHYTO-06-13-0179-R

Alabi, O. J., Martin, R. R., & Naidu, R. A. (2010). Sequence diversity, population genetics and potential recombination events in grapevine rupestris stem pitting-associated virus in Pacific North-West vineyards. *Journal of General Virology*, *91*(1), 265–276. https://doi.org/10.1099/vir.0.014423-0

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., & Tyson, G. W. (2012). Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, *40*(12). https://doi.org/10.1093/nar/gks251

Arora, A. K., Clark, N., Wentworth, K. S., Hesler, S., Fuchs, M., Loeb, G., & Douglas, A. E. (2020). Evaluation of rna interference for control of the grape mealybug pseudococcus maritimus (Hemiptera: Pseudococcidae). *Insects*, *11*(11), 1–15. https://doi.org/10.3390/insects11110739

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., & O'Grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, *33*(3), 296–302.

https://doi.org/10.1038/nbt.3103

Atallah, S. S., Gómez, M. I., Fuchs, M. F., & Martinson, T. E. (2012). Economic impact of grapevine leafroll disease on Vitis vinifera cv. Cabernet franc in Finger Lakes vineyards of New York. *American Journal of Enology and Viticulture*, *63*(1), 73–79. https://doi.org/10.5344/ajev.2011.11055

BCGSC. (2020). *NanoSim*. GitHub. https://github.com/bcgsc/NanoSim

Bonavia, M., Digiaro, M., Boscia, D., Boari, A., Bottalico, G., Savino, V., & Martelli, G. P. (1996). Studies on “corky rugose wood” of grapevine and on the diagnosis of grapevine virus B. *Vitis*, *35*(1), 53–58.

Borgo, M., Bazzo, I., Bertazzon, N., & Angelini, E. (2006). Sanitary controls for grapevine virus diagnosis [Vitis vinifera L.]. *Informatore Fitopatologico*, *56*(4), 15–17.

Boža, V., Brejová, B., & Vinař, T. (2017). DeepNano: Deep recurrent neural networks for base calling in MinION Nanopore reads. *PLoS ONE*, *12*(6), 1–13. https://doi.org/10.1371/journal.pone.0178751

CABI. (2020). *Detailed coverage of invasive species threatening livelihoods and the environment worldwide*. Invasive Species Compendium. https://www.cabi.org/isc/datasheet/26189#65ABB953-1B7D-4D2E-B1AB-6EBB984D7CF3

Cardwell, K., Dennis, G., Flannery, A. R., Fletcher, J., Luster, D., Nakhla, M., Rice, A., Shiel, P., Stack, J., Walsh, C., & Levy, L. (2018). Principles of diagnostic assay validation for plant pathogens: A basic review of concepts. *Plant Health Progress*, *19*(4), 272–278. https://doi.org/10.1094/PHP-06-18-0036-RV

Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., Lachman, O., Dror, O., Nissan, G., & Manulis-Sasson, S. (2019). Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology*, *68*(2), 229–238. https://doi.org/10.1111/ppa.12957

Chevalier, S., Greif, C., Clauzel, J. -M, Walter, B., & Fritsch, C. (1995). Use of an Immunocapture-Polymerase Chain Reaction Procedure for the Detection of Grapevine Virus A in Kober Stem Grooving-Infected Grapevines. *Journal of Phytopathology*, *143*(6), 369–373. https://doi.org/10.1111/j.1439-0434.1995.tb00277.x

Comeaux, B. L. (2013). *Taxonomy of the Native Grapes of North Carolina Author (s): Barry Lynn Comeaux , William B . Nesbitt and Paul R . Fantz Published by : Southern Appalachian Botanical Society Stable URL : http://www.jstor.org/stable/4033527 . Taxonomy of the Native Grape*. *52*(3), 197–215.

Crnogorac, A., Panno, S., Mandic, A., Gašpar, M., Caruso, A. G., Noris, E., Davino, S., & Matic, S. (2021). Survey of five major grapevine viruses infecting Blatina and Žilavka cultivars in Bosnia and Herzegovina. *PLoS ONE*, *16*(1 January), 1–20. https://doi.org/10.1371/journal.pone.0245959

David, M., Dursi, L. J., Yao, D., Boutros, P. C., & Simpson, J. T. (2017). Nanocall: An open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, *33*(1), 49–55. https://doi.org/10.1093/bioinformatics/btw569

Dicke, M. (2016). Plant phenotypic plasticity in the phytobiome: A volatile issue. *Current Opinion in Plant Biology*, *32*, 17–23. https://doi.org/10.1016/j.pbi.2016.05.004

Drake, J., & Holland, J. (1999). Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. USA*, *96*(24), 13910–13913. https://doi.org/10.1073/pnas.96.24.13910

Dumschott, K., Schmidt, M. H. W., Chawla, H. S., Snowdon, R., & Usadel, B. (2020). Oxford Nanopore sequencing: new opportunities for plant genomics? *Journal of Experimental Botany*, *71*(18), 5313–5322. https://doi.org/10.1093/jxb/eraa263

Engel, E. A., Escobar, P. F., Rojas, L. A., Rivera, P. A., Fiore, N., & Valenzuela, P. D. T. (2010). A diagnostic oligonucleotide microarray for simultaneous detection of grapevine viruses. *Journal of Virological Methods*, *163*(2), 445–451. https://doi.org/10.1016/j.jviromet.2009.11.009

Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, *17*(8), 459–469. https://doi.org/10.1038/nrg.2016.57

Espindola, A., Schneider, W., Hoyt, P. R., Marek, S. M., & Garzon, C. (2015). A new approach for detecting fungal and oomycete plant pathogens in next generation sequencing metagenome data utilising electronic probes. *International Journal of Data Mining and Bioinformatics*, *12*(2), 115–128. https://doi.org/10.1504/IJDMB.2015.069422

FAO. (2020). *World Food and Agriculture - Statistical Yearbook*. https://doi.org/https://doi.org/10.4060/cb1329en

Fellers, J., Webb, C., Fellers, M., Shoup, J., & De Wolf, E. (2019). Wheat Virus Identification Within Infected Tissue Using Nanopore Sequencing Technology. *Plant Disease*, *103*(9), 2199–2203.

Filloux, D., Fernandez, E., Loire, E., Claude, L., Galzi, S., Candresse, T., Winter, S., Jeeva, M. L., Makeshkumar, T., Martin, D. P., & Roumagnac, P. (2018). Nanopore-based detection and characterization of yam viruses. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-36042-7

Frith, M. C., Wan, R., & Horton, P. (2010). Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Research*, *38*(7). https://doi.org/10.1093/nar/gkq010

Fuller, K. B., Alston, J. M., & Golino, D. A. (2019). Economic benefits from virus screening: A case study of grapevine leafroll in the North Coast of California. In *American Journal of Enology and Viticulture* (Vol. 70, Issue 2). https://doi.org/10.5344/ajev.2018.18067

Galiakparov, N., Tanne, E., Sela, I., & Gafny, R. (2003). Functional analysis of the grapevine virus A genome. *Virology*, *306*(1), 42–50. https://doi.org/10.1016/S0042-6822(02)00019-3

Gambino, G., Navarro, B., Vallania, R., Gribaudo, I., & Di Serio, F. (2011). Somatic embryogenesis efficiently eliminates viroid infections from grapevines. *European Journal of Plant Pathology*, *130*(4), 511–519. https://doi.org/10.1007/s10658-011-9770-x

Garau, R., Prota, V. A., Boscia, D., Fiori, M., & Prota, U. (1995). Pseudococcus affinis Mask., new vector of grapevine trichoviruses A and B. *Vitis*, *34*(1), 67–68.

Goszczynski, D. E. (2015). Brief report of the construction of infectious DNA clones of South African genetic variants of grapevine virus A and grapevine virus B. *SpringerPlus*, *4*(1), 1–8. https://doi.org/10.1186/s40064-015-1517-2

Goszczynski, D. E., du Preez, J., & Burger, J. T. (2008). Molecular divergence of Grapevine virus A

(GVA) variants associated with Shiraz disease in South Africa. *Virus Research*, *138*(1–2), 105–110. https://doi.org/10.1016/j.virusres.2008.08.014

Goszczynski, D. E., & Habili, N. (2012). Grapevine virus A variants of group II associated with Shiraz disease in South Africa are present in plants affected by Australian Shiraz disease, and have also been detected in the USA. *Plant Pathology*, *61*(1), 205–214. https://doi.org/10.1111/j.1365-3059.2011.02499.x

Goszczynski, D. E., & Jooste, A. E. C. (2003). Identification of grapevines infected with divergent variants of Grapevine virus A using variant-specific RT-PCR. *Journal of Virological Methods*, *112*(1–2), 157–164. https://doi.org/10.1016/S0166-0934(03)00198-8

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J. M., Dodd, R., Mulembakani, P., Schneider, B. S., Muyembe-Tamfum, J. J., Stramer, S. L., & Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*(1), 1–13. https://doi.org/10.1186/s13073-015-0220-9

ICTV. (2020). *Betaflexiviridae*. ICTV 9th Report (2011). https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/241/betaflexiviridae

IndexBox. (2020). *Global Grape Market Overview*. IndexBox. https://doi.org/https://app.indexbox.io/report/080610/0/

Izard, J. (2015). Steps in Metagenomics: Let's Avoid Garbage in and Garbage Out. In *Metagenomics for Microbiology*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-410472-

3.00001-4

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 1–11. https://doi.org/10.1186/s13059-016-1103-0

Jiao, W. B., & Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, *36*, 64–70. https://doi.org/10.1016/j.pbi.2017.02.002

Klaassen, V. A., Sim, S. T., Dangl, G. S., Osman, F., Al Rwahnih, M., Rowhani, A., & Golino, D. A. (2011). Vitis californica and Vitis californica × Vitis vinifera hybrids are hosts for Grapevine leafroll-associated virus-2 and -3 and Grapevine virus A and B. *Plant Disease*, *95*(6), 657–665. https://doi.org/10.1094/PDIS-09-10-0621

La Notte, P., Buzkan, N., Choueiri, E., Minafra, A., & Martelli, G. P. (1997). Acquisition and transmission of grapevine virus a by the mealybug Pseudococcus longispinus. *Journal of Plant Pathology*, *79*(1), 79–85.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome

Assembly. In *Genomics, Proteomics and Bioinformatics* (Vol. 14, Issue 5). The Authors. https://doi.org/10.1016/j.gpb.2016.05.004

M'hirsi, S., Fattouch, S., Acheche, A., Marrakchi, M., & Marzouki, N. (2001). Detection of Grapevine A vitivirus in Tunisian grapevines. *Bulletin OEPP*, *31*(4), 509–513. https://doi.org/10.1111/j.1365-2338.2001.tb01035.x

Maliogka, V. I., Martelli, G. P., Fuchs, M., & Katis, N. I. (2015). Control of viruses infecting grapevine. In *Advances in Virus Research* (1st ed., Vol. 91, Issue 1). Elsevier Inc. https://doi.org/10.1016/bs.aivir.2014.11.002

Martelli, G. P. (2014). Fleck complex. *Journal of Plant Pathology*, *96*(1 SUPPL.), 1–4. https://doi.org/10.4454/jpp.v96i1sup.3143

Martinelli, L., Candioli, E., Costa, D., & Minafra, A. (2002). Stable insertion and expression of the movement protein gene of grapevine virus A (GVA) in grape (Vitis rupestris S.). *Vitis*, *41*(4), 189–193.

Melcher, U., Verma, R., & Schneider, W. L. (2014). Metagenomic search strategies for interactions among plants and multiple microbes. *Frontiers in Plant Science*, *5*(JUN), 1–5. https://doi.org/10.3389/fpls.2014.00268

Meng, B., Johnson, R., Peressini, S., Forsline, P. L., & Gonsalves, D. (1999). Rupestris stem pitting associated virus-1 is consistently detected in grapevines that are infected with rupestris stem pitting. *European Journal of Plant Pathology*, *105*(2), 191–199. https://doi.org/10.1023/A:1008771713839

Meng, B., Martelli, G. P., Golino, D. A., & Fuchs, M. (2017). Grapevine viruses: Molecular biology,

diagnostics and management. *Grapevine Viruses: Molecular Biology, Diagnostics and Management*, 1–698. https://doi.org/10.1007/978-3-319-57706-7

Milne, R. G., Conti, M., Lesemann, D. E., Stellmach, G., Tanne, E., & Cohen, J. (1984). Clostero virus-Like Particles of Two Types Associated with Diseased Grapevines. *Journal of Phytopathology*, *110*(4), 360–368. https://doi.org/10.1111/j.1439-0434.1984.tb00076.x

Monette, PL; James, D. (1990). Detection of two strains of grapevine virus A. *Plant Disease*, *74*(11), 898–900. https://doi.org/10.1094/PD-74-0898

Monette, P. L., & James, D. (1990). Use of in vitro cultures of Nicotiana benthamiana for the purification of grapevine virus A. *Plant Cell, Tissue and Organ Culture*, *23*(2), 131–134. https://doi.org/10.1007/BF00035833

Moore, M. O. (1991). Classification and systematics of eastern North American Vitis L. (VITACEAE) north of Mexico. *SIDA Contributions to Botany*, *14*(3), 339–367.

Naidu, R. A., Maree, H. J., & Burger, J. T. (2015). Grapevine Leafroll Disease and Associated Viruses: A Unique Pathosystem. *Annual Review of Phytopathology*, *53*, 613–634. https://doi.org/10.1146/annurev-phyto-102313-045946

Ng, T. F. F., Duffy, S., Polston, J. E., Bixby, E., Vallad, G. E., & Breitbart, M. (2011). Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. *PLoS ONE*, *6*(4). https://doi.org/10.1371/journal.pone.0019050

Nobile, M. S., Cazzaniga, P., Tangherloni, A., & Besozzi, D. (2017). Graphics processing units in bioinformatics, computational biology and systems biology. *Briefings in Bioinformatics*, *18*(5), 870–885. https://doi.org/10.1093/bib/bbw058

OIE. (2018a). Chapter 1.1.6. Principles and Methods of Validation of Diagnostic Assays for Infectious Diseases. In *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals* (pp. 72–87). World Organisation for Animal Health.

OIE. (2018a). Chapter 1.1.6. Principles and Methods of Validation of Diagnostic Assays for Infectious Diseases. Enero 2018 - Enero 2018[Tabla]. Recuperado de: In *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals* (pp. 72–87). World Organisation for Animal Health.

OIE. (2018b). Chapter 2.2.3. Development and optimisation of nucleic acid detection assays. In *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals* (pp. 195–205). World Organisation for Animal Health.

OIE. (2018c). Chapter 2.2.5. Statistical approaches to validation. In *Manual of Diagnostic Tests and Vaccines* (pp. 210–221). World Organisation for Animal Health.

OIE. (2018d). Chapter 2.2.7. Principles and methods for the validation of diagnostic tests for infectious diseases applicable to wildlife. In *Manual of Diagnostic Tests and Vaccines* (pp. 231–237). World Organisation for Animal Health.

Oxford Nanopore Technologies. (2020). *MinION*. Oxford Nanopore Technologies. https://nanoporetech.com/products/minion

Raman, T., & Muthukathan, G. (2015). Field suppression of Fusarium wilt disease in banana by the combined application of native endophytic and rhizospheric bacterial isolates possessing multiple functions. *Phytopathologia Mediterranea*, *54*(2), 241–252. https://doi.org/10.14601/Phytopathol

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, *19*(1), 1–11. https://doi.org/10.1186/s13059-018-1462-9

Reisch, B., & Pratt, C. (1996). *Grapes. In Fruit breeding. 2nd edition* (2nd ed.). Wiley.

Reske, A., Pollara, G., Krummenacher, C., Chain, B. M., & Katz, D. (2007). Understanding HSV-1 entry glycoproteins. *Reviews in Medical Virology*, *17*(1), 205–215. https://doi.org/10.1002/rmv

Ricketts, K. D., Gomez, M. I., Atallah, S. S., Fuchs, M. F., Martinson, T. E., Battany, M. C., Bettiga, L. J., Cooper, M. L., Verdegaal, P. S., & Smith, R. J. (2015). Reducing the economic impact of grapevine leafroll disease in california: Identifying optimal disease management strategies. *American Journal of Enology and Viticulture*, *66*(2), 138–149. https://doi.org/10.5344/ajev.2014.14106

Rosciglione, B., Castellano, M., Martelli, G., Savino, V., & Cannizzaro, G. (1983). Mealybug transmission of grapevine virus A. *Vitis*, *22*(4), 331–347.

Sabella, E., Pierro, R., Luvisi, A., Panattoni, A., D'Onofrio, C., Scalabrelli, G., Nutricati, E., Aprile, A., De Bellis, L., & Materazzi, A. (2018). Phylogenetic analysis of viruses in tuscan vitis vinifera sylvestris (Gmeli) hegi. *PLoS ONE*, *13*(7), 1–16. https://doi.org/10.1371/journal.pone.0200875

Saldarelli, P., Montano, H. G., & Martelli, G. P. (1994). Non-radioactive molecular probes for the detection of three filamentous viruses of the grapevine. *Vitis*, *33*(3), 157–160.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing.

Human Molecular Genetics, *19*(R2), 227–240. https://doi.org/10.1093/hmg/ddq416

Stobbe, A. H., Schneider, W. L., Hoyt, P. R., & Melcher, U. (2014). Screening metagenomic data for viruses using the E-probe diagnostic nucleic acid assay. *Phytopathology*, *104*(10), 1125–1129. https://doi.org/10.1094/PHYTO-11-13-0310-R

Stobbe, Anthony H., Daniels, J., Espindola, A. S., Verma, R., Melcher, U., Ochoa-Corona, F., Garzon, C., Fletcher, J., & Schneider, W. (2013). E-probe Diagnostic Nucleic acid Analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. *Journal of Microbiological Methods*, *94*(3), 356–366. https://doi.org/10.1016/j.mimet.2013.07.002

Stoiber, M., & Brown, J. (2017). BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal. *BioRxiv*. https://doi.org/10.1101/133058

This, P., Lacombe, T., & Thomas, M. R. (2006). Historical origins and genetic diversity of wine grapes. *Trends in Genetics*, *22*(9), 511–519. https://doi.org/10.1016/j.tig.2006.07.008

Tsai, C. W., Chau, J., Fernandez, L., Bosco, D., Daane, K. M., & Almeida, R. P. P. (2008). Transmission of Grapevine leafroll-associated virus 3 by the Vine Mealybug (Planococcus ficus). *Phytopathology*, *98*(10), 1093–1098. https://doi.org/10.1094/PHYTO-98-10-1093

Visser, M., Burger, J. T., & Maree, H. J. (2016). Targeted virus detection in next-generation sequencing data using an automated e-probe based approach. *Virology*, *495*, 122–128. https://doi.org/10.1016/j.virol.2016.05.008

Vlahovic, B., Potrebic, V., & Jelocnik, M. (2012). Preferences of Wine Consumers on Serbian Market. *Economics of Agriculture*, *59*(1), 37–49.

Wick, R; Judd, L; Holt, K. (2018). *Comparison of Oxford Nanopore basecalling tools*. Zenodo. https://zenodo.org/record/1188469#.YDGo2439Zz1

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *BioRxiv*, 1–10. https://doi.org/10.1101/543439

Yang, C., Chu, J., Warren, R. L., & Birol, I. (2017). NanoSim: Nanopore sequence read simulator based on statistical characterization. *GigaScience*, *6*(4), 1–6. https://doi.org/10.1093/gigascience/gix010

Young, K., Lahmers, K., Sellers, H., Stallknecht, D., Poulson, R., Saliki, J., Tompkins, S., Padykula, I., Siepker, C., Howerth, E., Todd, M., & Stanton, J. (2019). Randomly primed, strand-switching MinION-based sequencing for the detection and characterization of cultured RNA viruses. *Journal of Veterinary Diagnostic Investigation*.

Zeng, J., Cai, H., Peng, H., Wang, H., Zhang, Y., & Akutsu, T. (2020). Causalcall: Nanopore Basecalling Using a Temporal Convolutional Network. *Frontiers in Genetics*, *10*(January), 1–11. https://doi.org/10.3389/fgene.2019.01332

Zeng, J., Cai, H., Peng, H., Wang, H., Zhang, Y., & Akutsu, T. (2020). Causalcall: Nanopore Basecalling Using a Temporal Convolutional Network, 20 Enero 2020 - 20 Enero 2020[Tabla]. Recuperado de: *Frontiers in Genetics*, *10*(January), 1–11.

Zhang, P., Ganesamoorthy, D., Nguyen, S. H., Au, R., Coin, L. J., & Tey, S. K. (2020). Nanopore sequencing as a scalable, cost-effective platform for analyzing polyclonal vector integration sites following clinical T cell therapy. *Journal for Immunotherapy of Cancer*, *8*(1), 1–13. https://doi.org/10.1136/jitc-2019-000299

Anexos