PEDECIBA BIOINFORMÁTICA

“Variabilidad genética y mecanismos evolutivos de virus ARN: aproximación al análisis de cuasiespecies”

Martín Sóñora

Tesis de Maestría

Laboratorio de Virología Molecular
Centro de Investigaciones Nucleares
Facultad de Ciencias
Universidad de la República

2018

Orientadores: Dr. Gustavo Guerberoff y Dr. Fernando Alvarez
Co-orientador: Dr. Juan Cristina

Tribunal: Dr. Rodney Colina
Dr. Sergio Pantano
Dr. Pablo Smircich

ÍNDICE

Resumen

Los virus exploran todos los mecanismos conocidos para la generación de variabilidad genética. Entre estos encontramos a la mutación como fuerza principal, sin embargo los virus también se valen de la recombinación y el reordenamiento para la producción de variabilidad génica. Este último mecanismo se puede apreciar en virus ARN con genoma segmentado, como es el caso del virus de la gripe. Es más que conocido que los virus ARN circulan como un complejo enjambre de mutantes genéticamente relacionadas. Dicho conjunto de variantes es denominado como cuasiespecie. La dinámica evolutiva de las cuasiespecies virales es el producto de las características intrínsecas de los virus ARN. Las altas tasas de mutación, los pequeños tamaños genómicos, los cortos tiempos generacionales y los grandes tamaños poblacionales le confiere a los virus ARN, y a algunos virus ADN, un potencial adaptativo sorprendente, representando la principal dificultad para la prevención y tratamiento de enfermedades asociadas a estos patógenos. El entendimiento de las relaciones entre los individuos de una población viral y su evolución es de gran utilidad para afrontar los eventos epidémicos. El estudio de la dinámica poblacional y los patrones evolutivos de las poblaciones virales es crucial para la comprensión de los procesos evolutivos. Profundizar en el estudio de los componentes de una población viral intra hospedero es fundamental para evidenciar posibles cepas resistentes a antivirales o que resulten muy patogénicas.

Introducción General

VARIABILIDAD GENÉTICA VIRAL

Los virus, cuyo genoma está constituido por ARN, exploran todos los mecanismos conocidos para la generación de variabilidad genética. La mutación puntual, así como la recombinación génica, constituyen dos de los mecanismos más importantes involucrados durante este proceso. Sin embargo, dentro del mundo de los virus también se encuentran aquellos con genoma segmentado. Estos explotan, además, otro de los mecanismos para la generación de variabilidad genética, el reordenamiento.

Esta sección trata de explicar de forma general los principales mecanismos involucrados en la generación de variabilidad genética viral. Se tocan como temas centrales: mutación, recombinación y reordenamiento.

Mutación:

La mutación es un mecanismo fundamental para la generación de diversidad genética y es la fuente de poder principal para la evolución de cualquier organismo. Con respecto a los virus ARN, la mutación puntual es principalmente atribuida a la falta de actividad exonucleasa 3'→5' correctora de errores de las ARN polimerasas ARN dependientes [Nobusawa & Sato, 2006; Sanjuán y cols., 2010]. Esta carencia lleva a una tasa de error del orden de 10^{-3} a 10^{-5} sustituciones por nucleótido copiado, mientras que las tasas que presentan las ADN polimerasas se estiman en el orden de 10^{-8} a 10^{-11} [Holland y cols., 1982; Meyerhans & Vartanian, 1999; Vignuzzi y cols., 2005].

Estas elevadas tasas de mutación conducen a la generación de un gran número de variantes virales en el transcurso de una infección [Domingo, Sheldon & Perales, 2012]. Muchos de estos cambios son sinónimos (también conocido como mutaciones silenciosas), por lo que no presentan impacto directo a nivel de la secuencia aminoacídica. Sin embargo, estas mutaciones pueden afectar la estructura secundaria

del ARN viral. Además, también pueden impactar a nivel fenotípico, directamente actuando sobre la traducción de proteínas. Más precisamente, existe evidencia de que el uso de diferentes codones sinónimos puede alterar la cinética de la síntesis proteica y como consecuencia su plegamiento [Goymer, 2007].

Otro tipo de mutaciones son las denominadas no sinónimas. Estas son responsables de los cambios en la secuencia primaria de aminoácidos, y por consiguiente, de la generación de nuevas variantes virales. Asimismo, estos cambios pueden también llevar a la generación de genomas defectivos o letales [Le Guillou-Guillemette y cols., 2007]. Todo este universo de variantes, generadas en el curso de una infección viral, es lo que constituye la nube de mutantes también conocida como cuasiespecie viral [Eigen & Schuster, 1978a]; concepto que se abordará más adelante. De esta forma es como circulan en la naturaleza los virus ARN; como una constelación dinámica de mutantes inter-relacionados que cooperan y/o compiten, y en donde la unidad de selección es la población en su conjunto y no una variante en forma individual [Domingo y cols., 2006; Vignuzzi y cols., 2006; Domingo & Gomez, 2007].

Recombinación:

La existencia de diferentes poblaciones virales dentro de un mismo hospedero, ya sea subpoblaciones de una misma cuasiespecie o cepas de diferentes genotipos que coexistan dentro de una misma célula, posibilitan la generación de variabilidad mediante la recombinación genética. Este mecanismo consiste en la generación de nuevos mutantes a partir de la combinación genética de variantes parentales distintas [Worobey & Holmes, 1999]. En virus con genoma ARN la recombinación puede ocurrir mediante un mecanismo de salto de molde, mayormente conocido como “copy choice” [Coffin JM, 1979; Lai MM, 1992]. Se piensa que su mecanismo de acción sigue la siguiente lógica: cuando al menos dos virus diferentes infectan una misma célula la enzima ARN polimerasa ARN dependiente viral, que se encuentra polimerizando nuevos ARN genómicos, se disocia del primer genoma parental y continúa replicando utilizando un molde genómico distinto. Este salto de molde resulta en la generación de genomas tipo mosaico con regiones de ambos parentales [Colina R y cols., 2004; Worobey y Holmes, 1999; Simon-Loriere y Holmes, 2011], tal como se vio para algunas formas recombinantes circulantes en el caso de Virus de la Inmunodeficiencia Humana

(VIH) [Lau y Wong, 2013]. Esto puede dar lugar a variantes difícilmente diferenciables en el caso de recombinación de variantes parentales genéticamente muy similares, como lo son los casos de recombinación intra-subtipo, o más complejo aún, intra-cuasiespecie [Moreno MP y cols., 2006; Sentandreu V y cols., 2008; Palmer BA y cols., 2012; Jacobi & Nordahl, 2006].

La recombinación puede ocurrir tanto en los virus segmentados como en los no segmentados, aunque existen reportes donde se discute fuertemente los casos de recombinación en virus con genoma segmentado debido a la presencia de artefactos de laboratorio y/o bioinformáticos [Boni MF y cols., 2008; Boni MF y cols., 2010]. Por otro lado, cuando la recombinación ocurre entre variantes de distinto genotipo como se ha visto en el caso del Virus de la Hepatitis C (VHC), este mecanismo podría dar origen a cepas con una capacidad replicativa, transmisión y virulencia diferente a la de sus parentales. En relación a este virus, la recombinación ha sido reportada tanto intra como inter genotipos. Reportes previos describen algunas cepas del VHC provenientes de Honduras, en donde las secuencias parciales correspondientes a regiones distantes del genoma viral resultaron de genotipos diferentes, lo que probó la primer evidencia de posible recombinación en el VHC [Yun Z y cols., 1996]. Sin embargo, no fue hasta el año 2002 que se publicó el primer reporte convincente de una cepa recombinante del VHC intergenotipo [González-Candelas F. y cols., 2011; Kalinina O y cols., 2002]. Los resultados de estos y otros estudios apoyan el hecho de que la recombinación no puede ser negada como mecanismo evolutivo para la generación de diversidad en virus con genoma ARN y en particular en el VHC [Moreno MP y cols., 2006; Colina R y cols., 2004; Moreno MP y cols., 2009]. Por todo lo antes mencionado, la recombinación es un mecanismo efectivo de generación de variabilidad, que brinda gran plasticidad y capacidad adaptativa frente a diferentes condiciones, tanto del hospedero (evasión del sistema inmunitario), como frente al tratamiento con drogas antivirales.

Reordenamiento:

Por último, pero no menos importante, el reordenamiento viral o más comúnmente conocido simplemente como reordenamiento o reassortment, por su nombre en inglés, es el proceso de intercambio de información genética exclusivo de virus con genoma ARN segmentado. Al igual que la recombinación, en el reordenamiento la coinfección

de una misma célula con múltiples virus diferentes es un pre-requisito. Esto podría conducir a la mezcla de segmentos génicos dando como resultado una progenie de virus que muestre una combinación genómica novedosa [Marshall N y cols., 2013]. El reordenamiento ha sido observado en miembros de todas las familias de virus con genoma segmentado, incluyendo por ejemplo al virus Bluetongue [Batten CA y cols., 2008], pero sin lugar a duda este mecanismo es mayormente descripto para el virus Influenza A (VIA) como uno de los principales mecanismos vinculados a la transmisión inter especie y en la emergencia de cepas pandémicas [Webster RG y cols., 1992; Smith GJD y cols., 2009a; Neumann G y cols., 2009]. Es bastante intuitivo ver cómo el reordenamiento acelera la tasa de adquisición de marcadores genéticos. La aparición de nuevos genes de influenza en la población humana, y su establecimiento posterior para causar pandemias, se ha vinculado con el reordenamiento de segmentos génicos [Smith GJD y cols., 2009b; Neumann G y cols., 2009; Smith GJD y cols., 2009a]. Una de las ventajas principales que brinda la segmentación del genoma a los virus, y en particular a los Virus de Influenza, es la alta probabilidad de sufrir reordenamientos genéticos. Esto último contribuye a la gran variabilidad inmunológica de los principales determinantes antigénicos de superficie, Hemaglutinina (HA) y Neuraminidasa (NA), de los virus del tipo A [Hillerman MR, 2002; De Jong y cols., 2000]. Cabe destacar que actualmente hay descriptos 18 y 11 subtipos diferentes de HA y NA respectivamente [Tong y cols., 2013]. El salto antigénico ocurre cuando el virus adquiere una HA y/o NA pertenecientes a un subtipo de VIA diferente, mediante el proceso denominado rearreglo génico o reordenamiento. Estos constituyen cambios mayores en la estructura de los antígenos de superficie, lo que produce subtipos completamente nuevos de virus, ante los cuales la población tiene poca o ninguna inmunidad [Ferguson y cols., 2003; McHardy & Adams, 2009]. Es de resaltar, que este mecanismo ha sido el instrumento requerido en los principales eventos de transmisión de VIA endémicos de cerdos y aves, hacia humanos. Esto es ilustrado por ejemplo en las pandemias de Asia de 1957 y Hong Kong de 1968, que fueron ambas asociadas con virus reordenantes, e involucraron segmentos génicos de virus circulantes en humanos y aves [Scholtissek y cols., 1978; Kawaoka y cols., 1989]. De forma similar la pandemia del año 2009 resultó del reordenamiento entre virus porcinos muy divergentes. Involucró virus de cerdos americanos con virus de cerdos del linaje eurásico [Neumann y cols., 2009; Garten y cols., 2009]. En estas tres pandemias, el evento de reordenamiento resultó en un nuevo virus humano. Estos nuevos virus humanos expresan los antígenos HA y NA, siendo poco reconocidos por el sistema inmune

adaptativo. Por otro lado, el reordenamiento entre cepas de influenza co-circulantes en humanos, con el mismo subtipo de HA y NA, también es importante para la evolución y emergencia de las cepas estacionales de influenza A [Nelson MI y cols., 2007]. Esto incluye a cepas que son antigénicamente nuevas [Rambaut A, y cols., 2008], aquellas con transmisión potenciada [Nelson MI y cols., 2008] y a las resistentes a drogas antivirales [Simonsen L y cols., 2007]. Nuevas pandemias de virus de la Influenza A pueden emerger a través de este proceso de reordenamiento génico, entre cepas de virus que circulan en reservorios aviares y/o suinos [Taubenberger y cols., 2006].

CUASIESPECIES VIRALES

Características generales:

Como mencionamos brevemente en la sección anterior, el término cuasiespecie viral refiere a un conjunto de variantes virales estrechamente relacionadas desde un punto de vista genético [Andino & Domingo, 2015]. De hecho, las altísimas tasas de mutación y el gran tamaño poblacional de las cuasiespecies, permiten a los virus con genoma ARN evolucionar y adaptarse a cambios y desafíos durante una infección [Biebricher & Eigen, 2006]. Por consiguiente, el blanco de selección de los eventos evolutivos es la población viral en su conjunto, y no una estirpe viral de forma individual. La evolución de las cuasiespecies es el resultado de las características intrínsecas de los virus ARN. Entre ellas se destacan: las altas tasas de mutación, los pequeños tamaños genómicos, los cortos tiempos generacionales y los tamaños poblacionales significativamente grandes.

Como fue mencionado anteriormente, los mecanismos de generación de variabilidad como la mutación, recombinación y el reordenamiento de segmentos génicos juegan un papel muy importante en su evolución. Muchas de las características biológicas de estas poblaciones, tales como la de cambiar su tropismo celular, ampliar su rango de hospedero o evadir presiones selectivas, subyacen en el repertorio de variantes virales [Domingo y cols., 2012]. Esta dinámica poblacional le confiere a los virus ARN, y a algunos virus ADN, un potencial adaptativo sorprendente, representando la principal

dificultad para la prevención y tratamiento de enfermedades asociadas a estos patógenos.

Conceptos básicos de la teoría de las cuasiespecies:

La organización genética de las poblaciones virales puede ser representada utilizando el concepto de “espacio de secuencia”. En términos generales, el espacio de secuencia es una representación matemática de todas las posibles secuencias (permutaciones) de un genoma. Por ejemplo, supongamos que partimos de un genoma viral único: tras sucesivos ciclos de replicación se producirán miles de millones de “genomas hijos” que se diferenciarán del “genoma parental” en aproximadamente una o más posiciones. Asimismo, a medida que la población se replica la distribución de variantes incrementará en su grado de complejidad [Lauring & Andino, 2010] (Figura 1).

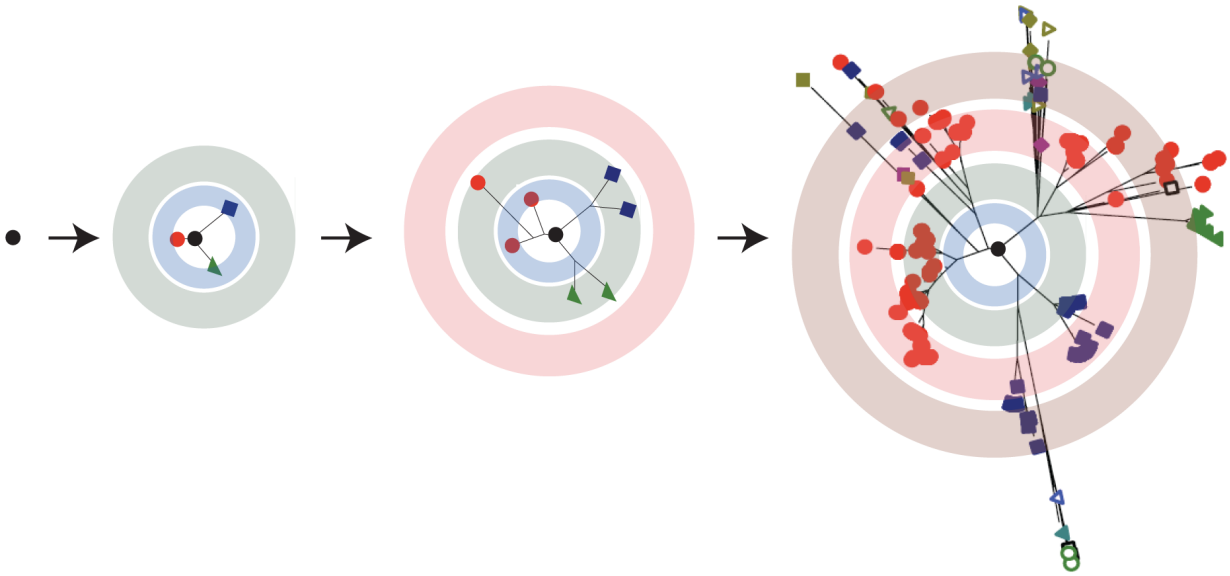


Figura 1. Replicación de virus con genoma ARN.
En el curso de una infección viral los virus ARN generan un repertorio de mutantes muy diverso en unos pocos ciclos de replicación. Los círculos concéntricos representan cada ronda de duplicación mientras que cada rama de los árboles indican variantes virales determinadas por ciertas mutaciones puntuales. La distribución resultante es representada como una nube que se centra en una secuencia maestra.
Tomado de Lauring A, Andino R. (2010).

De acuerdo con la teoría de genética de poblaciones, la frecuencia de una variante en una población puede ser aproximada a su capacidad de sobrevivir y replicar en un ambiente determinado, es decir al “fitness” o eficacia biológica de ese individuo [Lauring & Andino, 2010]. En el marco de las cuasiespecies, una variante de bajo “fitness” puede ser mantenida a mayor frecuencia debido a que puede estar asociada a una variante con mayor “fitness”. Este comportamiento es conocido como fenómeno de acoplamiento mutacional, una de las características que definen las cuasiespecies virales [Domingo E. 2005; Briones y cols., 2006].

Una posible representación de las poblaciones virales puede realizarse gracias al concepto de paisaje adaptativo o “fitness landscape” por su nombre en inglés (Figura 2A). En biología evolutiva, los paisajes adaptativos se utilizan para visualizar las relaciones entre genotipos y fenotipos (más precisamente éxito reproductivo). En otras palabras, el concepto de fitness landscape es muy útil para visualizar el efecto de la epistasis en las trayectorias adaptativas. Introducido por Sewall Wright, el paisaje adaptativo es un medio para visualizar el mapa genotipo-fitness [Wright, 1932].

Los conceptos de altura y distancia son suficientes para conformar el concepto de paisaje adaptativo:

-Altura: la altura del paisaje adaptativo es atribuida a la eficacia o éxito reproductivo (para nuestro enfoque será el fenotipo).

-Distancia: en el eje horizontal se representan las frecuencias de los diferentes genotipos de la población así como también toda posible secuencia (espacio de secuencia).

Sin embargo, el fitness landscape se ilustra clásicamente como un paisaje montañoso tridimensional, en el que el espacio genotípico se representa en el plano x-y y el fitness se representa en el eje z [Wright, 1932].

En un paisaje adaptativo, los genotipos que son muy similares están muy cerca el uno del otro, mientras que aquellos más diferentes están muy lejos entre sí. La adaptación de una cuasiespecie a nuevos ambientes implica la exploración del espacio de secuencia de un punto a otro del paisaje adaptativo abarcando picos y valles (Figura 2B) [Lauring & Andino, 2010]. Consideremos el siguiente ejemplo hipotético: existen dos poblaciones, una generada por replicadores rápidos y la otra por replicadores lentos. En un ambiente de baja tasa mutacional, el replicador rápido “triunfará” debido a que su descendencia será en mayor medida genéticamente idéntica y será generada más rápidamente. Esta población ocupará regiones de picos altos y estrechos del “fitness landscape” donde existe poca diversidad genética y “fitness” máximo [Lauring

& Andino, 2010] (Figura 2A). La teoría de las cuasiespecies predice que los replicadores lentos se verán favorecidos si dan surgimiento a progenies que en promedio presenten mayor “fitness” que su parental. Estas poblaciones ocuparán regiones chatas y pequeñas del “fitness” landscape [Wilke, 2005] (Figura 2A). Este efecto ha sido denominado “supervivencia del más chato”. Una cuasiespecie “chata” será más flexible e incorporará mutaciones sin un efecto aparente en su “fitness”. En otras palabras, esta población será más robusta desde un punto de vista mutacional, ya que amortiguará los cambios a nivel genotípico, sin grandes variaciones en su fenotipo. Una cuasiespecie “chata”, con una extensa nube de mutantes, podrá explorar en mayor profundidad el espacio de secuencia adaptándose en mayor medida a los cambios del ambiente. Esto puede explicar muchos de los fenómenos observados con implicancia clínica directa. A modo de ejemplo, los Arbovirus (término utilizado para referirse a cualquier virus que es transmitido por un vector artrópodo) deben adaptarse a hospederos invertebrados (por ejemplo los insectos como el mosquito *Aedes Aegypti*) y a vertebrados (en el caso de los mamíferos), por ende modular su paisaje adaptativo. Siguiendo con estos ejemplos, estudios retrospectivos que utilizan aislados primarios del VIH, sugieren que la cuasisespecie de estos virus podría estar explorando regiones chatas y de menor “fitness” en el espacio de secuencia [Quinones-Mateu y cols., 2006; Ariën y cols., 2005; Richman DD., 2006; Rolland y cols., 2007]. En el caso del Virus Influenza A, un mapeo detallado en uno de los principales determinantes antigénicos del virus, la glicoproteína HA, sugiere que las cepas interpandémicas permanecen antigénicamente estables por años. Esta acumulación estable de diversidad genética es puntuada por cambio en antigenicidad [Koelle y cols., 2006].

Por todo lo mencionado anteriormente, una población bien adaptada a un ambiente determinado replicará más rápidamente que otras menos adaptadas, poblando picos altos y estrechos del “fitness” landscape, mientras que las poblaciones menos adaptadas, aunque genéticamente más diversas, habitarán picos más bajos y menos estrechos [Lauring & Andino, 2010].

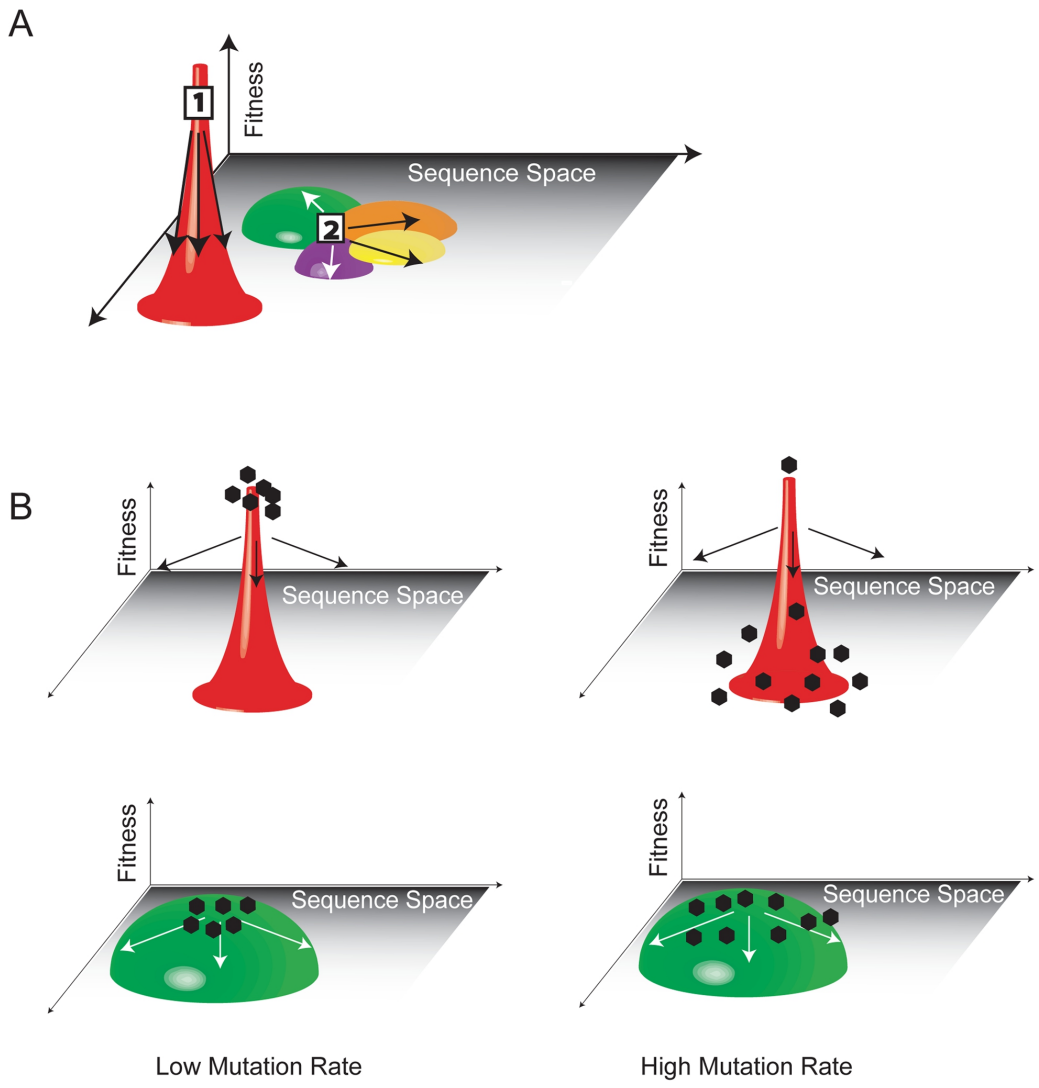


Figura 2. Paisaje adaptativo y sobrevivencia del “flattest”

Representación esquemática del fitness landscape de una población viral hipotética.

A) La población 1 tiene alto fitness aunque se encuentra “atrapada” en el espacio de secuencia ya que las mutaciones llevan a una pérdida de la capacidad replicativa. Por otro lado, la población 2 es mutacionalmente robusta dado que la mutación lleva a una menor pérdida del fitness. Se sitúa idealmente para moverse a través del espacio de secuencia y acceder a otros picos locales a través de redes mutacionales vecinas (indicado por los colores verde, amarillo, naranja y violeta). B) En un ambiente con bajas tasas de mutación, las variantes serán genotípicamente estables y se agruparan en la parte superior del pico de fitness. La variante con mayor fitness superará fácilmente a todas las demás. A altas tasas de mutación las variantes se dispersan sobre los picos correspondientes. Las variantes del pico más plano (verde) permanecen cerca de su fitness óptimo y tienen un fitness medio más alto que la población localizada en el pico más empinado (rojo). Prevalece la población más plana. Tomado de Lauring A, Andino R. (2010).

Otra característica fundamental de las cuasiespecies virales es la presencia de una distribución ordenada y estable de mutantes dominados por un genoma principal. Dicho genoma se denomina “secuencia maestra” y puede ser minoritario en una población de replicones, en cuyo caso la población está formada esencialmente por una constelación de mutantes [Domingo y cols., 2012]. Cada uno de los genomas que componen la cuasiespecie, presenta una eficacia biológica (capacidad replicativa) distinta, siendo éste el parámetro que determina su representación en la población viral. La secuencia con mayor eficacia biológica es lo que se denomina secuencia maestra. El resto de los genomas constituyen lo que se llama el espectro de mutantes de la cuasiespecie. Dado que los métodos de secuenciación clásicos no son lo suficientemente sensibles para poder detectar secuencias que estén en bajas proporciones, la secuencia que se determina experimentalmente es la promedio o consenso, que corresponde a los nucleótidos que están más representados en cada posición de la secuencia genómica del ARN viral [Lázaro & Homs, 2002].

Formulación del marco teórico de las cuasiespecies virales:

El concepto de cuasiespecie tuvo su origen con la formulación matemática de la evolución molecular. Ésta hizo hincapié en la replicación propensa a errores de moléculas de ARN simple hebra o replicones tipo ARN, como una de las características esenciales de la auto organización y adaptabilidad de las formas de vida primitivas. [Eigen, 1971]. Manfred Eigen fue pionero en el estudio teórico de la evolución de las macromoléculas biológicas. Eigen integró conceptos de la teoría de la información junto con conceptos de la selección natural darwiniana. Según su teoría, una molécula auto-replicativa denominada “secuencia maestra”, produce variantes con cierta distribución de probabilidad. Esta distribución dependerá de ciertos factores de calidad. Éstos determinan la fracción de copias que llevan a una réplica exacta del molde y la fracción que conduce a copias del molde con errores. Eigen utilizó el termino “cola de cometa” para referir a copias del molde con error. Posteriormente se introdujo el termino “cuasiespecies” y “espectro de mutantes”, usados comúnmente en virología [Andino & Domingo, 2015; Eigen & Schuster, 1977; Eigen & Schuster, 1978b]. La formulación de la teoría de las cuasiespecies refiere a otras formulaciones de la evolución darwiniana, pero su principal característica distintiva es el énfasis puesto en la ocurrencia de mutaciones durante la replicación [Page & Nowak, 2002]. Además, para definir en términos cuantitativos un sistema replicativo con altas tasas de mutación, la teoría de cuasiespecies establece una condición necesaria para asegurar

la conservación estable de la información genética. Esta condición fue formulada como umbral de error, básicamente significa que para cualquier complejidad de la nube hay una tasa de error máxima compatible con el mantenimiento de la información genética. El valor del umbral de error depende de la precisión en la replicación y del “fitness” de la secuencia maestra. En síntesis, el umbral de error corresponde a la tasa de mutación promedio (por sitio y replicación) a la cual la frecuencia de la secuencia dominante resulta muy pequeña y desaparece a efectos prácticos [Page & Nowak, 2002].

Repensando el concepto de memoria molecular en las cuasiespecies virales:

Según varios virólogos, la selección positiva y la memoria molecular son características determinantes de las cuasiespecies virales. La selección positiva implica la replicación de ciertos componentes de la nube de mutantes con alguna característica seleccionable, como ser la presencia de una mutación de resistencia a un antiviral. Sumado a esto, y dado que la replicación generalmente conlleva a una ganancia en el fitness [Escarmís, Dávila, & Domingo, 1999; Lorenzo-Redondo, Bordería, & Lopez-Galindez, 2011; Novella y cols., 1995], la frecuencia de la subpoblación aumentará. Según Ruiz-Jarabo y colaboradores, este escenario de replicación bajo selección positiva, además de llevar a un aumento en la capacidad replicativa, permitiría la perpetuación de la subpoblación como “genomas memoria” [C M Ruiz-Jarabo, Arias, Baranowski, Escarmís, & Domingo, 2000]. Esteban Domingo y colaboradores han planteado que el espectro de genomas en una cuasiespecie viral pueden contener variantes que representen registros “fósiles” o huellas de aquellos genomas que fueron dominantes en fases previas del linaje evolutivo [Esteban Domingo y cols., 2012].

Según ciertas publicaciones, parecería que la memoria molecular en poblaciones virales es pura consecuencia de la dinámica de cuasiespecies. En primer lugar la memoria cae (o directamente es eliminada) bajo la fuerza purificadora del efecto fundador o también conocido como cuello de botella. La acción de este último va de la mano con el hecho de que la memoria es una característica de la población como un todo y no de genomas individuales [E Domingo, 2000; C M Ruiz-Jarabo y cols., 2000]. Por otro lado la frecuencia de un genoma memoria en una población dependerá del fitness relativo del virus que será convertido en memoria. Hay estudios que muestran cómo ciertos mutantes con un fitness inicial alto alcanzan frecuencias mayores cuando

dicho mutante se convierte en memoria [Ruíz-Jarabo y cols., 2002]. Otro atributo de los genomas memoria es que estos ganan capacidad replicativa en paralelo con los genomas más frecuentes de una cuasiespecie replicativamente activa [Arias, Ruiz-Jarabo, Escarmís, & Domingo, 2004]. Por último pero no menos importante, el nivel de memoria puede continuar en el tiempo pero va disminuyendo gradualmente en poblaciones replicativas. Este hecho se ve reflejado en trabajos previos que sugieren un comportamiento determinístico de las poblaciones virales. Esto se debe a que la pérdida en memoria sigue una tendencia idéntica en linajes paralelos [Ruiz-Jarabo, Miller, Gómez-Mariano, & Domingo, 2003].

No sólo los estudios experimentales apoyan esta teoría de los genomas memoria, sino que también es fuertemente apoyada por estudios teóricos de comportamiento de la evolución de las cuasiespecies [Briones, de Vicente, Molina Paris, & Domingo, 2006; Briones, Domingo, & Molina Paris, 2003]. Según estas y otras publicaciones, de alguna forma, los genomas memoria adscriben a las poblaciones virales la capacidad de responder a las presiones selectivas sufridas con anticipación por el mismo linaje evolutivo [Briones & Domingo, 2008; E. Domingo, 2000; Perales, Lorenzo, López Galíndez, Martínez, & Domingo, 2010]. El uso de drogas antivirales para el tratamiento de un agente infeccioso es el ejemplo más claro en la dinámica de obtención de memoria. Se espera que opere *in vivo* mientras un mutante resistente es seleccionado y puesto a competir con otros componentes del espectro de variantes. La evolución hacia un genoma memoria se podría alcanzar mediante la interrupción del tratamiento o por la utilización de un tratamiento alternativo. Esto ha sido evidenciado en el caso de HIV-1 y es esperado que ocurra en el caso de nuevos agentes antivirales para el tratamiento de infecciones virales crónicas [Briones y cols., 2006; Domingo y cols., 2012].

En este sentido, la administración de ciertas drogas antivirales a pacientes que portan una población viral con cepas resistentes, podría ser evitada con el fin de atenuar la selección de los mutantes resistentes y por ende ampliar el beneficio del tratamiento. Las tecnologías de secuenciación de nueva generación, como se discutirá en la siguiente sección de esta introducción, abren la posibilidad de estudiar en mayor profundidad las variantes presentes dentro de un paciente infectado. Esto ayuda, entre otras cosas, a la identificación de variantes minoritarias resistentes a los antivirales. De esta forma es posible colaborar en la mejora del tratamiento así como en la calidad de vida de los pacientes.

ADVENIMIENTO DE LAS TECNOLOGÍAS DE SECUENCIACIÓN MASIVA Y SU APLICABILIDAD EN VIROLOGÍA AL ESTUDIO DE CUASIESPECIES VIRALES

La diversidad genética y la dinámica evolutiva de las poblaciones virales fue estudiada, muy recientemente, basándose en el secuenciado de genomas consenso. Mediante este abordaje, sólo es posible identificar la secuencia viral principal o dominante presente en una muestra, lo cual resulta poco informativo acerca del espectro de mutantes de variantes minoritarias en la población [Morelli y cols., 2013; Wright y cols., 2011]. Las secuencias de genomas consenso dan una visión incompleta sobre la evolución y diversidad viral intra e inter hospedero [Schönherz y cols., 2016].

El clonado de los productos de PCR y su subsecuente secuenciación por el método de Sanger han sido ampliamente utilizados en el análisis de la heterogeneidad de las cuasiespecies virales. No obstante, este procedimiento es extremadamente costoso y laborioso cuando se quieren analizar cientos de muestras [Beerenwinkel & Zagordi, 2011]. En contraste, las tecnologías de secuenciación de próxima generación (del inglés “Next-generation Sequencing”, NGS) brindan un análisis muchísimo más profundo de la diversidad en poblaciones virales. De esta manera se obtienen resultados en un tiempo y a un costo considerablemente menor que las técnicas de secuenciación clásicas [Borucki y cols., 2013; Eriksson y cols., 2008; Kampmann y cols., 2011; Morelli y cols., 2013].

Los estudios de secuenciación masiva se caracterizan por secuenciar de cientos a miles de veces una posición genómica determinada, lo que generalmente se denomina cobertura o coverage por su nombre en inglés. La misma puede ser calculada de forma aproximada a partir del largo del gen (G), el número de lecturas/reads (N) que alinean con la referencia y el largo de read promedio (L). Esta enorme profundidad posibilita investigar en mayor detalle variantes presentes en la muestra que se encuentren a una frecuencia ínfimamente pequeña (por debajo de 1%). Sin embargo, debido a que la tasa de error asociada a las plataformas de NGS caen en ese orden de magnitud, resulta extremadamente desafiante distinguir los errores de secuenciación de las variantes verdaderas. En este sentido, es importante tener en cuenta las posibles fuentes de error y la forma de minimizar el impacto asociado. Estos factores son fundamentales para diferenciar variantes reales de aquellas originadas debido al los errores de secuenciación [McElroy y cols., 2014].

Con el advenimiento de las tecnologías de secuenciado profundo se pudo superar las

limitaciones impuestas en el secuenciado de clones por el método de Sanger obteniendo cobertura de 10000 o más reads por par de base [Beerenwinkel & Zagordi, 2011]. La gran profundidad de las nuevas plataformas de secuenciación, permite la investigación de la evolución viral a una escala intra e inter hospedero [Schönherz y cols., 2016]. Con el uso de estas nuevas tecnologías es posible, además de detectar mutaciones de baja frecuencia, brindar información sobre la estructura poblacional, es decir, el conjunto de variantes dentro de la población y su frecuencia relativa [Beerenwinkel & Zagordi, 2011].

Por otra parte, las plataformas NGS tienen muchas aplicaciones, que van desde la medición de niveles de expresión génica [Esteller M, 2011; Skalsky & Cullen, 2010; Wilk E y cols., 2015; Wu L y cols., 2015], el descubrimiento de nuevos virus [Kriesel y cols., 2012; Singh K y cols., 2012] hasta estudios metagenómicos de comunidades de microrganismos [Balzola y cols., 2010; Gilbert & Dupont, 2011]. La utilización de estas nuevas tecnologías ha revolucionado los estudios en virología, en particular en el universo de virus con genoma ARN. Previamente se mencionó que estos virus evolucionan a velocidades extremas y generan gran diversidad poblacional. En este sentido, NGS resulta fundamental para el estudio de la diversidad intra paciente y sus consecuencias a nivel clínico y evolutivo. Las tecnologías de NGS han sido utilizadas para la caracterización de la evolución intra-paciente del virus Hepatits C durante la fase aguda temprana de una infección [Bull y cols., 2011; Hajarizadeh y cols., 2014]. Asimismo, con la utilización de estas plataformas, se ha visto cómo la transmisión de viriones a nuevos hospederos representan grandes eventos de cuellos de botella. Unas pocas variantes virales son las responsables de una nueva infección, como fue demostrado en el caso de infecciones de VIH en pacientes usuarios de drogas inyectables [Bar y cols., 2010].

Por otro lado, la secuenciación masiva de genomas completos ha sido ampliamente utilizada para la caracterización de infecciones agudas como es el caso de VIH-1. Henn y colaboradores, secuenciaron en forma masiva el genoma completo de VIH-1 y detectaron la rápida emergencia de variantes de escape de baja frecuencia que surgen como respuesta adaptativa a las células T $CD8^+$ del hospedero [Henn y cols., 2012]. Además, la detección de variantes minoritarias tiene profundas implicancias en el contexto de resistencia a drogas antivirales. Con el uso de NGS, fue posible la detección de mutaciones de resistencia a inhibidores de la proteasa de VIH que presentaron frecuencias menores a 1% [Fisher y cols., 2012]. Otros estudios han

mostrado la utilización de estas plataformas para la detección de mutaciones de baja frecuencia resistentes a drogas antivirales como es el caso del Virus Hepatitis C [Svarovskaia y cols., 2012; Verbinnen y cols., 2010], el Virus Hepatitis B [Nishijima y cols., 2012; Solmone y cols., 2009], el Virus Influenza A [Ghedin y cols., 2011; Ghedin y cols., 2012], entre otros agentes infecciosos de gran relevancia en salud pública humana.

Las tecnologías de NGS además de proporcionar conocimiento respecto a la evolución intra hospedero, brindan información a nivel poblacional acerca de la epidemiología molecular y los eventos de transmisión entre poblaciones de individuos infectados [Varble y cols., 2014]. Estos estudios mostraron como los eventos de cuellos de botella, en la transmisión del Virus Influenza A, difieren significativamente dependiendo de la ruta de infección. De igual modo, evidenciaron fuertes presiones genéticas responsables del efecto fundador durante la adaptación a diferentes especies de hospederos. Además, encontraron que la ruta de infección modula fuertemente la astringencia del cuello de botella, en donde la transmisión por aerosoles impone mayor selección que por contacto directo [Varble y cols., 2014]. Por otro lado Sim y colaboradores, para el caso del virus Dengue, estudiaron la diversidad genética durante eventos de transmisión de humano a mosquito. Mediante secuenciación masiva observaron cómo más del 90% de los “single nucleotide variants” (SNVs por sus siglas en inglés) se pierden durante eventos de transmisión de humano a mosquito, así como desde la zona abdominal del mosquito hacia sus glándulas salivales [Sim y cols., 2015].

Para resumir esta sección relacionada a las posibles aplicaciones de NGS a la virología, es importante remarcar que estas tecnologías proporcionan gran velocidad y rendimiento. Son capaces de producir enormes volúmenes de datos con muchas aplicaciones posibles en el área de la investigación y el diagnóstico. Las técnicas de NGS se han utilizado ampliamente para: secuenciar genomas virales completos, detectar la variabilidad genómica viral, estudiar la evolución viral intra-hospedero, así como para monitorear mutaciones de baja frecuencia asociadas a resistencia a drogas antivirales [Barzon y cols., 2011].

RELEVANCIA DE LA INVESTIGACIÓN PROPUESTA

El tronco común del presente trabajo se basó en el estudio de la dinámica evolutiva de las poblaciones virales. Si bien se sabe que los virus circulan como una compleja nube de variantes íntimamente relacionadas, muchas de las cuestiones que subyacen en tales poblaciones aún son un misterio sin resolver. Entender como los virus se relacionan y evolucionan es de gran utilidad para afrontar futuras epidemias con mayor precisión. El estudio de la dinámica poblacional y los patrones evolutivos de las poblaciones virales es crucial para la comprensión de los procesos evolutivos. Profundizar en el estudio de las enfermedades virales puede ayudar a mejorar la composición de las cepas vacunales y contribuir al desarrollo de nuevos fármacos antivirales, lo que resulta trascendental para mejorar las políticas de salud a nivel de la población humana. El conocimiento en profundidad de los componentes de una población viral intra hospedero es fundamental para evidenciar posibles cepas resistentes a antivirales o que resulten muy patogénicas. En este sentido, conocer las variantes intra-cuasiespecie y sus relaciones, nos permiten brindar mayores herramientas a la población con el fin de dilucidar los problemas específicos de cada paciente y poder diseñar terapias antivirales paciente específicas.

Este estudio se divide en tres capítulos principales. El primero se enfocó en la implementación de programas computacionales y al desarrollo de algoritmos bioinformáticos que ayuden al procesamiento y al análisis de los datos de NGS. Como modelo biológico se utilizó el Virus Influenza A H3N2 para el estudio de las variantes presentes en muestras de pacientes uruguayos infectados. Este virus pertenece a la familia Ortomixoviridae, presenta un genoma segmentado de ARN simple hebra con polaridad negativa. Al igual que la gran mayoría de virus con genoma ARN, tiene grandes consecuencias a nivel de salud pública humana e implicancias (humanas y económicas) devastadoras como se vio en los eventos pandémicos. Es de real interés el estudio en profundidad de los principales determinantes antigénicos de este virus como son la Hemaglutinina (HA) y la Neuraminidasa (NA). Estos determinantes son los más cambiantes dentro del genoma y son utilizados para la clasificación de estos virus. En particular se buscaron variantes minoritarias (detección SNPs y su frecuencia asociada) capaces de producir cambios aminoacídicos característicos de las cepas resistentes a antivirales. Cabe destacar que el desarrollo de estas metodologías, tanto experimentales como bioinformáticas, es de fundamental importancia a la hora de la detección de variantes de resistencia antiviral. Esto ayudaría enormemente en el

diseño de terapias antivirales paciente específico.

El capítulo número dos trata principalmente sobre la implementación de algoritmos bioinformáticos para el ensamblado de las variantes intra cuasiespecie presentes en las muestras tratadas en el capitulo previo. Mediante éstos, se buscó dilucidar los haplotipos, sus frecuencias e inter-relaciones.

Por último, en el capítulo tres de este trabajo se realizaron análisis bayesianos de coalescencia en donde se utilizó, como modelo biológico, el virus de la enfermedad de Newcastle (NDV). Se buscó investigar sus tasas evolutivas, la dinámica poblacional y los patrones de evolución de las variantes de este virus a nivel mundial. Los resultados son discutidos en términos del posible rol que juega la Antártida en el mantenimiento de poblaciones virales de NDV, potencialmente emergentes o re-emergentes en todo el mundo.

Capítulo 1

Desarrollo de un pipeline bioinformático para el pre-procesamiento de datos de deep seq

RESUMEN

La llegada de las plataformas de secuenciación masiva permitió la exploración, sin precedentes, de la variabilidad genética viral. La gran profundidad de dichas tecnologías ayudó a la detección de variantes minoritarias y mayoritarias dentro de una muestra. La inmensa cantidad de datos emitidos por estas plataformas condujo al desarrollo, en paralelo, de algoritmos computacionales que ayuden en la manipulación y al procesamiento de tales enormes volúmenes de datos. Conocer en profundidad la diversidad viral puede ayudar a la detección de variantes resistentes al tratamiento antiviral así como de escape a la respuesta inmune del paciente. Por estas razones, se realizó la construcción de un pipeline bioinformático para el manejo de datos masivos derivados de muestras de pacientes infectados con virus Influenza A subtipo H3N2. El mismo consta de varias etapas de limpieza de datos, alineamientos y ensamblados. Asimismo, se enfocó a la detección de variantes virales (mayoritarias y minoritarias) y su posible asociación hacia resistencia a antivirales (para el caso de la NA). Los resultados se pueden dividir en tres partes. Por un lado, los resultados generales para el desarrollo de un pipeline bioinformático son discutidos en términos del performance de cada programa utilizado. Por otro lado se buscó identificar una posible asociación entre las variantes de cuasiespecies de distintas muestras analizadas mediante la deteción de SNPs característicos. Asimismo, se exploró el posible rol de las variantes, intra e inter cuasiespecie, para la mantención de variantes resistentes a los antivirales. Se determinó un pipeline bioinformático para el procesamiento de datos masivos de muestras virales. Mediante la utilización de herramientas de alineamiento y ensamblado se pudo reconstruir las secuencias consenso. Por otra parte se determinó las variantes mayoritarias y minoritarias, dentro de la nube de mutantes, y su posible asociación entre muestras de distintos años. Por último se enfocó a la búsqueda de variantes de la Neuraminidasa con resistencia

antiviral. Se detectó variantes virales con ciertas mutaciones en posiciones genómicas que se sabe confieren resistencia a los inhibidores de la Neuraminidasa.

INTRODUCCIÓN

El advenimiento de las tecnologías de NGS, tales como las plataformas Roche GS FLX e Illumina Genome Analyzer, permitió la detección de variantes minoritarias que, como su nombre lo indica, se encuentran a muy baja frecuencia dentro de una muestra [Watson y cols., 2013]. Estas plataformas brindan una enorme cantidad de datos, en el orden de los GygaBytes. En paralelo a la llegada de las plataformas de NGS, fue necesario el desarrollo de algoritmos computacionales que ayuden en la manipulación y al procesamiento de enormes volúmenes de datos [Schadt y cols., 2010; Pandey y cols., 2011; Schmieder R. & Edwards R., 2011; Patel RV. & Jain M., 2012]. La cuantificación y cualificación de la diversidad viral puede tener importantísimas implicancias clínicas. Algunos ejemplos son: la detección de variantes resistentes al tratamiento con antivirales y la emergencia de variantes minoritarias que escapen a la respuesta inmunitaria del hospedero.

Como ya se mencionó, NGS provee muchísimas aplicaciones las cuales incluyen secuenciación de ADN y ARN, entre otras [Lee y cols., 2013]. Estas herramientas metodológicas pueden ser aplicadas a un amplio rango de cuestiones biológicas que abarcan, desde estudios en pacientes con cáncer hasta estudios de agentes infecciosos, e involucran tanto el secuenciamiento de genomas completos (WGS, Whole Genome Sequencing, por sus siglas en inglés) como el estudio dirigido hacia una región específica del genoma o gen de interés [Lee y cols., 2013]. El principal objetivo en este tipo de secuenciación es la búsqueda de variabilidad genómica en términos de SNVs, INDELS, o cualquier otro tipo de variación genómica que pueda estar asociada a enfermedades encontradas en humanos, tanto intrínsecas como causadas por algún agente patógeno [Gonzaga-Jauregui y cols., 2012; Snyder y cols., 2010]. Por otro lado, la secuenciación del RNA, que mide los cambios en la expresión génica, puede ser utilizada entre otras cosas para el descubrimiento de nuevos transcriptos, en la búsqueda de RNAs no codificantes, o hasta en el deslumbramiento de isoformas de splicing alternativo [Weikard y cols., 2013; Liu y cols., 2017].

Una corrida de NGS puede producir millones de cortas secuencias llamadas lecturas o reads, que dependiendo de la plataforma utilizada nos brindará distintos resultados. La Tabla 1 (Anexo I), muestra las principales características de las diferentes plataformas

de secuenciación masiva. Algunas características son: la cantidad máxima de datos generados, largo de reads, tasa de error, entre otras muy relevantes en el diseño experimental, ya que dependiendo del objetivo convendrá usar una u otra plataforma [Mardis, 2013; Barzon y cols., 2011; Lee y cols., 2013].

Como ya se mencionó, estos secuenciadores de nueva generación pueden generar cientos de millones de secuencias en una simple ronda de secuenciación. Antes de analizar estas secuencias para producir resultados y conclusiones biológicas, es aconsejable realizar controles de calidad para evitar problemas de sesgo en los análisis ulteriores. En las secciones siguientes se detallan los principales pasos y algortimos bioinformáticos más comúnmente usados en el pre-procesamiento y procesamiento de los datos masivos. Estos incluyen la obtención de reads de buena calidad, el alineamiento, ensamblado de las lecturas y la detección de SNPs y sus frecuencias, entre otros.

Programas para el Pre-procesamiento de datos masivos:

En la etapa de pre-procesamiento, lo primero que se debería evaluar sobre los datos de secuenciación es la calidad de los mismas. Cuando hablamos sobre calidad de secuencia, en un contexto de secuenciación, nos referimos a los valores de calidad en escala de Phred. Esta escala representa la confianza o probabilidad en la asignación de cada base por la plataforma de secuenciación [Ewing y cols., 1998a]. La escala de Phred, fue originalmente utilizada para representar los valores de calidad de las bases a comienzos del proyecto Genoma Humano [Ewing y Green, 1998b]. Actualmente, esta escala se aplica para representar las probabilidades y valores de confianza en un contexto genómico. Por convención, los valores de calidad de base en escala de Phred más usados caen en el rango de 2 a 40, con variaciones en el rango dependiendo del origen de los datos de secuencia. Sin embargo, estos valores pueden variar de cero a infinito. Se podría interpretar este valor como un estimativo del error, en donde el error es por ejemplo la probabilidad que una base sea incorrectamente nombrada por el secuenciador. Pero también se podría interpretar como un estimativo de la precisión, donde la precisión es por ejemplo la probabilidad de que la base sea identificada correctamente por el secuenciador. Un valor de Phred de 20 o superior se supone aceptable, ya que esto significa que presenta una precisión de un 99% con una chance de error de 1%.

Existen disponibles muchísimas herramientas para el control de calidad de los datos de secuenciado masivo. Entre ellas encontramos a FastQC [Andrews S., 2010], un programa desarrollado en Java que provee una descripción general sobre el aspecto y problemas encontrados en los datos de secuencias. Para asegurar un resultado final coherente, es necesario filtrar los datos de baja calidad, remover secuencias o bases de baja calidad ("trimming"), eliminar adaptadores y posible contaminación. Mejorar la calidad de nuestros datos de secuenciación es un paso sumamente complejo y fundamental para continuar con nuestros análisis. Como se puede ver en la Tabla 1 (Anexo I), cada protocolo de secuenciación introduce algún tipo de sesgo específico, en otras palabras, cada tipo de secuenciador será susceptible a generar algún tipo de ruido o error. Son varios los factores que pueden afectar este tipo de secuenciación. Están los artefactos que surgen durante la construcción de las librerías (ej.: sesgo en la composición de bases y el tamaño de los fragmentos) y aquellos vinculados al proceso de secuenciación en sí mismo (ej.: los errores sistemáticos en la calidad de las lecturas). Estos factores pueden impactar negativamente en la calidad de los datos y por consiguiente en los análisis subsiguientes. La variación en la proporción de secuencias duplicadas, introducidas por sesgo en la amplificación por PCR, y la contaminación de especies conocidas o desconocidas, diferentes a nuestro target, también pueden influenciar negativamente en los resultados [Schmieder & Edwards, 2011; Zhou y cols., 2013; Triverdi y cols., 2014]. Además, incluso la especie bajo investigación y contexto biológico de las muestras, son capaces de influenciar sobre los resultados e introducir algún tipo de sesgo.
En resumen, FastQC reporta varias fuentes de sesgo, como ser, contenido en GC, enriquecimiento por PCR, errores producidos durante la secuenciación, entre otros. Además de FastQC existen muchos programas utilizados para el control de calidad de los reads, como ser, HTSeq [Anders y cols., 2014], SAMStat [Lassmann y cols., 2011], RSeQC [Wang y cols., 2012], entre otros.

Podemos mejorar aún más la calidad de nuestras secuencias "trimando" los reads, es decir, removiendo los adaptadores y bases de mala calidad. Para ello hay disponibles en forma gratuita gran variedad de algoritmos bioinformáticos que permiten, de manera rápida y confiable, este tipo de procesamiento. Entre los programas encontrados en la bibliografía se destaca Scythe [https://github.com/vsbuffalo/scythe], que utiliza una aproximación Bayesiana para la clasificación de pequeños "substrings" encontrados en las lecturas. Además, Scythe considera la información de calidad. Esta

característica lo hace más robusto en la eliminación del adaptador localizado hacia el extremo 3', que usualmente presenta bases de baja calidad. Esta es una propiedad muy común de las lecturas provenientes de los secuenciadores NGS. Lo conveniente es realizar el trimado basado en calidad antes de proseguir con los análisis subsiguientes, como el mapeo y ensamblaje. Sin embargo, el trimado basado en calidad podría remover bases útiles en la identificación de contaminación por adaptadores hacia el extremo 3'. Por esta razón, es muy recomendable utilizar Scythe, o cualquier otra herramienta de corte de adaptadores, antes del trimado basado en calidad.

Otro programa muy útil en el pre-procesamiento de datos es Trimmomatic (http://www.usadellab.org/cms/index.php?page=trimmomatic) [Bolger AM. y cols., 2014]. Este programa realiza una variedad de tareas de recorte para datos de Illumina paired-end y single-end. Es una herramienta de pre-procesamiento flexible, optimizada para los datos de NGS de Illumina. El software incluye varios pasos de procesamiento para el recorte y el filtrado de lecturas. Éste utiliza una arquitectura basada en tuberías que permite aplicar "pasos" individuales a cada par de lectura/lectura, en el orden especificado por el usuario. Otras características que Trimmomatic ejecuta son la remoción de adaptadores, la eliminación de bases en posiciones específicas basándose en umbrales de calidad, el corte de reads a un largo particular, la conversión de los puntajes de calidad a Phred-33/64, ente otras aplicabilidades.

Por otro lado encontramos a Sickle [Joshi y Fass, 2011], que trima las lecturas basado en la calidad de las mismas. Este programa se vale de ventanas deslizantes según umbrales de calidad y longitud de secuencia. Determina cuándo la calidad es suficientemente baja para "trimar" el extremo 3' de los reads y también evalúa cuándo la calidad es lo suficientemente alta para recortar el extremo 5'. Además descarta reads que caen debajo de cierto umbral de longitud [Joshi y Fass 2011, https://github.com/najoshi/sickle].

Otro programa ampliamente utilizado para la remoción de los adaptadores en los datos de secuenciación masiva es cutadapt [Martin M., 2011]. Esta herramienta soporta la salida de las plataformas Illumina, SOLiD y 454, y es usada especialmente cuando el largo de read de la máquina de secuenciación es más largo que la molécula secuenciada, como en el caso de los microRNAs.

FASTX Toolkit [http://hannonlab.cshl.edu/fastx_toolkit/] es otra herramienta de línea de comando que integra un conjunto de programas para el pre-procesamiento de los archivos de lecturas tanto en formato FASTA como en FASTQ. Algunas de las

características disponibles en este paquete son la conversión de formatos de archivos FASTQ a FASTA, brindar información sobre estadísticas de calidad, remoción de secuencias de adaptadores, filtrado y corte de secuencias basados en la calidad de las mismas y conversión de DNA a RNA.
Es importante tener en mente, a medida que se procesan los reads de secuenciación masiva, realizar chequeos de calidad en cada uno de los pasos intermedios mediante algunos de los programas mencionados anteriormente como ser FASTQC [Andrews 2010], HTSeq [Anders y cols., 2014], SAMStat [Lassmann y cols., 2011], RSeQC [Wang y cols., 2012].

Programas para el alineamiento y ensamblado:

Posteriormente al pre-procesamiento llega la etapa de alineamiento o mapeo de los reads contra un genoma de referencia, si es que está disponible, o alternativamente realizar un ensamblado *de novo* (sin secuencia de referencia). Básicamente hay dos tipos de algoritmos de mapeo principales, los basados en tablas de hash (indexado de las lecturas o del genoma de referencia) y los basados en la transformada de Burrows-Wheeler que se valen de estructuras tipo arrays de sufijos y prefijos (FM-index). Existen varios tipos de programas para el alineamiento de datos de deep sequencing. Dentro de estos encontramos dos variantes, aquellos que permiten la inserción de gaps y los que no. Las herramientas disponibles capaces de alinear reads en forma continua, es decir sin la inserción de espacios o gaps, se basan principalmente en dos clases de métodos. Están aquellos que utilizan la transformada de Burrows-Wheeler [Burrows & Wheeler 1994] como ser Bowtie [Langmead y cols., 2009] o su segunda versión Bowtie2 [Langmead & Salzberg, 2012] y BWA [Li & Durbin 2009]. Por otro lado tenemos aquellos basados en la extensión de una semilla, que utilizan los algoritmos Needleman-Wunsch [Needleman & Wunsch, 1970] o Smith-Waterman [Smith & Waterman, 1981]. Los primeros métodos suelen ser más rápidos, mientras que los segundos, como lo es el caso de BFAST (https://github.com/nh13/BFAST) [Homer N. y cols., 2009], a pesar de su mayor consumo en términos de tiempo computacional, parecen ser más sensibles generando mayor cantidad de reads correctamente alineados [Needleman & Wunsch, 1970] o Smith-Waterman [Smith & Waterman, 1981]. Bowtie es un programa de alineamiento ultra rápido, muy eficiente en el manejo de la memoria, usado para alinear las cortas secuencias de ADN o reads contra grandes genomas. Esta herramienta indexa el genoma de referencia utilizando un esquema

basado en la transformada de Burrows-Wheeler [Burrows & Wheeler 1994] y en el algoritmo de Ferragina y Manzini [Ferragina & Manzini 2000; Ferragina & Manzini 2001]. Para tener una idea del potencial que tiene este programa, dando como ejemplo el indexado del genoma humano, el algoritmo Burrows-Wheeler le permite a Bowtie alinear más de 25 millones de reads por hora computacional, utilizando una cantidad de memoria de aproximadamente 1.3 gigabytes. Además Bowtie utiliza una variante de este algoritmo que permite la introducción de mismatches y favorece alineamientos de alta calidad.

Por otro lado, BWA o Burrows-Wheeler Aligner, es un paquete computacional creado para el mapeo de secuencias de baja divergencia contra grandes genomas de referencia como el genoma humano. Este programa consiste de tres algoritmos: BWA-backtrack, BWA-SW y BWA-MEM. El primer algoritmo es diseñado para reads de secuencia de Illumina de hasta 100pb, mientras que los otros dos algoritmos procesan reads en el rango de 70pb a 1Mpb. En términos de tolerancia a errores de secuencia, BWA-backtrack soporta tasas de error que caen por debajo del 2%, mientras que BWA-SW y BWA-MEM toleran errores en el orden de 2, 3, 5 y 10% para alineamientos de 100, 200, 500 y mil o más pares de bases, respectivamente [Li & Durbin, 2009].

Además de alinear los reads contra un genoma de referencia, lo que se suele hacer es un ensamblado *de novo.* En este caso los reads son ensamblados directamente en los "transcriptos". Existen dos tipos de abordajes para el ensamblaje de secuencias derivadas de ARNs. Por un lado están los métodos que no requieren de un genoma guía o referencia como molde. En este grupo de métodos encontramos a Trinity [Haas y cols., 2013], un novedoso algoritmo para la reconstrucción *de novo,* eficiente y robusta, de transcriptos a partir de datos de RNA-seq. Trinity combina tres algoritmos (Inchworm, Chrysalis y Butterfly) independientes aplicados secuencialmente para el procesamiento de grandes volúmenes de datos de RNA-seq. Básicamente Trinity particiona los datos de secuencia en grafos de Bruijn individuales, representando en cada uno la complejidad transcripcional para un gen o locus determinado. Luego se procesa cada grafo independientemente para extraer las isoformas de splicing de largo completo y para separar transcriptos derivados de genes parálogos [Haas y cols., 2013].

Otro programa sumamente útil para este caso es Velvet (velveth, velvetg) [Zerbino & Birney, 2008] que también utiliza los grafos de Bruijn para ensamblar cortas lecturas. Velvet se compone de un conjunto de algoritmos que manipulan eficientemente los grafos de Bruijn para eliminar errores y resolver repeticiones. Estas dos tareas se

realizan por separado: en primer lugar, el algoritmo de corrección de errores fusiona las secuencias que pertenecen entre sí, y luego el solucionador de repetición separa las rutas que comparten superposiciones locales.
Por otro lado están los métodos que utilizan un genoma de referencia para alinear y ensamblar los reads en ARNs completos. Dentro de estos encontramos a Cufflinks [Trapnell y cols., 2010], un ensamblador de transcriptos que estima sus abundancias y prueba la expresión diferencial y regulación en muestras de ARN. Este programa acepta reads alineados y los ensambla generando el conjunto más parsimonioso de transcriptos. Luego estima la abundancia relativa de esos transcriptos basado en cuantos reads sustentan cada unos de los transcriptos y tiene cuenta el sesgo en la preparación de las librerías.
Asimismo, es de destacar que el programa Trinity [Haas y cols., 2013] puede ser utilizado para ensamblar con genoma guía seteando la opción “genome_guided_bam” que usa como input el mapeo previo de los reads contra la referencia.

Evaluación y Visualización de los alineamientos y ensamblados:

Evaluación de alienamientos y ensamblajes:

Respecto a la evaluación y registro de la fidelidad de los alineamientos para datos de deep sequencing, podemos encontrar varias herramientas computacionales que garanticen este paso del análisis. Dentro de estas encontramos por ejemplo a Qualimap [García-Alcalde y cols., 2012] y su versión actualizada Qualimap2 [Okonechnikov y cols., 2016]. Ésta es una aplicación, tanto “friendly” como de linea de comandos, desarrollada en el lenguaje de programación Java que soporta el control de calidad de los datos mapeados (alineados contra un genoma de referencia). Esta herramienta toma en cuenta características de las secuencias y propiedades de los genomas. En resumen, Qualimap toma los datos de secuencias alineadas y brinda análisis estadísticos y gráficos para la evaluación de los mismos. Dicho control de calidad es vital para descubrir problemas, tanto en el proceso de secuenciamiento como en el de mapeo, que deben ser solucionados previo al resto de los análisis [García-Alcalde y cols., 2012; Okonechnikov y cols., 2016].
Por otro lado, en el caso de la evaluación de los ensamblajes encontramos a Quast [Gurevich y cols., 2013]. Una herramienta que evalúa la calidad y compara ensamblajes de genomas. QUAST puede evaluar ensamblajes tanto con un genoma

de referencia como sin una referencia. Éste produce muchos informes, tablas de resumen y gráficos, para ayudar a los científicos en su investigación y en sus publicaciones.

Visualización:

Luego de alinear y/o ensamblar los reads, lo conveniente es visualizar nuestro alineamiento para tener idea de la fidelidad de nuestros análisis. Existen muchísimas herramientas de visualización. Dentro de estas podemos destacar a Tablet [Milne y cols., 2013], un visualizador gráfico de alto desempeño usado para la observación de alineamientos y ensamblajes de secuencias de NGS. Es una herramienta de alto rendimiento, muy útil para la navegación y visualización de datos masivos. Tablet muestra los reads, tanto en formato empaquetado como apilado, y soporta los siguientes formatos de archivos: ACE, AFG, MAQ, SOAP2, SAM, BAM, FASTA, FASTQ, y GFF3. El mismo busca y localiza reads por su nombre o subsecuencia a través del conjunto entero de datos. Además, permite la visualización de paired-end reads (formato SAM/BAM) y una simple visualización de las variantes.
Otra herramienta ampliamente utilizada para la visualización de datos de deep sequencing es IGV (Integrative Genomics Viewer) [Robinson y cols., 2011; Thorvaldsdóttir y cols., 2013]. IGV es un visualizador de alto desempeño que maneja eficientemente enormes y heterogéneos conjuntos de datos. Éste genera una experiencia de uso muy amigable e intuitiva en todos los niveles de resolución genómica. Una característica clave de IGV es su foco en la naturaleza integrativa de estudios genómicos. El mismo soporta, datos basados en arrays, secuencias de nueva generación y la integración de datos clínicos y fenotípicos.

Procesamiento post-alineamiento:

Las actividades más comúnmente aplicadas en la etapa de procesamiento post-alineamiento incluyen: la conversión de formatos de los archivo de salida, la creación de reportes del proceso de alineamiento y la remoción de artefactos derivados de la etapa de PCR. Estas etapas pueden ser llevadas a cabo por ejemplo utilizando el paquete SAMtools [Li y cols., 2009]. La conversión de formatos con este tipo de herramienta permite no solo la reducción en tamaño de los data set de salida, sino que

también garantiza la compatibilidad en los análisis subsiguientes. Ejemplo de ello es el caso del “variant calling”, tal como ocurre con el programa UnifiedGenotyper del paquete Genome Analysis Toolkit [McKenna y cols. 2010]. En general, los programas de alineación generan un resumen que describe el proceso de alineamiento. Esta descripción incluye: el número total de reads alineados, el número de pares de reads alineados correctamente y el número de reads alineados exactamente una vez. Dicho resumen de cuestiones estadísticas son fundamentales para la evaluación de la calidad global y exactitud del alineamiento. En caso que el reporte no sea generado por el propio programa, tal como ocurre con BWA, es posible generar dicho resumen usando por ejemplo la herramienta samtools flagstat [Li. y cols., 2009].
Otro paso importante del post-procesamiento es la remoción de las dúplicas de PCR. Estas dúplicas representan el mismo par de reads ocurriendo repetidas veces en los datos crudos y cambia el estimativo verdadero de la cobertura del genoma de referencia. La herramienta samtools utiliza la opción “rmdup” para la remoción de este tipo de artefacto [Li y cols., 2009].

Detección de Variantes:

Es importante recordar que las tecnologías de NGS aún son altamente sensibles a errores técnicos. El uso de las mismas lleva consecuentemente a una dependencia directa con el uso de herramientas bioinformáticas para su manipulación [Medvedev P. y cols., 2009; Horner DS. Y cols., 2010]. Técnicamente, dos de los principales pasos bioinformáticos en el procesamiento de las secuencias generadas por NGS incluyen: alineamiento de las lecturas al genoma de referencia y la detección de SNPs [Mielczarek & Szyda, 2016].

Procesamiento previo al variant calling:

Para obtener resultados confiables del estudio de polimorfismos o variant calling son necesarios pasos de procesamiento adicional específicos, los cuales son muy recomendados realizar previamente a la detección de variantes. Los pasos más importantes involucran la corrección de artefactos en los alineamientos y/o ensamblajes y la recalibración de los valores de calidad de las secuencias.
Durante el proceso de alineación podrían ocurrir fallas que resultarían en el “mismatche” de una o varias bases cerca del sitio erróneamente alineado. Tal

“mismatche” puede ser fácilmente confundido como un SNP. La presencia de InDels en los genomas individuales, con respecto al genoma de referencia, requiere la realineación local para la solución de dicho problema.
La precisión de un alineamiento juega un rol fundamental en la detección de variantes, dado que reads incorrectamente alineados podrían llevar a errores en dicho proceso. Es importante que los algoritmos de alineamiento sean capaces de hacer frente a errores de secuenciación, así como a diferencias potencialmente reales (tanto mutaciones puntuales como indels) entre el genoma de referencia y el genoma secuenciado. Además, es importante que los programas de alineamiento produzcan valores de calidad de alineación bien calibrados, ya que las variantes y sus probabilidades posteriores dependen de esas puntuaciones [Nielsen R. y cols., 2011].
Por tal motivo se han diseñado herramientas de alineamiento local para realinear los reads en la cercanía de un artefacto previamente identificado. Más aún, es importante remarcar que las plataformas de secuenciación podrían calcular de forma errónea la calidad de las bases (expresadas en valores de Phred). Para solucionar esto existen herramientas provistas por el paquete recalibrate de GATK [McKenna y cols. 2010] que recalibra los puntajes de calidad de las bases, lo que resulta en estimativos más precisos de las probabilidades actuales de “mismatches” contra el genoma de referencia.
Dado que el realineamiento es un proceso costoso en términos de tiempo computacional, han sido desarrollados otro tipo de abordajes. Entre estos encontramos a SAMtools [Li y cols., 2009a], un paquete bioinformático que utiliza una aproximación basada en la combinación de ambos tipos de correcciones. El programa asigna un puntaje de calidad de base (BAQ, base alignment quality) a cada posición genómica. BAQ se calcula como la probabilidad escalada de Phred de que dicha base esté mal alineada. Si la base es alienada a una base de referencia diferente en un alineamiento sub-óptimo su BAQ será bajo. La estimación del BAQ es implementada en SAMtools como opción por defecto.

Breve descripción sobre Variant-calling:

Avances concomitantes, de las tecnologías de secuenciación y del desarrollo de algoritmos computacionales, han provisto nuevas oportunidades para la identificación de un gran número de variantes. En teoría, la secuenciación de los genomas completos permite, potencialmente, el descubrimiento de todos los polimorfismos

existentes. Consecuentemente, no solo posibilita la identificación de SNPs comunes, sino también de variantes raras. Sin embargo, es un pre-requisito contar con alta cobertura para una detección confiable de las variantes, ya que el número de reads alineados a cada base es usado para diferenciar entre errores de secuencia y polimorfismos verdaderos.

Muchos algoritmos y paquetes computacionales han sido dedicados a la identificación de SNPs en datos de NGS. Cuando un único genoma es analizado, SNP calling y genotype calling son prácticamente el mismo proceso. Esto se debe a que genotipos distintos a la referencia implican la presencia de un SNP. En el análisis simultáneo de múltiples muestras, un SNP es identificado si al menos presenta una posición diferente a la referencia. Así, SNP calling podría ser definido como el proceso de identificación de sitios que difieren de la secuencia de referencia [Nielsen y cols. 2011]. Existen varios programas gratuitos disponibles para el proceso de variant-calling. En particular, el paquete samtools mpileup recaba información contenida en los archivos de entrada BAM (binary version of the Sequence Alignment Map format, SAM) y calcula la probabilidad de los datos dado para cada posible genotipo. Luego bcftools aplica la probabilidad a priori y calcula el SNP calling [Li y cols. 2009a]. Otro ejemplo de programas ampliamente utilizados son GATK [McKenna y cols. 2010], Atlas-SNP2 [Shen Y. Y cols., 2010], SOAPsnp [Li R. y cols., 2009b] y SNVer [Wei y cols. 2011].

OBJETIVOS

Objetivo General:

Profundizar en el conocimiento de la diversidad de las cuasiespecies virales intra e inter hospedero de muestras provenientes de pacientes uruguayos infectados con el Virus Influenza A circulantes en Uruguay durante las temporadas invernales 2011-2013.

Objetivos Específicos:

1) Construcción de librerías genómicas para las muestras de VIA y posterior secuenciación masiva.

2) Construcción de un pipeline bioinformático para el pre-procesamiento de los datos masivos.

Trimado de adaptadores, primers (utilizados en la amplificación inicial) y extremos de baja calidad.

Alineamiento y ensamblado.

Validación de los Alineamientos y ensamblados.

Visualización de los datos.

3) Detección de Variantes minoritarias (SNPs calling).

4) Búsqueda de una posible correlación entre poblaciones virales.

5) Mapeo de mutaciones (SNPs) de interés en la estructura proteica de la Neuraminidasa y su posible asociación de resistencia a los antivirales.

METODOLOGÍA

1) Recolección de muestras y extracción del ARN genómico viral.

Con el fin de profundizar en el estudio de las relaciones genéticas entre las cepas de Influenza que circularon en Uruguay, se recolectaron veintiocho muestras de pacientes uruguayos (hospitalizados y ambulatorios) con Enfermedad tipo Influenza (ETI), Insuficiencia Respiratoria Aguda (IRA) e Insuficiencia Respiratoria Aguda Grave (IRAG). La información correspondiente a cada uno de los pacientes se detalla en la Tabla 2 (Anexo I). Dichas muestras fueron provistas por la Asociación Española Primera de Socorros Mutuos (AEPSM), en el marco del proyecto Alianza PE_ALI_2009 1 1603 - ANII - AESPM - Facultad de Ciencias, UdelaR. El ARN total fue extraído mediante el uso del kit comercial High Pure Viral Nucleic Acid Kit de ROCHE de acuerdo con instrucciones suministradas por los fabricantes. El material genético fue almacenado a -80°C para ser luego utilizado en el paso de síntesis de ADN complementario (ADNc).

2) Transcripción Reversa (TR).

Se realizó la transcripción reversa del ARN presente en cada muestra. La síntesis del ADNc se realizó con cebadores cortos al azar (hexámeros). La TR se realizó en un volumen final de 20 µL conteniendo: 5 µL de ARN de concentración conocida en ng/µL, 4 µL de buffer de reacción 5x (el cual contiene: 250 mM TrisHHcl, pH 8.3 a temperatura ambiente; 375 mM KCl; 15mM MgCl2), 1 µL de dNTPs 10mM, 1 µL de hexameros (50-250 ng), 1 µL de transcriptasa reversa SuperScript II (200 U/µL; Invitrogen), 2 µL de DTT 0,1 M y 6 µL de H_20 libre de RNAsa. La reacción se inició con el calentamiento del ARN junto con los dNTP's y los hexámeros, durante 10 minutos (min.) a 65°C; a continuación se colocaron en hielo por un minuto. Seguidamente se agregó a la mezcla de reacción el buffer 5X y el DTT y se incubó todo a 25°C por 2 min. para luego agregar 1µL de la enzima SuperScript II Reverse Transcriptase (RT). Finalmente se incubó la mezcla de reacción en el termociclador Corbett modelo CAS 1200 en la siguientes condiciones: 25°CH10 min.//42°CH50 min.//70°CH15min. El ADNc fue almacenado a -80°C hasta su posterior utilización.

3) PCR con enzimas de alta fidelidad para los genes blanco (HA y NA).

El ADNc sirvió como molde para un primer round de amplificación de los genes blanco utilizando la enzima Platinum Taq DNA Polyerase High Fidelity de Invitrogen siguiendo las especificaciones del fabricante. La amplificación se realizó en un volumen final de 50 µL conteniendo: 5 µL de Buffer 10X -1X concentración final- (el cual contiene 60 mM Tris-SO_4 (pH 8.9), 180 mM Sulfato de Amonio), 1µL de dNTP`s 10 mM, 1,5 µL de $MgCl_2$ 50 mM, 1 µL del primer forware, 1 µL del primer reverse, 0,3 µL de la enzima Taq Platinum, 35,3 µL de agua ultra pura y 5 µL del molde de DNAc. Finalmente se incubó la mezcla de reacción en el termociclador Corbett modelo CAS 1200 en la siguientes condiciones: 94°C - 2 min. // 94°C - 30 seg. / TM-HA53yNA55 °C – 30 seg. / 68ºC – 2 min // 68 ºC – 10 min. // 20 ºC infinito. Los amplicones fueron almacenados a -20°C hasta su posterior utilización.
Los amplicones generados en este ciclo de amplificación fueron el material de partida para una segunda etapa de amplificación por PCR utilizando la técnica de Heminested, con un juego de cebadores internos específicos de cada gen. Para esta ronda de amplificación se utilizaron exactamente las mismas cantidades y tipos de reactivos que para el primer round, con la excepción del molde que, en este caso, se usó 2 µL del primer round y 39,2 µL de agua. Las condiciones de ciclado y termociclador utilizados fueron idénticas que para el primer round. Los amplicones fueron almacenados a -20°C hasta su posterior utilización.

4) Electroforesis en gel de agarosa 1% para confirmar banda única.

Los productos de PCR fueron separados electroforéticamente en geles de agarosa al 1% teñidos con Gel red (Nucleic Acid Stain). Los tamaños de banda esperados para el primer round de los genes HA y NA fueron 1692 y 1353, respectivamente. Para el segundo round fue 902 y 907 para el caso de HA; 806 y 813 para el caso de NA.

5) Cuantificación ADNc con nanodrop:

Para la cuantificación del producto de ADN purificado se utilizó la técnica de espectrofotometría UV-visible. En el caso de ADN, 1 UA 260 nm = 50 µg/mL. Se midió la absorbancia de la muestra a 280 nm para determinar la pureza de la misma, considerando relaciones Abs260/Abs280 > 1,8 como muestras de buena calidad.

6) Secuenciación por el metodo de Sanger de todos los amplicones:

Las bandas obtenidas fueron cortadas, purificadas y procesadas con el kit de purificación de Invitrogen PureLink Quick Gel Extraction Kit. El producto obtenido de dicha purificación se utilizó para su sencuenciación por el método de Sanger. Posteriormente fueron analizados los cromatogramas con el software Lasergen, y se ensamblaron las secuencias consenso obteniendo los genes completos.

7) Cuantificación por Qubit

La concentración del ADN de partida para la construcción de las librerías NexteraXT fue medida con el instrumento Qubit® y el kit Qubit® 2.0 Fluorometer dsDNA BR Assay (Life Technologies, CA, USA) con las especificaciones del fabricante. Éste usa 2 µL de cada muestra de ADN con 198 µL de la solución de trabajo Qubit.

8) Construcción de librerías NexteraXT para Illumina

La preparación de las librerías paired-end se realizó usando el protocolo NexteraXT bajo las especificaciones del fabricante. Este kit fue optimizado para concentraciones iniciales de ADN de 1 ng, por lo que es aconsejable cuantificar el material de partida previamente a la preparación de las librerías. Diluir el material de partida en agua de grado molecular o 10 mM de Tris HCL, pH 7.5-8.5.
Se utilizaron 9 muestras de pacientes infectados con VIA. Se cuenta con 4 amplicones por muestra de paciente (2 por cada gen), lo que da un total de 36 amplicones de unos 900 nucleótidos cada uno aproximadamente (902 y 907 para el caso de HA; 806 y 813 para el caso de NA).

9) Bioanalyzer Check Libraries

Este procedimiento es fundamental para corroborar el estado de las librerías. Se corrió 1 µL de la librería no diluída en la plataforma Agilent Technology 2100 Bionalyzer con el chip “High Sensitivity DNA chip” y se siguió las especificaciones del fabricante. Luego de chequear el estado de nuestras librerías, se procedió a normalizarlas para asegurar una representación equitativa de cada una de las librerías en el “pooleado” de las mismas.

10) Secuenciación Masiva:

La secuenciación fue llevada a cabo en reads de 150 nucleótidos (preparado con el kit NexteraXT) de forma paired-end (2X150). Dicha secuenciación se condujo con el instrumento de Illumina MiSeq, perteneciente al servicio de secuenciación del Instituto Pasteur de Motevideo, bajo la indicación del Dr. Gonzalo Greif (al igual que la construcción de la librería, normalización, y todas las etapas pertinentes a esta parte). La misma produjo un volumen aproximado de 9 GB para el total de las muestras.

11) Análisis de Datos Masivos

a) Pre-procesamieto.

Los datos crudos fueron analizados con el programa FastQC con el fin de chequear la calidad de los mismos y tener una visión global del estado de estos [Andrews S., 2010].

Para poder continuar con los análisis bioinformáticos siempre es necesario realizar una limpieza de los datos crudos y eliminar la contaminación que dificulta los análisis ulteriores. El primer paso fue el trimado de los adaptadores remanentes, que muchas veces quedan ligados a los reads. Seguidamente se procedió a trimar la secuencia de los primers utilizados en la etapa de amplificación de los genes de interés.

Para estas etapas de trimado de utilizó la herramienta Trimmomatic siguiendo las específicaciones de los desarrolladores.

Cabe destacar que posterior a cada etapa de limpieza se realizaron los chequeos de calidad pertinentes con la herramienta FastQC. Luego son eliminadas las secuencias de baja calidad con el programa Sickle-pe. Este trimma los extremos de mala calidad según umbrales de calidad. Como ya se mencionó, se considera una base de mala calidad si ésta presenta un valor de “q” menor a 30, lo que corresponde a un error de secuenciación de una base en 1000, nivel de ruido aceptado por la mayoría de los ensambladores.

Nuevamente se chequeó la calidad y estado de nuestros datos con la herramienta FastQC.

b) Alineamiento y ensamblado.

Concluida la limpieza de datos el paso siguiente implica el alineamiento y ensamblado de los mismos contra un genoma de referencia, si es que está disponible, y el ensamblado *de novo*.

Los reads obtenidos a partir del pre-procesamiento, que llamaremos de ahora en adelante reads de buena calidad, fueron mapeados en forma paralela con Bowtie, Bowtie2 y con BWA. En todos los casos se usaron como referencia los genomas obtenidos en la secuenciación de Sanger.

Los archivos de salida de todos los alineadores file.sam fueron tratados paralelamente. Primero se procedió a convertir estos archivos a su versión binaria, file.bam, con la herramienta samtools view opciones -bS. Luego se procedió a ordenarlos con la herramienta samtools sort que generó los sort_files.bam. Luego se continuó con la eliminación de las dúplicas de pcr y dúplicas de secuenciación con la herramienta Picard. Los archivos .sam, resultantes de ese paso, fueron filtrados con la herramienta samtools view con el fin de seleccionar únicamente aquellas secuencias que mapean a la referencia. Este output se procesó con la herramienta picard, opción SamtoFastq, para generar los fastq de entrada para los ensambladores. Se procedió a eliminar los reads que no mapean con la referencia y se generaron los archivos.sam correspondientes. La salida de este paso es el material de partida para el ensamblaje *de novo* por parte de dos ensambladores, Trinity y SPAdes. Previo al ensamblaje los archivos sort.bam se convirtieron con la herramienta samtools rmdup a sort_rmdup.bam y este último se convirtió a fastq con la herramienta bam2fq del paquete samtools. Paralelamente se llevó a cabo el ensamblado de los reads de buena calidad.

Se realizó el ensamblado *de novo,* independiente de referencia, en donde se usó tanto el programa Trinity como el programa SPAdes. Para ambas herramientas se setearon los parámetros por defecto. A partir de los contigs generados por estos ensambladores se construyeron las secuencias consenso para cada gen. Éstas fueron utilizadas como secuencias de referencia para analizar sus diferencias contra las secuencias generadas por la secuenciación de Sanger y para los análisis de variantes.

c) Validación de los alineamientos y ensamblados.

Posteriormente los alineamientos y ensamblajes deben ser validados. Para chequear la calidad de los mapeos se usó Qualimap [Okonechnikov y cols., 2016], mientras que para los ensamblajes se utilizó QUAST [Gurevich y cols., 2013].

d) Visualización de datos masivos.

Para comprobar cada etapa del procedimiento los alineamientos obtenidos fueron visualizados con el programa Tablet [Milne y cols., 2013].

e) SNPs calling

Para la detección de SNPs y profundización del estudio de las variantes minoritarias se utilizó la herramienta SAMtools [Li y cols., 2009a].

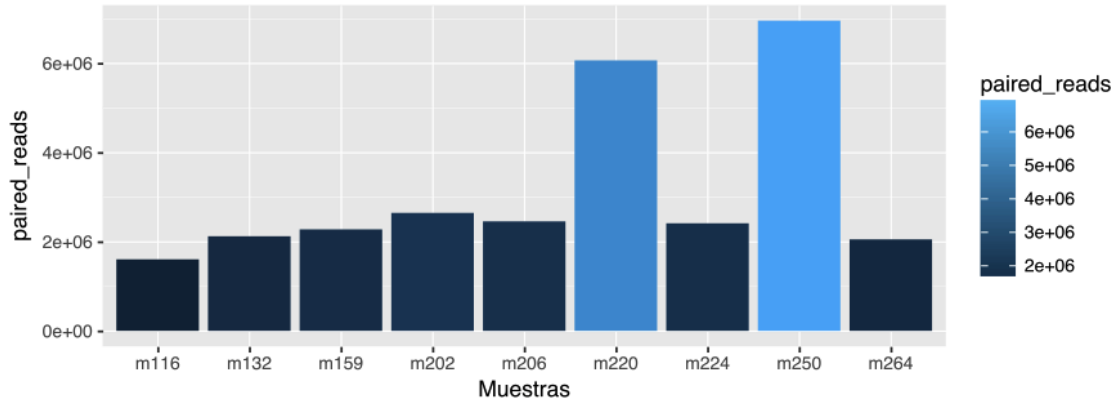
f) Mapeo de mutaciones sobre estructura cristalográfica

Se utilizó el programa VMD [Humphrey y cols. 1996] para realizar el mapeo de las mutaciones sobre la estructura 3D de la Neuraminidasa 4gzp.pdb, publicada en la plataforma del Protein Data Bank [https://www.rcsb.org/].

Secuenciación Masiva y Pre-procesamiento de datos:

La secuenciación masiva produjo un volumen aproximado de 9 GB, sumando (28.564.050 reads) 28.5 millones de pares de reads para todas las muestras. Se obtuvo el conteo y porcentaje de reads totales emitidos por el secuenciador MiSeq de Illumina, Figura 3 (A) y (B), respectivamente.

(A)



(B)

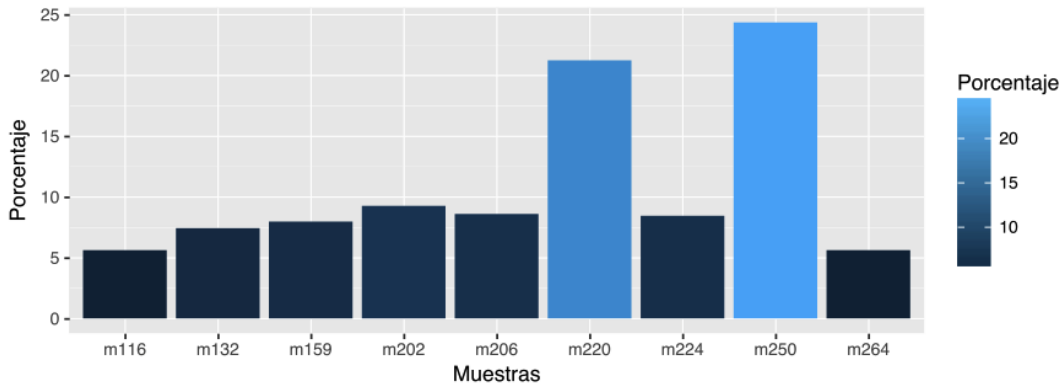


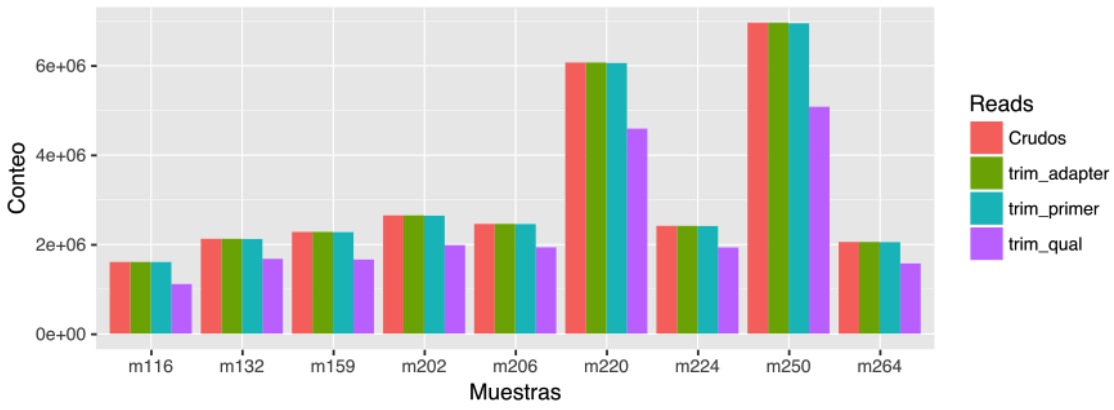
Figura 3. Salida del secuenciador MiSeq de Illumina.
(A) Conteo de reads totales por muestra. (B) Porcentaje de reads secuenciados para cada muestra. A la derecha de cada figura se observa la escala de colores que denota cantidad y porcentaje de reads, respectivamente.

El rendimiento teórico de la secuenciación con el kit NexteraXT es de 15 millones de secuencias (30 millones de pares de reads). Se obtuvo aproximadamente esa cantidad (28.5 millones de pares de reads) luego de los filtros de calidad propios del secuenciador MiSeq de Illumina. La calidad fue muy buena en general, con aproximadamente 96% de las bases con Q mayor a 30.

El indexado de las secuencias anduvo bien, aunque dos de las nueve muestras (m220 y m250) quedaron con más reads de lo deseado (20 y 25 %, respectivamente). El resto de las muestras obtuvieron entre 5.5 y 9.5 % del total de reads secuenciados (Figura 3). La muestra que presentó menor cantidad de lecturas fue la m116 con 1.650.000 de pares de reads, y dado que los amplicones sumados dan aproximadamente 3Kb, y la longitud de read fue 150, la cobertura en promedio quedó en 82500X (1.650.000 x 150 / 3000).

Luego de ejecutado el proceso de trimado de las secuencias adaptadoras, la cantidad de reads se mantuvo, el output de este análisis arrojó un total de 28.564.050 de paired-end reads para todas las muestras, lo que indicó la ausencia de secuencias adaptadoras. Por otro lado, el trimado de las secuencias de los primers, utilizados en la etapa de amplificación de los genes de interés, dio un total de 28.508.160 de paired-end reads para todas las muestras, lo que indicó un 0.2% de secuencias con extremo de primer. Por último, el filtrado por calidad de los reads resultó en un total de 21.496.102 de paired-end reads (llamados “reads de buena calidad”), lo que significó aproximadamente un 75.3% de lecturas de buena calidad. En la Figura 4 A y B se puede observar el conteo y porcentaje de reads, respectivamente, para los reads crudos y trimados. En la Tabla 3 (Anexo I) se puede apreciar en mayor detalle la cantidad de reads previo y posterior de las diferentes etapas de trimado.

(A)



(B)

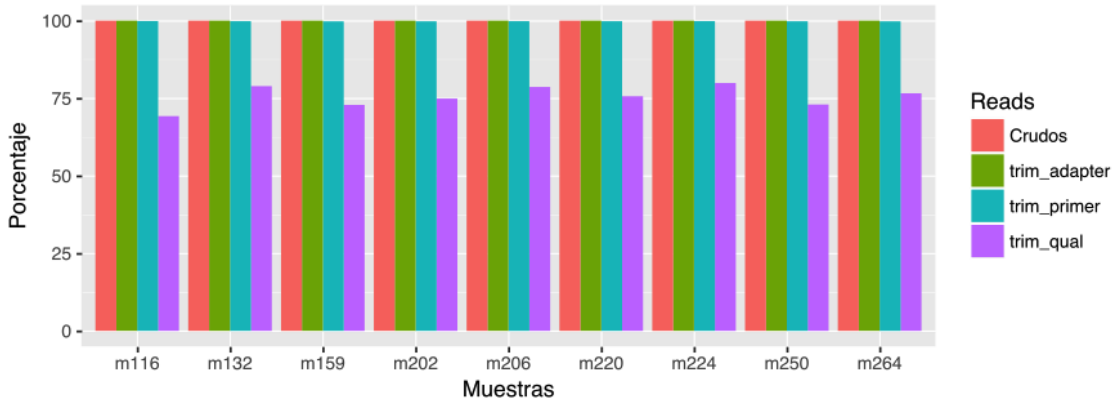


Figura 4. Comparación de datos masivos durante la etapa de Pre-procesamiento.
(A) Conteo de reads para datos crudos (rojo), trimado de adaptadores (verde), trimado de primers (cian) y trimado por calidad (violeta), por muestra analizada. (B) Porcentaje de reads para cada muestra, mismo patrón de colores que (A).

El porcentaje de lecturas que pasan la etapa de filtrado resultó entre el 70 y 80 % de la cantidad de reads iniciales por muestra (Figura 4).

Para cada paso en la etapa de pre-procesamiento se utilizó el programa FastQC con el fin de chequear el estado y calidad de los datos. En la Anexo II se puede apreciar la salida del programa FastQC. En éste se pueden observar los datos crudos emitidos por el secuenciador MiSeq de Illumina, así como los datos para todas las etapas de pre-procesamiento (trimming de adaptadores, primers y calidad). A groso modo, los datos crudos y procesados de todas las muestras analizadas mostraron, para cada posición de reads, puntajes de calidad superiores a 30. Además se vio que la distribución de calidad media por read presentó gran proporción de lecturas con scores mayores a 37 (entre 100000 y 600000 lecturas con Prhed Score entre 37 y 38). También se pudo apreciar cómo luego de ejecutado el trimado por calidad y largo de secuencia, la gran mayoría de los reads presentó longitudes entre 148 y 150 nucleótidos (Anexo II: Figura 1). Por otro lado, se puede apreciar, en mayor detalle, el porcentaje de reads que obtuvo cada muestra luego del trimado por calidad (los llamados reads de buena calidad). Se puede ver como todas las muestras presentaron porcentajes de reads de buena calidad superiores a 70% (Anexo II: Figura 2).

Alineamientos:

Se obtuvo la cantidad de reads de buena calidad que alinean contra los genomas de referencia (ver secuenciación por Sanger). Para ello se utilizaron los alineadores Bowtie, Bowtie2 y BWA. Los resultados de este análisis se ven resumidos en la Tabla 4 A y B (Anexo I). En la misma se puede ver detalladamente la cantidad de lecturas utilizadas por cada programa para cada muestra y gen analizados.

Bowtie2

Esta prueba mostró que del total de reads analizados (11598408), el 85.34% (9897694 reads) de los mismos fue del tipo paired-end, por otro lado se vio que el 14.66% (1700714 reads) fue unpaired reads, es decir singles reads. Para el caso del gen HA se alinearon un total de 7024418 pares de reads, lo que equivale a un 60.56% de la cantidad de reads originalmente analizados. Por otro lado, el alineamiento para el gen NA mostró que 3931618 pares de reads mapearon contra la referencia utilizada, lo que equivale a un 33.89% de la cantidad de reads analizados (Figura 5). Se detalla el conteo y porcentaje de reads específico de cada muestra y gen analizados en este trabajo (Anexo II: Figura 3 y 4, respectivamente, Anexo I: Tabla 4A).

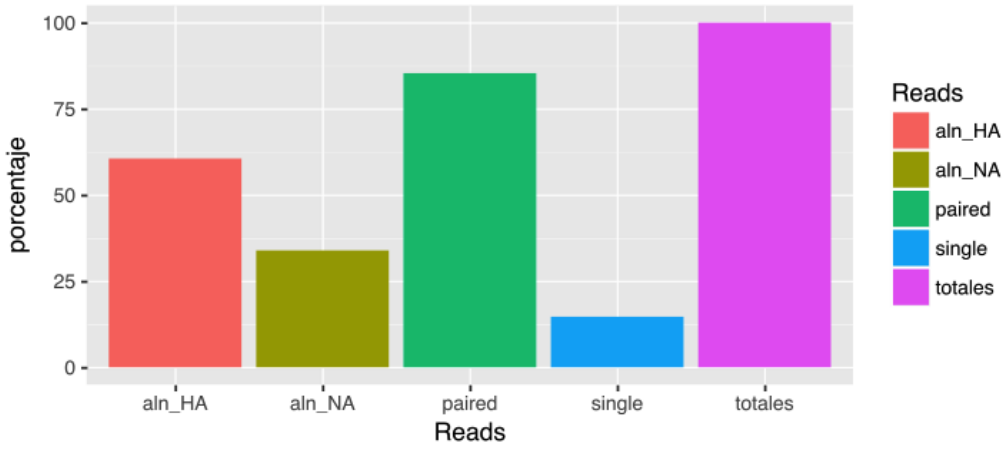


Figura 5. Resumen alineamiento Bowtie2.
Porcentaje de reads que mapean contra los genes HA y NA.

Bowtie

La cantidad de lecturas analizadas por este algoritmo fue exactamente la misma que para el caso del programa Bowtie2. Se obtuvieron 85.34 y 14.66% de paired y singles reads, respectivamente. El alineamiento contra las secuencias de referencia resultó en un 46.83% (5431869 pares de reads) que alineó contra HA y un 24.68% (2862371

pares de reads) que mapeó contra NA (Figura 6). El conteo y porcentaje de reads específico de cada muestra y gen se detalla en el Anexo II: Figura 3 y 4, respectivamente y en el Anexo I: Tabla 4A.

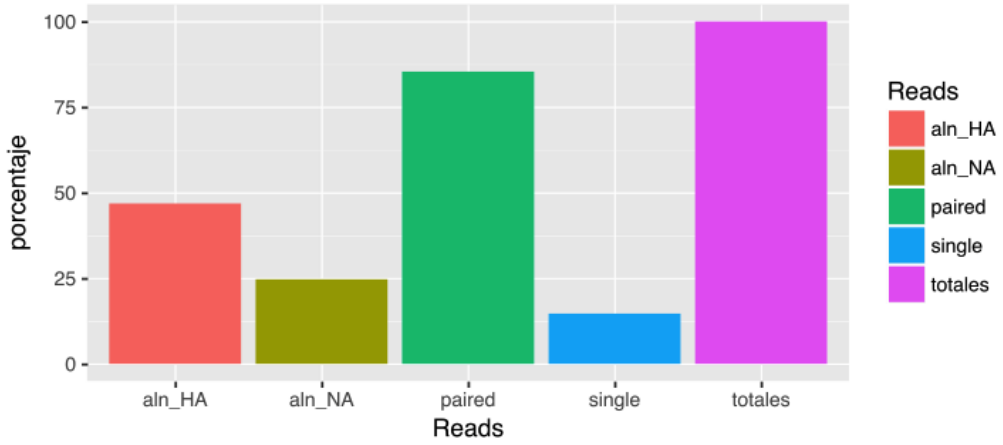


Figura 6. Resumen alineamiento Bowtie.
Porcentaje de reads que mapean contra los genes HA Y NA.

BWA

Este estudio mostró leves diferencias en la cantidad de reads evaluados por gen analizado. Por un lado, para el caso del gen HA, se vio que se procesó un total de 21509508 de reads, de éstos el 92.09% (19807560) fueron asignados a paired reads, mientras que un 7.91% (1701948) correspondieron a single reads. El alineamiento dio un total de 13.737.673 de lecturas mapeadas, lo que correspondió al 63.87% de la cantidad de reads inicialmente evaluados. Por otro lado, el procesamiento de las lecturas para el gen NA condujo a evaluar un total de 21.505.904 reads, de los cuales el 92.09% (19804171) fueron paired-end reads y el 7.91% (1701733) a singles reads. El mapeo de los reads a las referencias sumó un total de 7.581.050 lecturas, lo que equivale al 35.25% de la cantidad de reads analizados inicialmente (Figura 7). Se obtuvo el conteo y porcentaje de reads específico de cada muestra y gen analizados (Anexo II: Figura 3 y 4 respectivamente, y Anexo I: Tabla 4A).

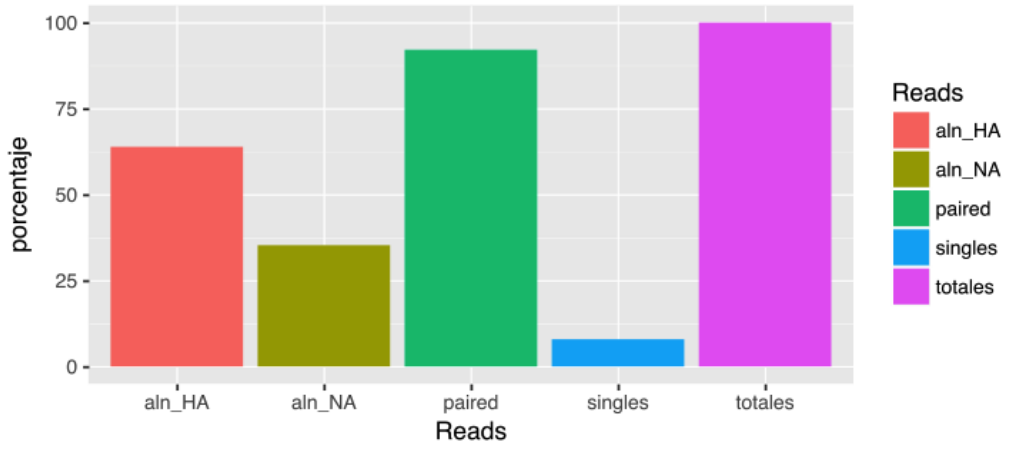


Figura 7. Resumen alineamiento BWA.
Porcentaje de reads que mapean contra el gen HA.

La comparación de los alineadores, en función de cada muestra analizada, mostró que para el caso de la HA y NA Bowtie2 alineó en promedio 61.6% y 32.9%, respectivamente. Por otro lado, Bowtie mapeó en promedio para HA 48.9% y para NA 32.1%. También se vio que BWA alineó en promedio para el gen HA 65.03 % y para el gen NA 34.2% (Anexo II, Figura 4).

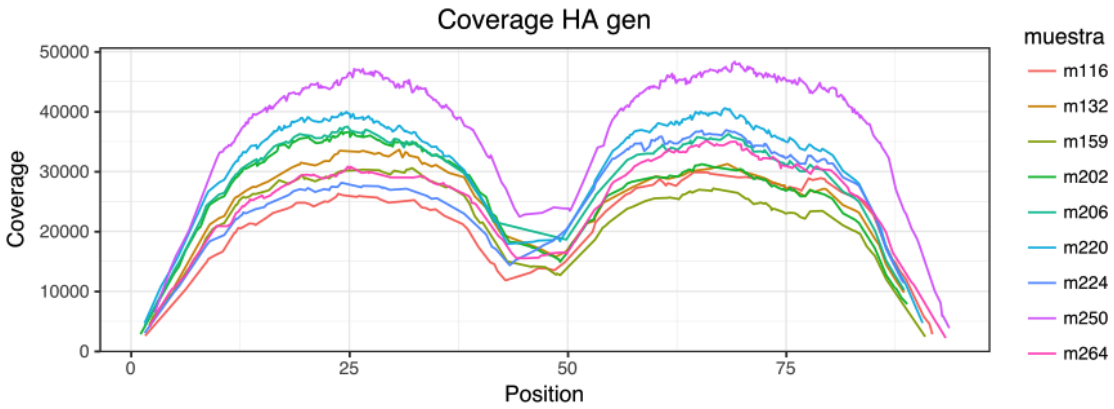
En base a estos resultados se decidió continuar los análisis usando los reads derivados de BWA ya que, además de tener un porcentaje de reads mapeados mayor, fue el alineador que mapeó mayor cantidad de reads en HA y NA. Esto se deduce debido a que el input para bowtie y bowtie2 presentó una cantidad de reads menor que el input para el caso de bwa (Anexo I, Tabla 4).

Los alineamientos generados con BWA fueron validados utilizando el programa qualimap [García-Alcalde y cols., 2012].

Cobertura y Calidad

La cobertura y calidad son parámetros que dan una idea de cuán bien un determinado target de secuencia fue secuenciado [Harismendy y cols., 2009]. En este sentido se obtuvieron, en términos generales, la cobertura y calidad de cada muestra y gen analizados. Para el gen HA la cobertura cayó en el rango de 10000 a 50000X, mientras que la calidad fue en promedio superior a un valor de 40 (Figura 8A y Figura 9A). Por otro lado, para el caso del gen NA la cobertura fluctuó en el rango de 10000 a 30000X y el valor de calidad fue en promedio de 45 (Figura 8B y Figura 9B).

(A)



(B)

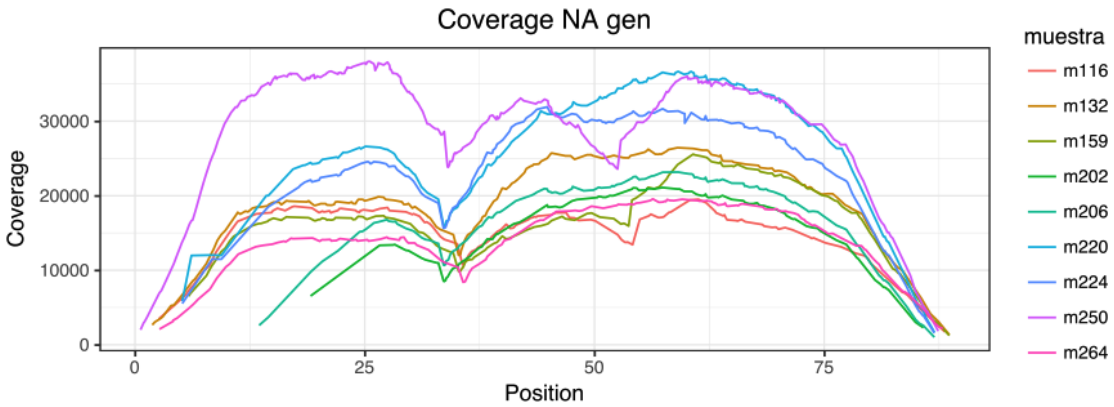
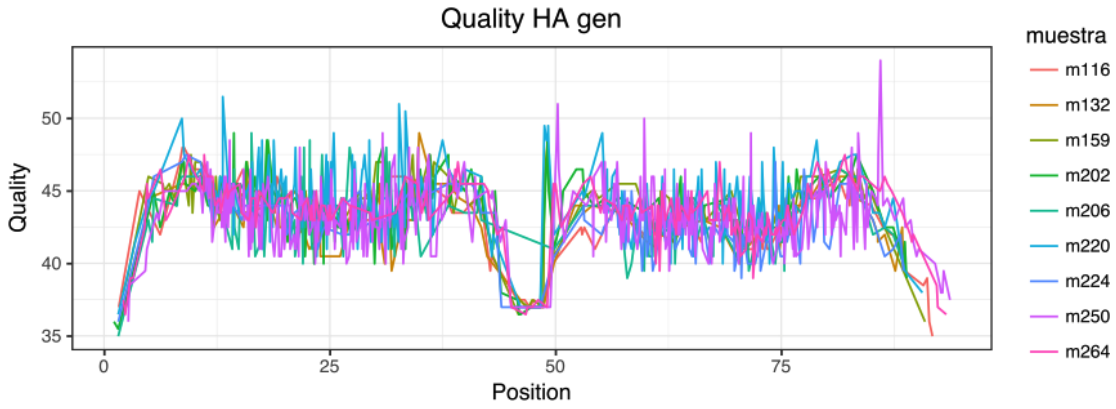


Figura 8. Cobertura del alineamiento con BWA.

(A) Cálculo de cobertura para el gen HA de todas las muestras analizadas.

(B) Cálculo de cobertura para el gen NA de todas las muestras analizadas.

(A)



(B)

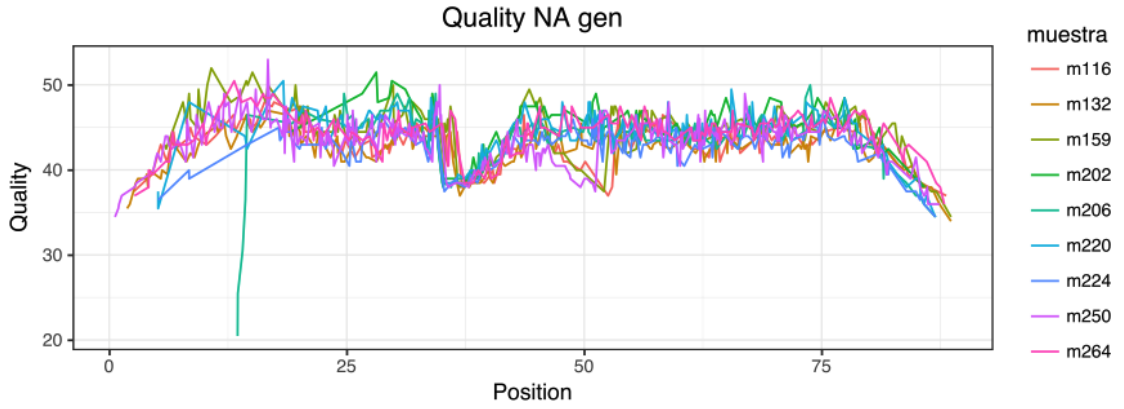


Figura 9. Calidad del alineamiento con BWA.

(A) Cálculo de calidad para el gen HA de todas las muestras analizadas.

(B) Cálculo de calidad para el gen NA de todas las muestras analizadas.

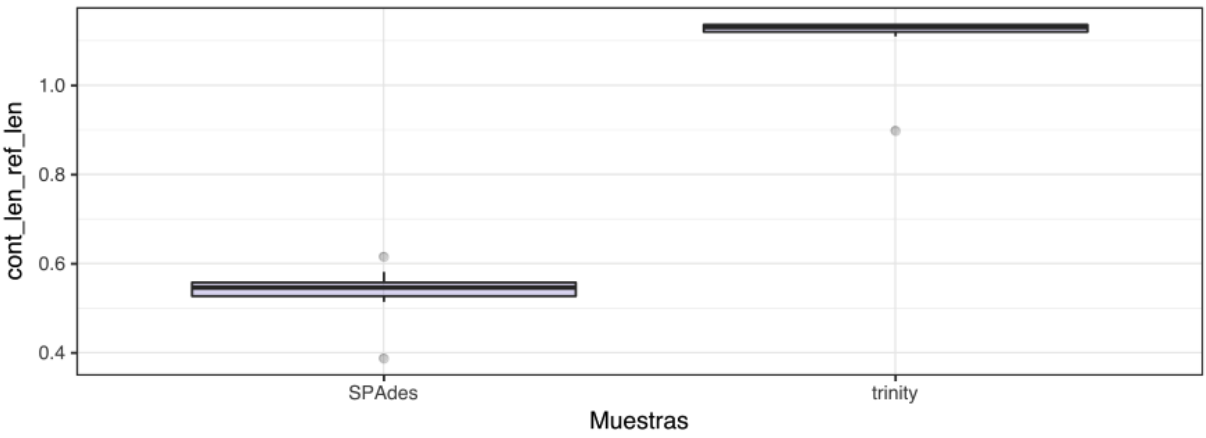
Más allá que la cobertura presentó cierta variación a lo largo de ambos genes, se obtuvieron valores superiores a 20000X en promedio, lo cual es muy bueno. Por otra parte, los valores de calidad también varían a lo largo de la secuencia de cada gen, sin embargo se obtuvieron en general valores de calidad superiores a 40, lo cual es excelente. Cabe mencionar la caída en cobertura y calidad que se aprecia aproximadamente en la sección media de los genes HA y NA. Este defecto podría deberse principalmente a un artefacto impuesto por la técnica de PCR, dado que ambos genes fueron amplificados en dos fragmentos de aproximadamente 800pb cada uno. Esto se explica por el hecho de que ambos, cobertura y calidad, caen drásticamente en los extremos de cualquier fragmento secuenciado.

Ensamblados:

Se seleccionaron los reads mapeados con BWA en la etapa de alineamiento y se realizó el ensamblado *de novo.* Se utilizaron los programas Trinity y SPAdes con todos sus parámetros por defecto. Se obtuvo para cada programa la proporción de largo de contigs en función del lago de gen analizado. Los resultados globales de este análisis se resumen en la Figura 10 (ver también: Anexo I, Tabla 5). El ensamblado con Trinity produjo contigs que cubren una sección mayor al 90% del largo de los genes HA de referencia (Figura 10a). Asimismo, el ensamblado para el gen NA produjo contigs con longitudes mayores al largo de las respectivas referencias utilizadas (Figura 10b). Por otro lado, para el caso del programa SPAdes se lograron construir contigs que cubren del 40 al 60% del gen HA (Figura 10a) y contigs que involucraron secciones del gen NA en un rango mayor al 50% (Figura 10b). Asimismo se detalla la relación entre largo de contig vs largo de gen específico de cada muestra y gen analizado (Anexo II, Figura 5).

La evaluación de los ensamblajes realizados por el programa Trinity fue realizado con la herramienta Quast [Gurevich y cols., 2013].

(a)



(b)

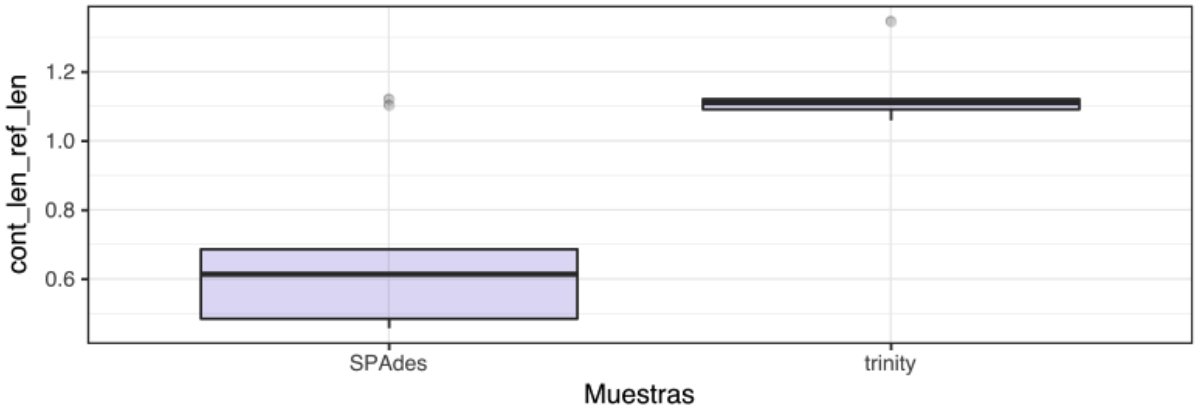


Figura 10. Resumen ensamblaje con los programas Trinity y SPAdes.
Porción relativa de genes ensamblados. (a) → HA ; (b) → NA.

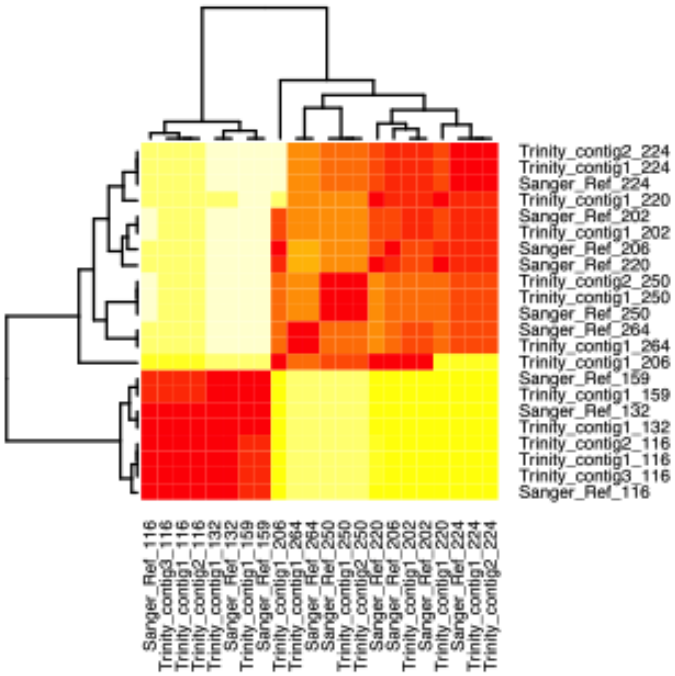
Si se analiza la relación entre largo de contigs en función del largo de la referencia utilizada, se puede ver cómo el programa Trinity produce contigs que rondan el orden de longitud de dichas referencias. Por otro lado, SPAdes produce en general contigs menores al 60% de la longitud de los genes de referencia utilizados. Más precisamente, si se observa esta relación para el caso del gen HA, se puede ver que en 8 de las 9 muestras, los contigs generados por el ensamblador Trinity alcanzan la totalidad de la longitud de las referencias. Incluso, se puede apreciar que los contigs generados presentan, en su mayoría, longitudes mayores a las referencias. La única excepción es la muestra 250 que cubre el 90% de su referencia. Por otro lado, los contigs producidos por el ensamblador SPAdes no alcanzan el 60% del largo de de referencia (Anexo II, Figura 5). Así también, la relación de longitudes para el caso del gen NA mostró que, en la totalidad de las muestras estudiadas, la cobertura de contigs generados con el programa Trinity fue mayor al 100%, es decir se obtuvo contigs con logitudes mayores a las respectivas referencias. Sin embargo, en este caso el programa SPAdes logra reconstruir, en 2 de las 9 muestras (m116 y m206), contigs de NA que alcanzan el 100% de sus referencias. Asimismo, el resto de las muestras analizadas cubren como máximo hasta el 65% de los genes de referencia (Anexo II, Figura 5).

Los resultados de los ensambladores, para cada gen y muestra en particular, apuntan a que en general Trinity logra capturar la longitud de gen deseado.

Generación de referencia basada en el ensamblado:

Se compararon los contigs, ensamblados en la etapa anterior por el programa Trinity, contra las secuencias de referencia generadas en este trabajo mediante el método de secuenciación de Sanger. Para ello se construyeron alineamientos que involucraron dichas secuencias. Posteriormente, éstos fueron utilizados para generar matrices de distancia con el método Maximum Composite Likelihood model [Tamura K., y cols. 2004], implementado en el programa MEGA5 [Tamura K., y cols. 2011]. Se muestran las matrices de distancia para los genes HA y NA. El rango de distancia entre pares de secuencias, las de Sanger contra los contigs deerivados de secuenciación masiva, se halla entre 0-0.0010 para HA y 0-0.0090 para NA (Figura 11A y 11B respectivamente).

(A)



(B)

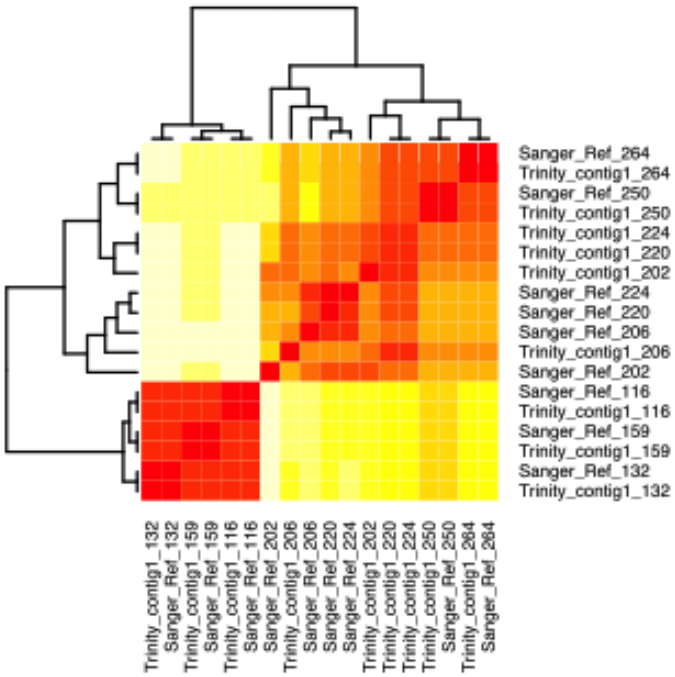


Figura 11. Generación de referencia basada en el ensamblado con Trinity.
Matrices de distancia de alineamientos entre secuencias generadas por el método de Sanger en comparación con los contigs ensamblados con el programa Trinity. (A) Heamglutinina; (B) Neuraminidasa. La escala de colores denota puntaje de distancia (distancia = 0 , rojo intenso; a medida que la distancia aumenta el color se va tornando mas claro). Margen derecho e inferior indican nombre de contigs y referencias. Margen izquierdo y superior cladogramas de agrupamiento.

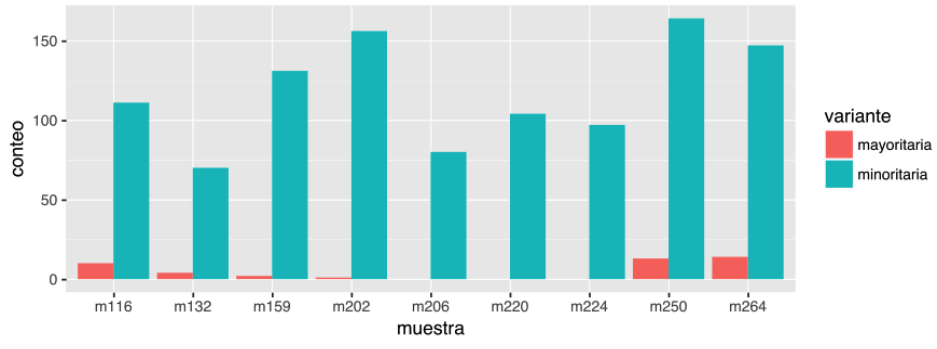
En el caso del gen HA el ensamblador Trinity produjo diferente número de contigs que varía entre las muestras. Para la muestra 116 los tres contigs generados muestran una distancia de 0.0010 en relación a su referencia. En el caso de las muestras 132, 159, 202, 206 y 264, para las cuales solo se ensambló un contig, no muestran diferencias con su respectiva referencia. En el caso de la muestra 220 para la cual se generó un único contig, éste presentó una distancia de 0.0010 con su referencia. Para las muestras 224 y 250 se ensambló dos contigs cada una, y su distancia fue de cero respecto a sus referencias. Por otro lado, para el caso del gen NA el ensamblador Trinity construyó un contig por muestra. Para el caso de las muestras 116, 132, 159, 250 y 264 se encontró una distancia de cero respecto a sus referencias. Asimismo, las muestras 202, 206, 220 y 224, presentaron un distancia de 0.007, 0.009, 0.004 y 0.005, respectivamente contra a sus referencias.

El análisis de distancias mostró que la gran mayoría de los contigs se parecen en gran medida a sus respectivas referencias, por lo que utilizar este tipo de abordaje y herramienta para la generación de secuencias consenso sería de utilidad cuando se cuenta con datos masivos.

SNPs: Variantes Mayoritarias y Minoritarias

Para cada muestra y gen analizados se identificaron variantes presentes en distinta frecuencia dentro de la población viral (Anexo I, Tabla 6. Cambios AAs). Estas variantes se clasifican en minoritarias si su frecuencia se encuentra entre 0.001 y 0.01, y en mayoritarias si su frecuencia es mayor a 0.01, a nivel poblacional [Isakov y cols. 2014]. Se representa globalmente el número total de variantes (conteo) para todas las muestras y genes procesados (Figura 11). Se obtuvo un rango amplio en el conteo de variantes tanto mayoritarias como minoritarias. Para el caso del conteo de variantes para el gen HA se reportó un rango entre 70-164 variantes minoritarias para las muestras analizadas, mientras que para el gen NA se reportó un máximo de 313 y un mínimo de 111 variantes minoritarias (Figura 11 A y B). Por otro lado, los conteos de variantes mayoritarias se ubicaron en el rango de 0-14 y de 0-20 para HA y NA, respectivamente (Figura 11 A y B).

(A)



(B)

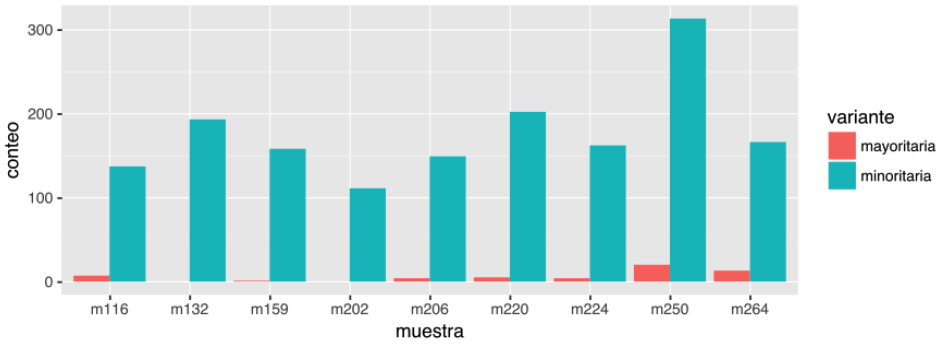


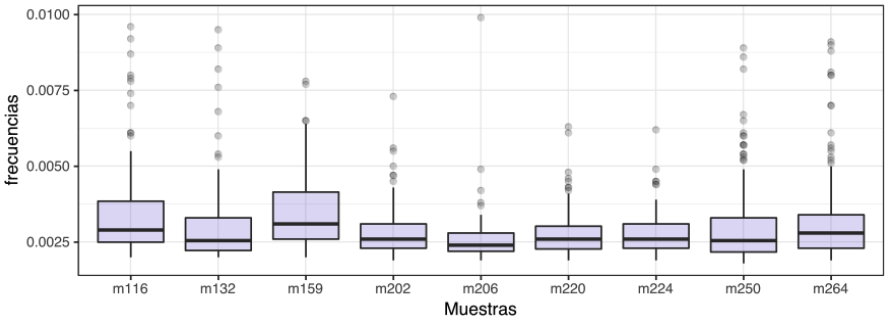
Figura 11. Conteo de SNPs de variantes mayoritarias y minoritarias.
(A) HA; (B) NA. Las variantes minoritarias se rotulan en color turquesa mientras que las mayoritarias se representan en color coral.

La gran cantidad de variantes encontradas en todas las muestras apoyan la idea de que los virus circulan como poblaciones virales heterogéneas. La variación en el conteo de variantes minoritarias, entre las muestras analizadas, señala que la diversidad viral dentro de cada población presenta cierta independencia una de la otra. En otras palabras, se puede ver que la distribución del conteo de variantes entre las muestras analizadas presenta cierta heterogeneidad. Sin embargo, respecto al conteo de variantes mayoritarias entre las muestras analizadas se ve menor grado de dispersión. Esto se explica por el hecho de que las cuasiespecies están compuestas por unas pocas variantes mayoritarias rodeadas de un enjambre de mutantes muy diverso.

Otro punto interesante es que a pesar de que el gen NA se llevó un 50 % menos reads que el gen HA (Figura 7), el conteo de variantes fue aproximadamente el doble (Figura 11).

Se obtuvo la frecuencia de las variantes minoritarias y mayoritarias para todas las muestras y genes analizados. Para el caso del gen HA, se encontraron variantes minoritarias principalmente en el rango de frecuencias de 0.0018-0.0037, sin embargo se encontraron para todas las muestras variantes minoritarias de hasta frecuencias de 0.01 inclusive. Se encontró que la distribución de frecuencias minoritarias varía ampliamente entre muestras (Figura 12A). Por otro lado, como era de esperar, las variantes mayoritarias muestran un distribución de frecuencias muchísimo más homogénea. Se encontraron frecuencias principalmente en el rango de 0.011-0.094, es decir entre 1 y 9 %, aproximadamente. Sin embargo, se encontraron variantes con frecuencias de 0.1101, 0.1199, 0.3506, 0.3657 y 0.9997 (Figura 12B).

(A)



(B)

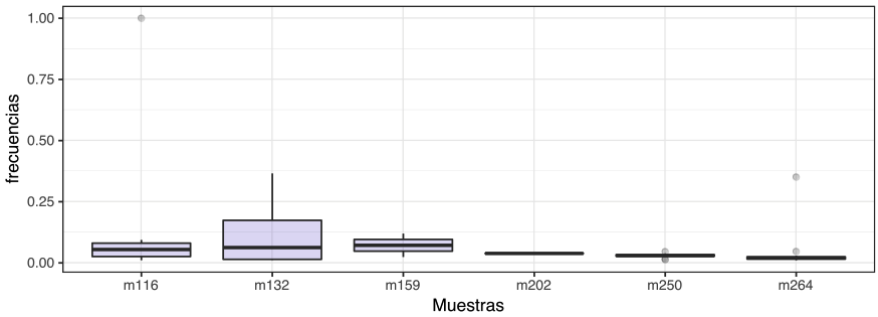
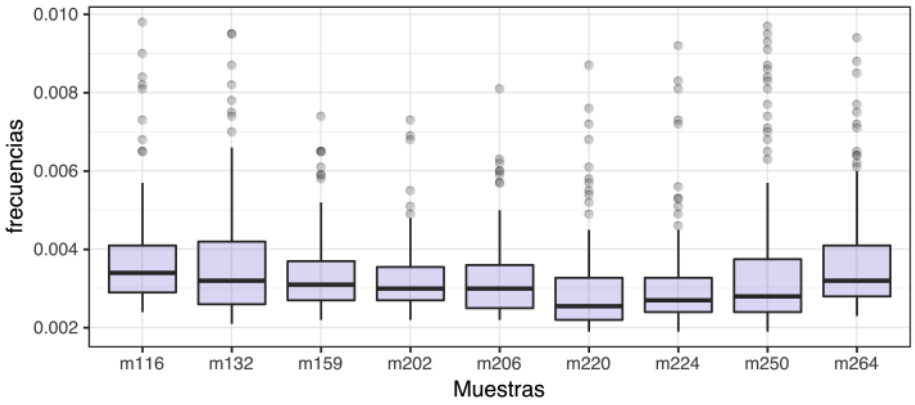


Figura 12. Variantes Minoritarias y Mayoritarias en HA.
(A) Frecuencias de variantes minoritarias; (B) Frecuencia de variantes mayoritarias.

El análisis de las variantes minoritarias en el gen NA reportó una población de frecuencias de variantes principalmente en el rango 0.0019-0.0042. Sin embargo, al igual que para el gen HA, se encontraron variantes minoritarias con frecuencias cercanas a 0.01 y una distribución general de frecuencias bastante heterogénea (Figura 13A). Por otro lado, se lograron detectar variantes mayoritarias con frecuencias principalmente en el rango de 0.0107-0.0521, es decir entre un 1-5 %, sin embargo, también se encontraron variantes con frecuencias de 0.9979, 0.9992, 0.9975, 0.9987, 0.9995, 0.9977 y 0.9971 (Figura 13B).

(A)



(B)

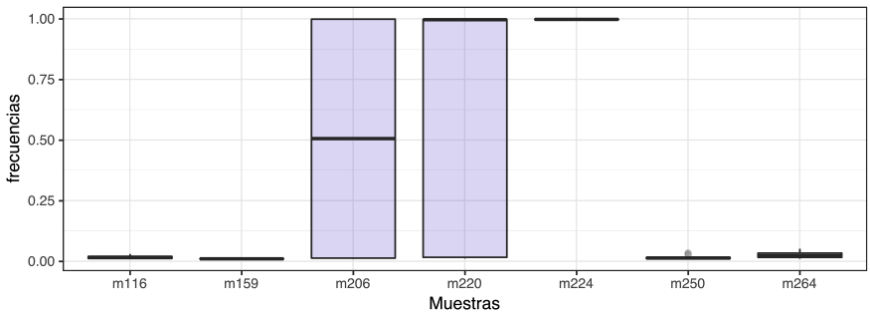


Figura 13. Variantes Minoritarias y Mayoritarias en NA.
(A) Frecuencias de variantes minoritarias; (B) Frecuencia de variantes mayoritarias.

El análisis de la cantidad de variantes (mayoritarias y minoritarias) y sus frecuencias da una idea de la variabilidad genética intra población. El rango de frecuencias que se vio en ambos genes fue similar, sin embargo el conteo de SNPs fue mucho mayor para el gen NA (Figura 11; Figura 12A; Figura 13A). En este sentido se podría decir que el gen NA presentaría mayor variabilidad intra población cuando se lo compara con su contraparte HA. El análisis de las frecuencias minoritarias de ambos genes discriminado por muestra arrojó que las variantes minoritarias para el gen HA presentan frecuencias restringidas principalmente a un valor de 0.0025 para la mayoría de las muestras (Figura 12A). Por otro lado, este mismo análisis pero respecto al gen NA, mostró que presenta variantes minoritarias con frecuencias ajustadas principalmente a un valor de 0.003, sin embargo presentó mayor dispersión en sus frecuencias en comparación con las del gen HA (Figura 12A; Figura 13A). En

términos generales, considerando la variabilidad como al conjunto de variantes y sus frecuencias, se observó mayor diversidad de variantes para el gen NA. Las diferencias encontradas en la cantidad de variantes y frecuencias entre HA y NA se ven apoyadas por el registro de las diferencias halladas en las tasas evolutivas entre ambos genes. Un estudio de análisis genómico sobre el reordenamiento y evolución de los virus influenza A H3N2, que circularon entre los años 1968 y 2011, mostró que el gen HA presentó una tasa evolutiva de 4.84×10^{-3} sustituciones/sitio/año, mientras que el gen NA mostró una tasa levemente mayor de 3.27×10^{-3} sustituciones/sitio/año. Esto indicaría una evolución más acelerada para el marcador NA [Westgeest KB y cols., 2014]. Tasas evolutivas similares fueron reportadas por otro trabajo, en el cual se estudiaron las relaciones filogenéticas de los genes HA y NA de VIA H3N2 entre cepas vacunales y estacionales [Kim J. Y cols., 2017].

El análisis de la diversidad poblacional mostró la presencia de variantes genéticas dentro de cada muestra y gen analizado. Las mismas podrían representar subpoblaciones caracterizadas por distintos mutantes que exploran el espacio de secuencia. Estos resultados se apoyan en trabajos previos que estudian la variabilidad viral intra paciente y revelan la presencia de variantes intra muestra que difieren en las propiedades antigénicas e incluso de resistencia a anitivirales [Ghedin y cols., 2009]. Sin embargo, hasta ahora no se obtuvieron indicios sobre qué relación existe entre los mutantes de las diferentes muestras analizadas. Conocer las relaciones entre dichas variantes genéticas es fundamental para entender la dinámica intra e inter cuasiespecies. Por tal motivo se dispuso a realizar un análisis de correlación para evidenciar vínculos entre variantes de las muestras que se procesaron.

Correlación entre muestras de pacientes de distintos años:

Entender como los VIA se transmiten y evolucionan, durante los brotes epidémicos y pandémicos, es crítico para mejorar las políticas de salud a nivel de la población humana. Reconstruir las relaciones entre las variantes puede ayudar a mejorar la composición de las cepas vacunales e incluso en el diseño de nuevos fármacos antivirales.

Los SNPs y sus frecuencias fueron utilizados para evaluar la posible relación de los mutantes inter-paciente. Con esto se busca de alguna forma evidenciar la presencia de variantes que se propagan a lo largo del tiempo. Para ello, se construyeron matrices de frecuencia de SNPs para cada gen y posición génica y se evaluó mediante el criterio de distancia “Euclidiana” usando el método de clustering de “Ward D.” Se

determinaron posibles relaciones entre las variantes de cada paciente visualizando los resultados mediante heatmaps.

El heatmap, para el gen HA, mostró ciertas posiciones genómicas para las cuales se encontró determinado SNP con una frecuencia similar. Como ser, el conjunto de muestras 159, 202, 132, 220, 206, y 224 mostraron un grupo de SNPs (A→G y T→C) posicionales que presentan frecuencias en el rango 0.0021-0.0059. De igual forma, se observó para las muestras 250 y 264 un pequeño grupo de variantes con frecuencias en el rango 0.0019-0.0276. Asimismo, se observa que ciertos conjuntos de muestras, como ser 202-220, 159-220, 202-220-206-224, comparten determinados cambios (Anexo I: Tabla 6; Figura 14A). También se ven determinadas variantes en posiciones particulares que son exclusivas de cada muestra. Entre estas se encontró un grupo de SNPs para la muestra 250, otro para la 116 y otro gran grupo de variantes para la muestra 264 (Figura 14A).

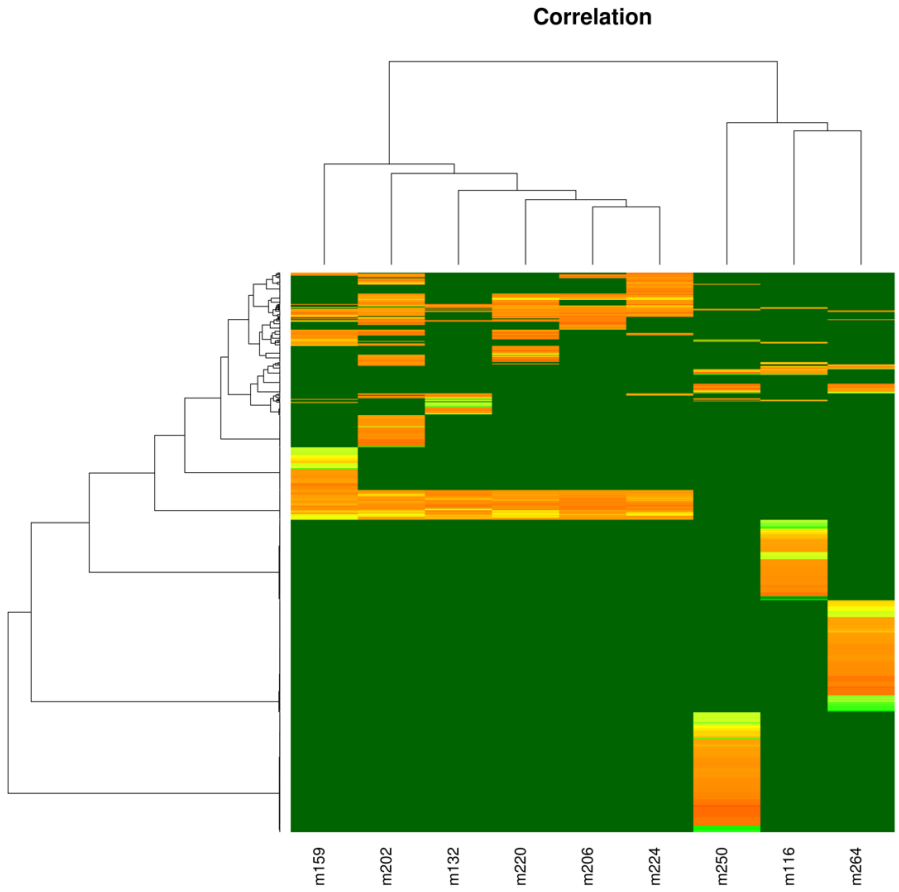
Se calculó, usando el criterio de distancia Euclidiana y método de clustering Ward.D, los heatmaps para el gen NA. Dicho análisis mostró un gran grupo de variantes posicionales conservadas entre las muestras 220, 224, 202 y 206. Fueron detectados otros conjuntos de variantes para ciertos grupos de muestras. Por ejemplo, se encontró un grupo para las muestras 132 y 159, otro para las muestras 250 y 116, otro conjunto de SNPs para el grupo 132-159-250, y otros para la dupla 250-264. No obstante, también se encontraron grupos de variantes en distintas posiciones característicos de cada muestra. Por ejemplo, se puede apreciar un gran grupo en la muestra 250, otro para la muestra 116 y finalmente un conjunto de SNPs en la muestra 264 (Anexo I: Tabla 6; Figura 14B).

Ahora bien, analizando las correlaciones de SNPs a nivel global se pudo ver que para el caso del gen NA se forman dos clados principales. En uno de ellos agrupan las muestras 220, 224, 202 y 206, todas circulantes durante el año 2012. Sin embargo, dentro de este clado se pudo ver que se forman dos sub-clados, uno para las muestras 220 y 224, y otro para las muestras 202 y 206. Por otro lado, el resto de las secuencias agrupan dentro del otro clado que también se subdivide. Dentro de éste último, se vio que las muestras 132 y 159 agrupan juntas (ambas del año 2011), mientras que la 250, 116 y 264 se agrupan en otro subclado. Asimismo, fue interesante evidenciar que la muestra 116 (año 2011) y la 264 (año 2013) agrupan juntas en un sub-clado (Figura 14B).

En esta misma linea, el análisis de correlación de SNPs a nivel global para el caso del gen HA mostró un cluster correspondiente a las muestras 250, 116, 264. Dentro de

éste se pudo ver, al igual que para el caso del gen NA, que la muestra 116 y 264 agrupan juntas en un sub-clado mientras que por afuera cae la muestra 250. El resto de las muestras agrupan en un clado totalmente separado que se subdivide. Dentro de éste último se encontró que las muestras 224, 206, 220, 132 y 202 presentan mayor relación entre ellas que con la muestra 159. Este grupo de muestras comparte determinados SNPs en sitios genómicos característicos (Anexo I: Tabla 6).

A)



(B)

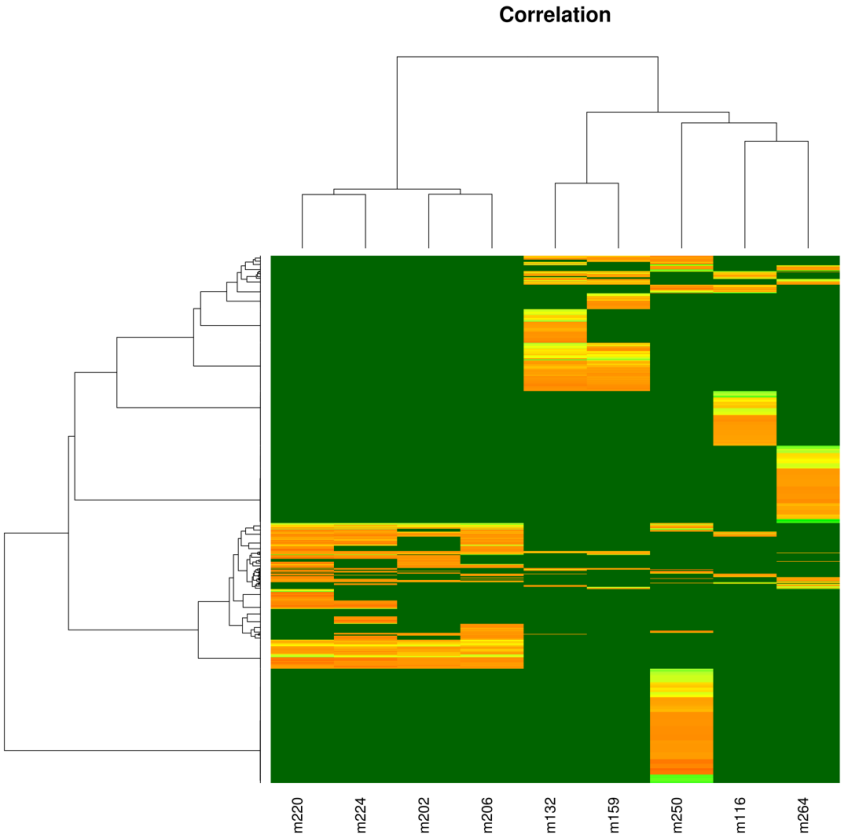


Figura 14. Correlación de SNPs y su frecuencia por posición genómica.

(A) y (B) heatmaps para los genes HA y NA, respectivamente. El degradé de color indica frecuencias (verde oscuro=1 – naranja=0.001).

El análisis de correlación de SNPs y sus frecuencias para las muestras de diferentes años no mostró fuerte agrupación de variantes respecto al año de circulación. De hecho, el caso más extremo se ve entre las muestras 116 y 264, que a pesar de ser muestras colectadas en años diferentes (2011 y 2013, respectivamente), el análisis de cada gen muestra que caen juntas. Sin embargo, este pequeño cluster (116-264) parece ser debido a un artefacto en el método ya que lo que comparten en su mayoría son las posiciones coloreadas en verde, que en realidad corresponde a sitios de secuencia conservados. Por otro lado, otro grupo de muestras, las colectadas durante el 2012 (202, 206, 220 y 224) y algunas del 2011 (132 y 159) comparten muchas variantes, aunque no necesariamente se cumple en todos los casos. Asimismo, se vieron cambios propios de cada muestra. Este estudio permitió identificar ciertos SNPs que circulan dentro de la nube de cuasiespecies de cada paciente. Estos SNPs

podrían estar jugando un rol significativo en el mantenimiento de las variantes de un evento epidémico a otro. Es necesario profundizar estos estudios para confirmar o refutar estas hipótesis.

Búsqueda de mutaciones de resistencia antiviral: mapeo de variantes sobre estructura cristalográfica:

Se han aprobado dos clases de fármacos anti VIA, los adamantanos y los inhibidores de la NA (INA, como oseltamivir). Estos fármacos se dirigen a los componentes virales del canal iónico M2 y NA, respectivamente. Sin embargo, VIA adquirió, en el transcurso de su evolución, resistencia contra ambas clases de fármacos al mutar dichos componentes virales. La mutación mayormente reportada que confiere resistencia al oseltamivir sobre la NA, de los subtipos H1N1 y H5N1, es el cambio H274Y. Por otro lado, los subtipos H3N2 y H7N9 presentan cambios de resistencia principalmente en las posiciones E119V y R292K [Meijer y cols., 2014]. Además, se ha visto en ciertos aislados de VIA H3N2 que llevan algunas simples mutaciones (I222K/T/R, N142S, D198E, S246G/N, N294S y G320E) y algunas dobles mutaciones (T156I+D213G, I222T+S331R, I222R/V+H274Y y S246N+H274Y) en su NA, sensibilidad reducida al oseltamivir y peraprevir [Eshaghi A. y cols., 2014; Hurt y cols., 2016]. Asimismo, la sensibilidad reducida o resistencia a zanamivir en aislados de H3N2 fue conferida por la mutación Q136K en la NA (Anexo I, Tabla 7) [Takashita y cols., 2015].

Las dos glicoproteínas de superficie principales de este virus juegan roles fundamentales para el ciclo viral. HA media la entrada del virus a la célula por su unión a residuos terminales de ácido siálico (N-acetyl neuraminic acid) de las proteínas de superficie del hospedero [Nayak DP., y cols 2009]. Por otro lado, NA facilita la liberación de la progenie viral por su actividad siálidasa. Esto lo hace a través del clivado de los ácidos siálicos terminales de las glicoproteinas virales y celulares [Nayak DP., y cols 2009].

La cabeza globular de la NA es un tetrámero compuesto de cuatro subunidades idénticas conformando una simetría tipo circular, (Figura 15A). Cada monómero contiene seis hojas beta compuestas de cuatro cadenas antiparalelas. El sitio activo de la NA está localizado centralmente en cada unidad y forma un bolsillo compuesto de residuos sumamente conservados entre todas las NA del Virus Influenza A [Colman PM. y cols., 1994]. Estos residuos delimitan la pared del bolsillo catalítico, e incluyen ocho residuos funcionales y once residuos estructurales. Los residuos funcionales

(R118, D151, R152, R224, E276, R292, R371 y Y406) interaccionan directamente con el “substrato” de ácido siálico, mientras que los residuos estructurales (E119, R156, W178, S179, D/N198, I222, E227, H274, E277, N2943 y E425) interaccionan con los residuos catalíticos estabilizando el sitio activo [Shtyrya YA. y cols., 2009], (Figura 15B). Los virus influenza A adquieren ciertas mutaciones alrededor del sitio activo que les otorga resistencia a los inhibidores de la NA. Dichos cambios son principalmente E119V, I222V, H274Y, R292K y N294S [Hussain y cols., 2017], (Anexo I: Tabla 7).

Con la finalidad de analizar la existencia de variantes, con potencial resistencia a los inhibidores de NA, se estudió la presencia de cambios aminoacídicos característicos del sitio activo y su entorno en la glicoproteína NA para las muestras de VIA (H3N2) procesadas. El análisis global de las variantes, halladas a partir de datos masivos, mostró la presencia de ciertos cambios aminoacídicos que podrían conferir resistencia a antivirales contra NA, (Anexo I: Tabla 7). Asimismo se mapearon, sobre la estructura cristalográfica 4gzp.pdb, los cambios aminoacídicos de potencial resistencia antiviral así como los residuos funcionales y estructurales del sitio activo (Figura 15). Esto se realizó para tener una idea general sobre los sitios estructurales con mayor relevancia biológica.

(A) (B)

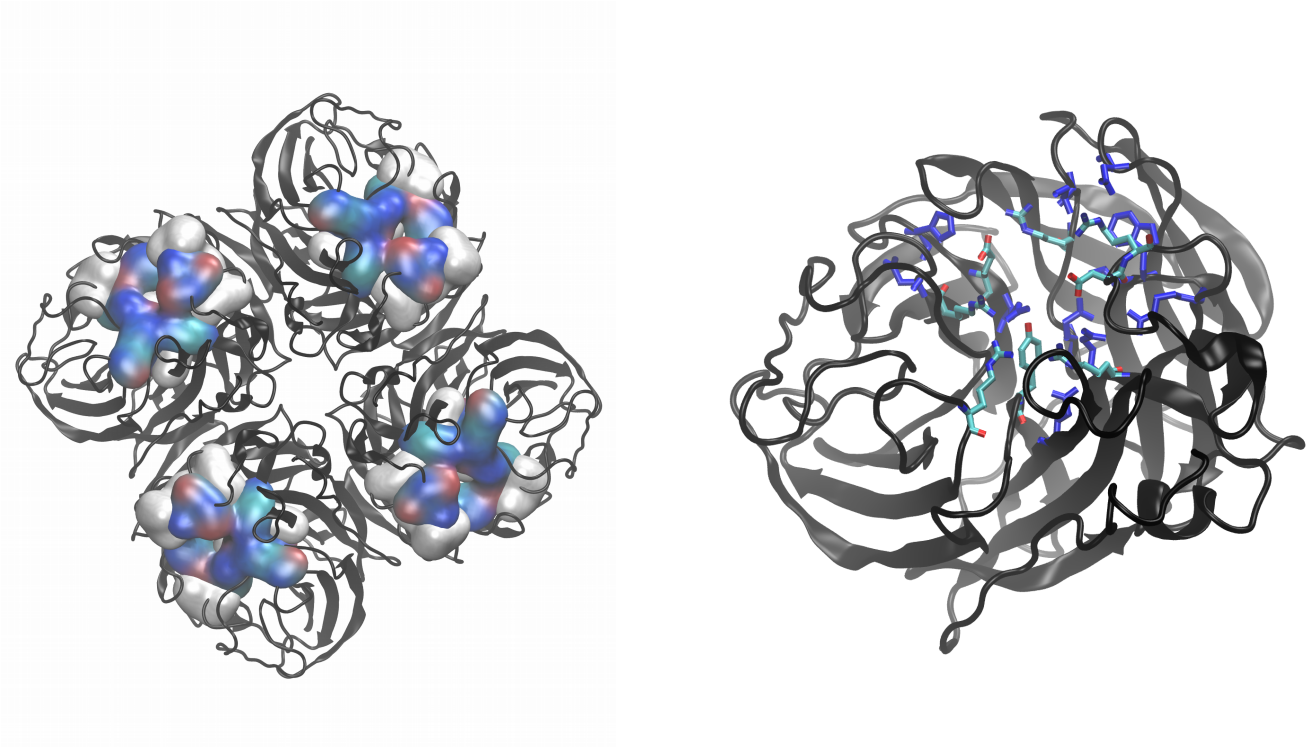


Figura 15. Mapeo de residuos funcionales y catalíticos sobre estructura cristalográfica de la NA. (A) Representación molecular de la estructura 4gzp.pdb completa. Visualización tipo superficie sobre los residuos estructurales (gris claro) y funcionales (tipo de átomo). (B) Representación molecular de la cadena A (estructura 4gzp.pdb). Se mapean los residuos estructurales (azul) y los del sitio catalítico (tipo de átomo).

El cambio en las propiedades fisicoquímicas, de ciertas posiciones aminoacídicas en la estructura de una proteína, puede tener consecuencias a nivel estructural [Schaefer & Rost, 2012]. También puede promover la resistencia a drogas antivirales [Hussain y cols., 2017].

Los residuos que forman la pared del bolsillo catalítico, y por ende interaccionan con el ác. Siálico, se pueden clasificar en tres grupos principales: básicos (R118, R152, R224, R292, R371), ácidos (D151, E276) y polar (Y406). Por otro lado, los residuos estructurales son aminoácidos ácidos (E119, D198, E227, E425, E277); básicos (R156); no polares (I222, W178) y polares (S179, N198, H274, N2943).

El contexto aminoacídico, en el cual un aminoácido se encuentre, va a ser crucial para la funcionalidad de la actividad de cierta proteína [Wang & Pollock, 2005].

Por tal motivo, con el objeto de tener una idea sobre la relevancia biológica que podrían presentar los cambios aminoacídicos sobre la estructura nativa de la proteína NA, se evaluó el tipo de cambio aminoacídico así como la distancia de dichos cambios

a los residuos funcionales y estructurales (Figura 16 y Figura 17).

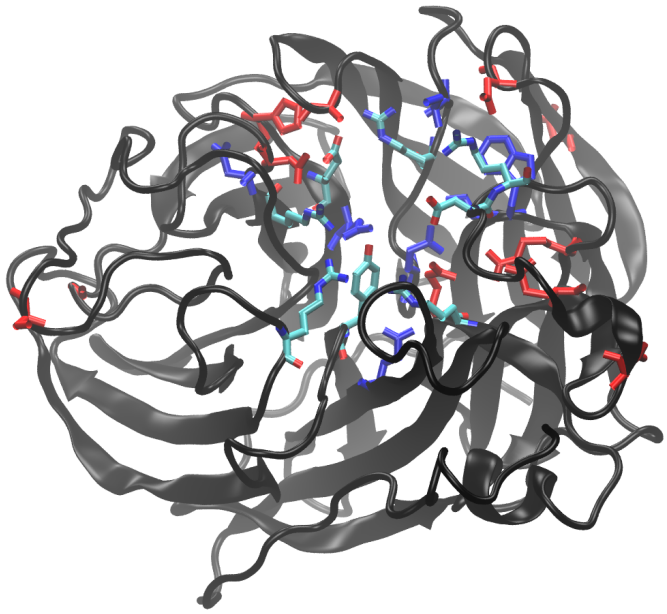


Figura 16. Mapeo de mutaciones sobre NA.
En rojo se resaltan las mutaciones encontradas en este estudio, que se sabe por bibliografía confieren posible resistencia contra INA, así como otros cambios relevantes. Se señalan los residuos estructurales (azul) y los aminoácidos del sitio catalítico (tipo de átomo).

Como se mencionó anteriormente, se han descripto ciertas mutaciones (E119V, Q136K, N142S, R292K, G320E, y la doble mutación I222T+S331R) en la NA que reducen notoriamente su sensibilidad a determinados antivirales, [Hussain y cols., 2017; Eshaghi A. y cols., 2014].

Según los cambios aminoacídicos encontrados en este estudio (Anexo I: Tabla 7) se pudo constatar que el cambio E119V no está presente en la muestra. Sin embrago, sí se encontró la sustitución E119G en siete de las nueve muestras analizadas (m116, m132, m159, m220, m224, m250 y m264) con cierta frecuencia minoritaria (0.0026, 0.0047, 0.0045, 0.0023, 0.0025, 0.0027, 0.0028). Es de destacar que el cambio E119G pasa de un aminoácido tipo ácido a uno tipo no polar y se encuentra a una distancia particular de ciertos residuos funcionales y estructurales. E119G se encuentra a 3.68 A de R118 y a 5.19 A de Y406, ambos funcionales. Por otro lado, se localiza a 4.37 A de

E227 (residuo de tipo estructural). El cambio a un aminoácido no polar y de pequeño tamaño como la glicina podría alterar los contactos con los residuos estructurales y del sitio catalítico, y modificar de cierta manera el bolsillo de la NA. El cambio H274Y no se detectó, sin embargo sí se logró identificar el cambio H274Q que mantiene sus propiedades fisicoquímicas (aminoácidos polares). Éste fue encontrado en la muestra m264 y presentó una frecuencia de 0.024. Además se identificó el cambio H274R (que pasa de polar a básico) en las muestras m116, m132, m159, m220, m224, m250 y m264 con las siguientes frecuencias 0.0068, 0.0021, 0.0061, 0.0020, 0.0023, 0.0071 y 0.0025, respectivamente. La posición 274 pertenece a los residuos estructurales del pocket, conservado entre todos los IAV [Hussain y cols., 2017]. El cambio H274R que se encuentra a 3.3 A de distancia del residuo E276 (sitio activo) podría modificar el ambiente electrostático, por transitiva cambiar el entorno aminoacídico, y por ende la funcionalidad de la enzima, posibilitando la evasión antiviral. Por otro lado, se encontró en el sitio activo el cambio R292K, que mantiene sus propiedades fisicoquímicas (aminoacidos básicos), en la muestra m220 con una frecuencia de 0.0026. Además también se evidenció el cambio R292G, que pasa de básico a no polar, en las muestras m132, m159, m206, m220, m224 y m250 con frecuencias 0.0028, 0.0024, 0.0023, 0.0024, 0.0021 y 0.0027, respectivamente. La posición aminoacídica 292 se encuentra a 2.98 A del residuo E277, a 5.4 A del residuo E276, a 3.10 A de Y406 y a 3.85 A de la posición R371. Cambios en la posición 292 que modifiquen el entorno electrostático podría llevar a la modificación de las distancias entre residuos y por ende alterar la actividad enzimática de la NA. Estos resultados se apoyan de estudios previos. En ellos se mutan ciertos residuos conservados en la NA por residuos con similares propiedades fisicoquímicas para mantener la carga original. Dicho estudio encontró una caída significativa en la actividad de la NA debido a las mutaciones introducidas [Yen H-L. y cols., 2006].

Por otro lado, se encontró el cambio N294S en la muestra 202 con una frecuencia de 0.0025. Dicho cambio no afecta la propiedad fisicoquímica del residuo, dado que ambos son aminoácidos polares, sin embargo podría modificar el vecindario aminoacídico. Otro cambio reportado para esa misma posición fue el N294D, que cambia a un residuo ácido, en las muestras m132, m220 y m264 a una frecuencia de 0.0025, 0.0022 y 0.0027, respectivamente. Asimismo, el cambio N294K, que pasa de un aminoácido polar a uno básico, se encontró en la muestra 264 a una frecuencia de 0.0026. La posición 294 se encuentra a 4.44 A del residuo E276 (que forma parte del sitio activo). Un cambio en este punto podría también llevar a la modificación de la

función (Figura 17).

Otros cambios detectados que también podrían modificar el ambiente electrostático del sitio activo son los cambio R156A y el D198G, con frecuencias en el rango 0.0022-0.0038. La posición 156 se encuentra a 2.8 A del residuo D151, ambos integrantes del core conservado de residuos del pocket de la NA. Estos cambios podrían alterar en gran medida el ambiente fisicoquímico de la enzima. El residuo 198 se encuentra a 3.69 A de R152, a 4.92 A de I222 y a 5.1 A de W178. Al igual que en el caso anterior, estas interacciones involucran a varios residuos del bolsillo catalítico. Cambios en el entorno electrostático podrían modificar, en cierta medida, la estabilidad y función de la proteína. Estos cambios también podrían evadir las presiones selectivas impuestas por las terapias antivirales.

El cambio G320E no fue detectado, sin embargo se encontró en la muestra 116, a una frecuencia relativamente considerable (0.0117), el cambio G320R. Esta sustitución modifica las propiedades fisicoquímicas de esa posición, pasa de un aminoácido no polar a uno básico. También se encontró la sustitución N142S (polar por polar), que no afecta las propiedades fisicoquímicas en gran medida, en las muestras 132 y 264 a una frecuencia de 0.0023 y 0.006, respectivamente. Sin embrago, se detectó en la muestra 264 el cambio N142D (polar por ácido) con una frecuencia de 0.0037. Otra mutación de resistencia a los INA es el cambio S331R. Dicha sustitución no fue reconocida entre las muestras analizadas, pero sí el cambio por una glicina. Esto se traduce en un cambio de un aminoácido polar a uno no polar. Dicha modificación se vio para todas las muestras analizadas con frecuencias en el rango de 0.0021-0.0038. La sustitución Q136K no se encontró, aunque un cambio similar (Q136R) se detectó para la gran mayoría de las muestras analizadas con frecuencias en el rango de 0.0026-0.0058. La sustitución S246N no se encontró entre los datos analizados, aunque sí se logró identificar el cambio S246A/T (polar a no polar) en todas las muestras analizadas con frecuencia de 1. Otros cambios que se buscaron fueron las dobles mutaciones asociadas a resistencia. No se encontró este tipo de sustituciones en las muestras analizadas. Sin embargo, sí se descubrieron los cambios T156A, D213G y D198G. Aquí todos cambian a aminoácidos polares, y las frecuencias de cambio rondaron el rango 0.0022-0.0038, (Anexo I: Tabla 7; Figura 17).

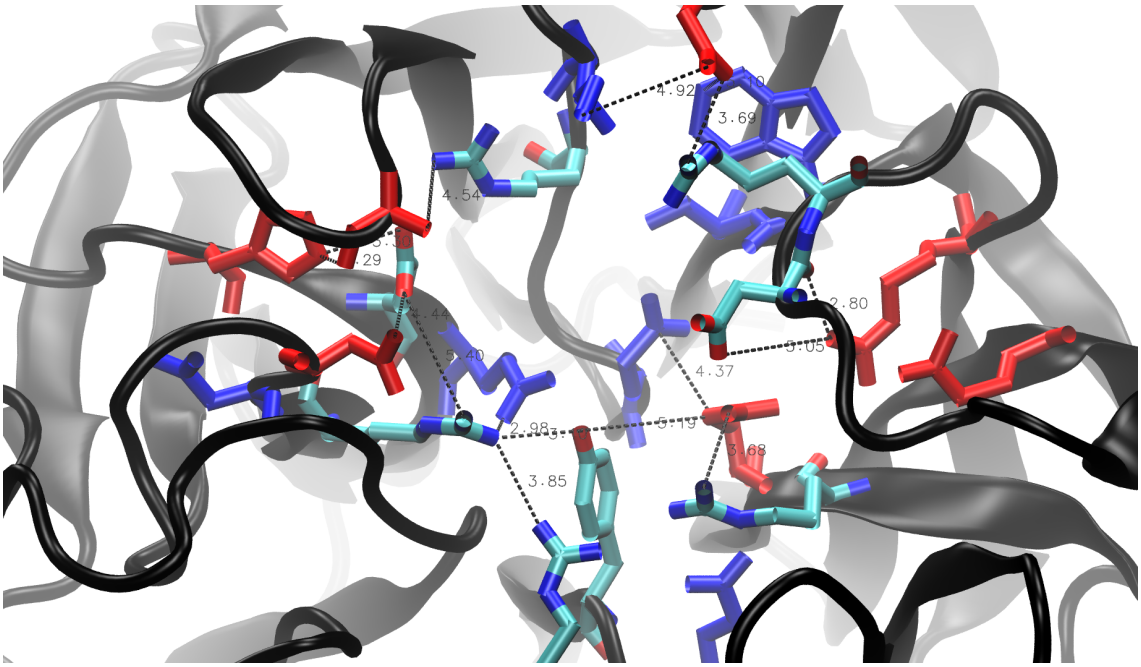


Figura 17. Distancias entre residuos mutados involucrados con el sitio activo de la NA.
Se puede apreciar los siguientes cambios y distancias: H274R/Q a 3.3 A de E276; S246A a 4.54 A de R224; S246A a 3.29 de H274; N294S/D/K a 4.44 A de E276; R292K/G a 5.4 A de E276, a 2.98 A de E277, a 3.85 A de R371 y a 3.10 A de Y406; R156 entre 2.8 y 5.05 A de D151; D198 a 3.69 A de R152, a 4.92 A de I222 y a 5.1 A de W178.

CONCLUSIONES

En base a los análisis realizados en el presente estudio, se logró el desarrollo de un pipeline bioinformático para el procesamiento de datos masivos derivados de la plataforma de secuenciación MiSeq de Illumina. Se obtuvieron reads de buena calidad y se pudieron ensamblar los genes de referencia para la Hemaglutinina y Neuraminidasa del VIA H3N2. La cobertura y calidad de los reads alineados resultó alta, lo que dio confiabilidad a los análisis. Se logró la construcción eficiente de las secuencias consenso de cada gen mediante el ensamblaje *de novo* de los datos de buena calidad. Dichas secuencias mostraron mínimas diferencias con las secuencias generadas por el método de secuenciación de Sanger lo que comprobó la eficiencia del método de ensamblaje. Se pudo identificar eficientemente los SNPs y sus frecuencias, que corresponden a las variantes mayoritarias y minoritarias dentro de cada muestra. No se encontraron fuertes indicios sobre una posible correlación de

SNPs entre muestras de distintos años, sin embargo sí se pudo ver cierto patrón de correlación entre muestras de un mismo año. Conocer las relaciones entre dichas variantes genéticas es fundamental para entender la dinámica intra e inter cuasiespecies. Finalmente se pudieron identificar cambios a nivel del sitio activo de la NA y su entorno que podrían alterar el ambiente electrostático y por ende la actividad enzimática de la NA. Estos cambios podrían estar participando en proceso de evasión de la respuesta inmune así como en mecanismos de resistencia a los antivirales. Entender como los VIA se transmiten y evolucionan, durante los brotes epidémicos año a año, es fundamental para mejorar las políticas de salud a nivel de la población humana. El estudio constante de las enfermedades virales puede ayudar a mejorar la composición de las cepas vacunales y contribuir al desarrollo de nuevos fármacos antivirales.

Capítulo 2

Implementación de algoritmos bioinformáticos para el ensamblado de haplotipos y estimación de sus frecuencias

RESUMEN

La gran mayoría de las investigaciones respecto a las cuasiespecies virales se han llevado a cabo mediante el análisis de los SNPs individuales. Uno de los mayores obstáculos en Bioinformática hoy en día, y de especial interés en la Virología, es la reconstrucción de los haplotipos virales a partir de un conjunto de reads.

Los errores de secuenciación en los reads, la corta longitud de los mismos, así como la corta distancia genética entre las variantes, hacen a la reconstrucción de la cuasiespecie un problema muy complejo de resolver. La reconstrucción de una cuasiespecie involucra tanto el ensamblaje de secuencias individuales dentro de una población así como la estimación de sus frecuencias.

La gran diversidad genética presente en poblaciones virales afecta sustancialmente las terapias antivirales y el diseño de vacunas apropiadas. Estas complicaciones motivan el desarrollo e implementación de algoritmo para el estudio en profundidad la variabilidad intra cuasiespecie. Aproximaciones recientes, sobre la resolución del problema de reconstrucción de haplotipos intra cuasiespecies incluyen varios tipos de métodos, como ser: inferencia Bayesiana, aproximaciones Bayesianas no-paramétricas basadas en modelos mixtos de procesos de Dirichlet y métodos basados en modelos de Markov ocultos.

La implementación de varios algoritmos de reconstrucción permitió evidenciar que la cantidad de haplotipos ensamblados varía entre los programas utilizados. Se estimó la frecuencia de cada variante y se encontró en todas las muestras, para ambos genes, la presencia de variantes mayoritarias y minoritarias. Un análisis multidimensional mostró la presencia de posibles subpoblaciones virales dentro de las cuasiespecies. Asimismo, un análisis por mapas de calor evidenció relaciones intra-cuasiespecie y posibles realaciones inter-cuasiespecies.

Conocer las variantes intra-cuasiespecie y sus relaciones, permiten brindar mayores

herramientas a la población con el fin de dilucidar los problemas específicos de cada paciente y poder diseñar terapias antivirales paciente específicas.

INTRODUCCIÓN

Con la reciente aparición de las tecnologías de NGS se pudieron superar las limitaciones impuestas en el secuenciado de clones por el método de Sanger [Beerenwinkel & Zagordi, 2011]. Este tipo de metodología proporciona un análisis muchísimo más detallado de la variabilidad viral [Borucki y cols., 2013; Eriksson y cols., 2008; Kampmann y cols., 2011; Morelli y cols., 2013]. La gran profundidad de las nuevas plataformas de secuenciación permitó la investigación de la evolución viral a una escala intra e inter hospedero [Schönherz y cols., 2016]. Con el uso de estas nuevas tecnologías de secuenciación es posible detectar mutaciones de baja frecuencia además de brindar información sobre la estructura poblacional, es decir, el conjunto de variantes dentro de la población y su frecuencia relativa [Beerenwinkel & Zagordi, 2011].

Uno de los mayores obstáculos o problemas en Bioinformática hoy en día es la reconstrucción de los haplotipos virales a partir de un conjunto de reads. Sin embargo, la aparición de las NGS nos permitió la detección de SNPs de haplotipos, incluso si presentan baja frecuencia. No obstante, hay dos problemas principales: primero es necesario distinguir entre SNPs reales de aquellos correspondientes a errores de secuenciación y segundo es necesario determinar qué SNPs ocurren dentro de un mismo haplotipo. Claramente los SNPs dentro de un mismo haplotipo no podrán ser inferidos si la distancia entre SNPs excede el largo de read observado [Schirmer M., y cols. 2014].

Hasta la fecha la gran mayoría de las investigaciones en relación a las cuasiespecies se han llevado a cabo mediante el análisis de las variantes o “single nucleotide variant” (SNV) presentes en una población viral [Wang C. y cols., 2007; Zagordi O. y cols., 2010a; Skums P y cols., 2012; Macalalad AR. y cols., 2012]. Este tipo de abordaje es muy útil en la identificación de variantes para todas las posiciones genómicas individuales y obtiene en forma precisa la frecuencia o proporción de una variante determinada. Sin embargo, estos abordajes no permiten la dilucidación de los haplotipos o variantes genómicas completas presentes en la cuasiespecie en un momento dado. La reconstrucción de una cuasiespecie involucra tanto el ensamblaje de secuencias individuales dentro de una población así como la estimación de sus

frecuencias.

Los errores de secuenciación en los reads, la corta longitud de los mismos, así como la corta distancia genética entre las variantes, hacen a la reconstrucción de la cuasiespecie un problema muy complejo de resolver. Aunque conceptualmente es similar al problema de resolver haplotipos individuales, la reconstrucción de la nube de mutantes presenta una mayor dificultad. En relación a lo mencionado anteriormente, el número de haplotipos individuales dentro de una cuasiespecie es a priori desconocido y las mutaciones puntuales son en general multialélicas en lugar de ser bialélicas [Töpfer A. y cols., 2014]. Desde hace un tiempo a la fecha se han desarrollado varios algoritmos para la reconstrucción de las variantes intra cuasiespecie, todos basados en el mapeo de un genoma de referencia y en grafos solapantes [Jojic V. y cols., 2008; Eriksson N. y cols., 2008; Prosperi M, y cols., 2011; Beerenwinkel & Zagordi, 2011; Zagordi O. y cols., 2010a; Beerenwinkel N. y cols., 2012; Mancuso N. y cols., 2011; Huang A. y cols., 2012; Westbrooks K. y cols., 2008; Zagordi O. y cols., 2011]. Estos algoritmos dependen de varios factores, como ser, el diseño experimental, la tecnología de secuenciación utilizada y obviamente la estructura poblacional de la cuasiespecie. En general la reconstrucción puede ser tanto local como global. La reconstrucción local implica la estimación del número local de haplotipos únicos y, concomitantemente, la corrección de errores de secuenciamiento. Para estas cuestiones han sido propuestos métodos de clustering probabilístico [Eriksson N. y cols., 2008; Macalalad AR. y cols., 2012; Quince C. y cols., 2011; Zagordi O. y cols., 2010b] y estadísticas de k-mean [Skums P. y cols., 2012]. La reconstrucción Global es más demandante en términos de tiempo computacional ya que requiere del ensamblaje de los reads, unos contra otros, sin la ayuda de un genoma de referencia [Salzberg SL. y cols., 2012]. El rendimiento de la reconstrucción global de haplotipos tiene una dependencia multifactorial. Estas incluyen la diversidad subyacente de la población, la distribución de los errores de amplificación y secuenciamiento, el largo de read y la distribución de la cobertura de reads a lo largo del genoma secuenciado [Schirmer M. y cols., 2014; Zagordi O. y cols., 2012; Prosperi MC. y cols., 2013].

Aproximaciones recientes, sobre la resolución del problema de reconstrucción de poblaciones de cuasiespecies, incluyen métodos de inferencia Bayesiana tal como ShoRAH [Zagordi O. y cols., 2011] y QuRe [Prosperi & Salemi, 2012]. Por otro lado también están las aproximaciones Bayesianas no-paramétricas, basadas en modelos mixtos de procesos de Dirichlet [Prabhakaran S. y cols., 2014], donde encontramos a PredictHaplo. También están: Quasirecomb basado en modelos de Markov ocultos

[Töpfer A. y cols., 2013] y HaploClique basado en técnicas de enumeración de máximo-clique en los gráficos de alineamiento de reads [Töpfer A. y cols., 2014]. Finalmente están los métodos heurísticos basados en graph-coloring como VGA [Mangul S. y cols., 2014], y los métodos de reconstrucción de ensamblaje de-novo asistidos por una secuencia de referencia como ViQuaS [Jayasundara D. y cols., 2014].

En general estos métodos pueden ser categorizados en aquellos basados en grafos de reads [Töpfer A. y cols., 2014; Mangul S. y cols., 2014], los basados en inferencias probabilísticas [Zagordi O. y cols., 2011; Prosperi & Salemi, 2012; Prabhakaran S. y cols., 2014; Töpfer A. y cols., 2013], abordajes basados en teoría combinatoria mediante el análisis de los grafos de solapamiento de reads [Eriksson N. y cols., 2008; Astrovskaya I. y cols., 2011; Mancuso N. y cols., 2011; O'Neil ST, & Emrich SJ, 2012] y las técnicas basadas en el ensamblaje de-novo [Jayasundara D. y cols., 2014].

Los métodos basados en grafos de reads [Töpfer A. y cols., 2014; Mangul S. y cols., 2014] utilizan aproximaciones combinatorias. Esto se lleva a cabo para el análisis de los grafos, en donde los vértices representan los reads y las aristas a los nodos correspondientes a reads que solapan. Específicamente, Töpfer y colaboradores (2014) formulan el problema de reconstrucción de una cuasiespecie como el de enumerar los cliques máximos en el grafo mencionado anteriormente, mientras que Mangul (2014) y Huang (2011) utilizan restricciones estructurales.

En los métodos probabilísticos se formula el problema de reconstrucción de una cuasiespecie como el problema de inferir variables ocultas que modelan la abundancia de ciertas variantes virales. En particular, PredictHaplo [Prabhakaran S. y cols., 2014] utiliza un modelo de mezcla infinito (infinite mixture model) para determinar el número de especies y reconstruir cada haplotipo mediante la maximización de la probabilidad de los reads observados. Por último, Jayasundara y colaboradores (2014) lleva a cabo la reconstrucción local de haplotipos. Esto lo hace mediante un ensamblaje de novo asistido por un genoma de referencia. Llega a una solución global a través de solapamientos concordantes con los haplotipos reconstruidos de forma local.

Como se comentó en la parte introductoria general de esta tesis, la inmensa diversidad genética presente en poblaciones virales afecta en gran medida las terapias con drogas antivirales y hace muy complejo el diseño de vacunas apropiadas. Dichos obstáculos motivan el desarrollo e implementación de este tipo de algoritmo con el fin de conocer en profundidad la variabilidad intra cuasiespecie. Mediante ello, será

posible colaborar en estos asuntos, tan importantes para la salud humana y animal, como por ejemplo en el desarrollo de terapias antivirales personalizadas. Para llegar a este objetivo es necesaria la reconstrucción de la nube de variantes. Como ya se comentó, involucra tanto el ensamblaje de las secuencias individuales así como la estimación de su abundancia o proporción en la muestra.

OBJETIVOS

Objetivo General:

Implementación de algoritmos bioinformáticos para la reconstrucción de las cuasiespecies virales intra hospedero de muestras de pacientes uruguayos infectados con el Virus Influenza A subtipo H3N2.

Objetivos Específicos:

1) Reconstrucción de cuasiespecies virales.

2) Estimación de frecuencias intra cuasiespecie.

3) Evidenciar relaciones entre haplotipos intra e inter muestra. Construcción de un análisis de Escalamiento Multidimensional y heatmap de correlación.

Reconstrucción de cuasiespecies:

La reconstrucción de las variantes intra cuasiespecie siguió tres abordajes diferentes. Se utilizó QuRe v0.9994 (http://sourceforge.net/projects/qure/) [Prosperi & Salemi, 2012], PredictHaplo (http://bmda.cs.unibas.ch/software.html) [Prabhakaran S. y cols., 2014] y Quasirecomb (https://github.com/cbg-ethz/QuasiRecomb) [Töpfer A. y cols., 2013].

El método de reconstrucción QuRe, es un programa especificamente desarrollado para analizar reads mayores a 100 pares de bases. El programa construye alineamientos de los fragmentos contra un genoma de referencia. Busca una división óptima del genoma en ventanas deslizantes, basándose en la cobertura y en la diversidad presente en la muestra. El mismo intenta reconstruir todas la secuencias individuales de la cuasiespecie viral utilizando su prevalencia. Esto lo hace mediante un algoritmo heurístico que machea distribuciones multinomiales de distintas variantes virales solapadas a través la división del genoma. QuRe trae un método de corrección de errores de Poisson y un método de clustering probabilistico post-reconstrucción. Ambos están parametrizados en las tasas de error de los sitios homopoliméricos y no-homopoliméricos. En resumen QuRe utiliza un método basado en construcción de grafo solapante sobre ventanas deslizantes. El mismo selecciona variantes candidatas mediante el uso de un algoritmo basado en solapamientos consistentes y en la similaridad de la distribución de frecuencias de las variantes en cada ventana.

El método de reconstrucción PredictHaplo es un software que busca la reconstrucción de haplotipos a partir de datos masivos. Desde una perspectiva estadística, este programa considera el análisis de datos como un problema de agrupamiento no estándar. Esto es debido a la falta de similitud entre pares de lecturas no superpuestas. Para superar este problema, propaga un Proceso de Dirichlet actualizando secuencialmente la información previa de análisis locales sucesivos. El modelo se verifica utilizando datos de secuencia simulada y real.

El método de reconstrucción Quasirecomb usa modelos de Markov ocultos. Presenta una implementación del algoritmo EM (Expectation Maximization). Este último se utiliza para encontrar estimaciones máximas a posteriori de los parámetros del modelo. También utiliza un método para estimar la distribución de cepas virales en la cuasiespecie.

Análisis de escalamiento Multidimensional (MDS):

El MDS es un método de análisis de una matriz de distancia establecida sobre un conjunto de individuos, en este caso variantes reconstruídas. El mismo tiene como objetivo modelizar las proximidades entre los individuos, de tal modo que pueda representarlos lo más exactamente posible en un espacio de baja dimensión (generalmente 2 dimensiones).

Reconstrucción de cuasiespecies virales:

Actualmente existen gran cantidad de algoritmos de reconstrucción de las variantes intra cuasiespecie [Jojic V. y cols., 2008; Eriksson N. y cols., 2008; Prosperi MC. y cols., 2011; Beerenwinkel & Zagordi, 2011; Zagordi O. y cols., 2010a; Beerenwinkel N. y cols., 2012; Mancuso N. y cols., 2011; Huang A. y cols., 2012; Westbrooks K. y cols., 2008; Zagordi O. y cols., 2011], y se han implementado varios de ellos en programas bioinformáticos [Zagordi O. y cols., 2011; Prosperi & Salemi, 2012; Prabhakaran S. y cols., 2014; Töpfer A. y cols., 2013; Töpfer A. y cols., 2014; Mangul S. y cols., 2014; Jayasundara D. y cols., 2014].

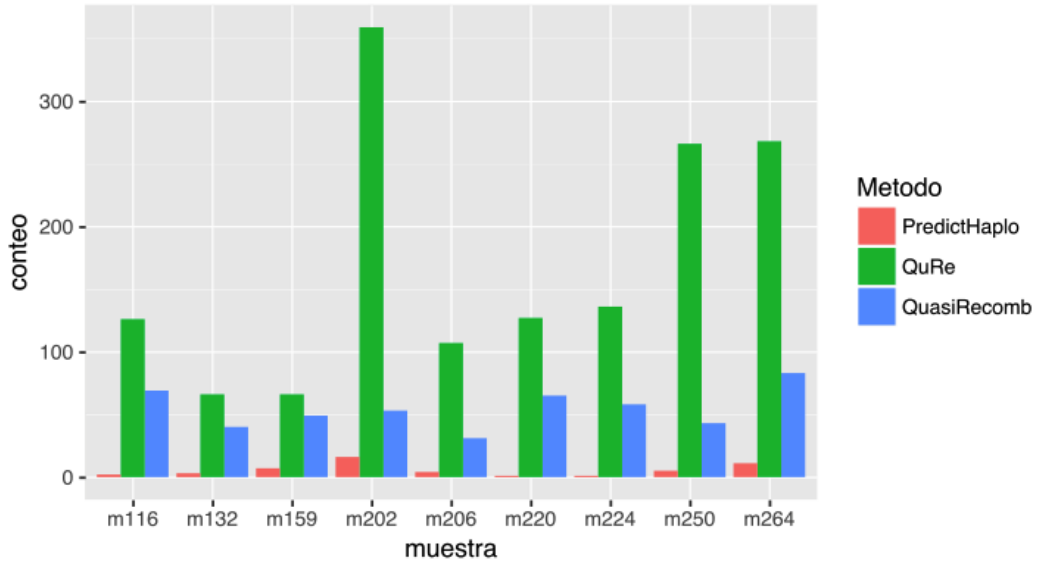
Estudios previos han utilizado varios de estos algoritmos para caracterizar diferentes haplotipos dentro de las poblaciones virales así como para la identificación de sus frecuencias [Giallonardo y cols., 2014]. Con el objetivo de dilucidar las variantes completas intra cuasiespecie, para los genes HA y NA, se implementaron tres algoritmos de reconstrucción utilizando los datos masivos de buena calidad del Capítulo 1. La evaluación de los algoritmos de reconstrucción de cuasiespecies utilizados en el presente trabajo mostró gran diversidad en términos de cantidad de variantes ensambladas (Figura 1, Anexo I: Tabla 8). Este resultado se ajusta con decenas de estudios previos que comparan varios programas de reconstrucción de haplotipos [Astrovskaya I. y cols., 2011; Mancuso N. y cols., 2011; Soyeon Ahn & Haris Vikalo, 2017].

Según los análisis llevados a cabo en el presente trabajo, el software QuasiRecomb produjo un rango entre 39 y 131 variantes ensambladas para el gen NA y un rango entre 31 y 83 variantes para el gen HA. Por otro lado, el algoritmo de reconstrucción PredictHaplo ensambló para HA un rango entre 1-16 haplotipos y para NA un rango entre 4-29 mutantes. Por último, el programa QuRe ensambló para el gen HA un total de variantes que caen en el rango 66-359, mientras que para el gen NA reconstruyó variantes en un rango de 78-273 haplotipos (Figura 1, Anexo I: Tabla 8). El programa QuRe [Prosperi & Salemi, 2012] fue el más eficiente en términos de cantidad de variantes reconstruídas. Esto podría deberse al hecho de que QuRe es un programa específicamente desarrollado para analizar reads mayores a 100 pares de base, lo que ajusta completamente con la longitud promedio de read (90-151 nt) analizado en este estudio [Anexo II].

La cantidad de variantes ensambladas para cada gen mostró que el gen NA es el que

presentó mayor cantidad de variantes. Sin embargo, a priori, el simple conteo de variantes no indica nada sobre aspectos relacionados con la estructura poblacional de cada gen. Por ello se procedió al cálculo de las frecuencias de las variantes detectadas.

(A)



(B)

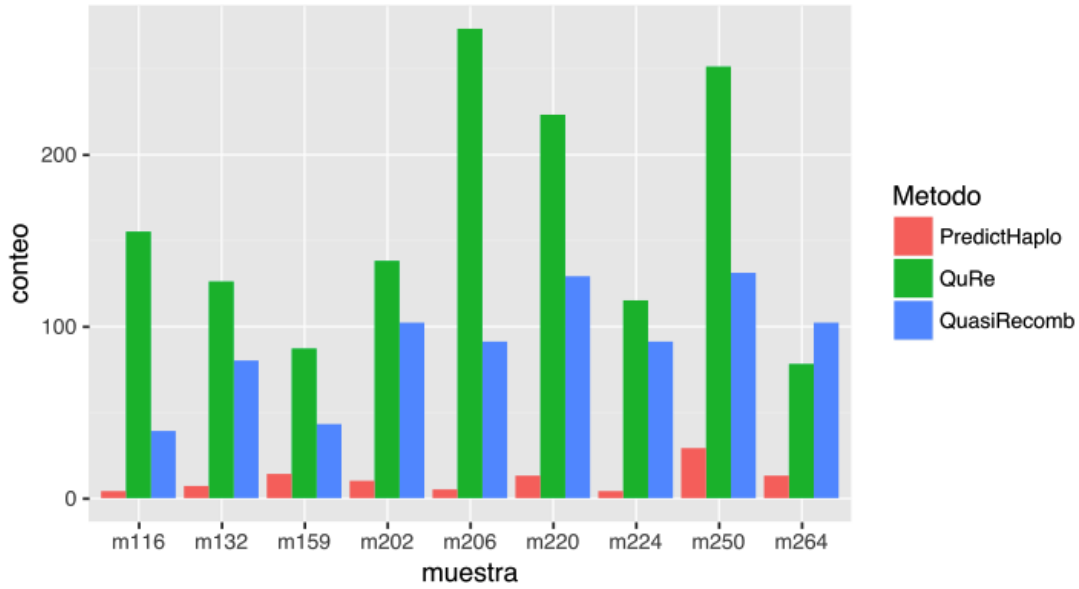


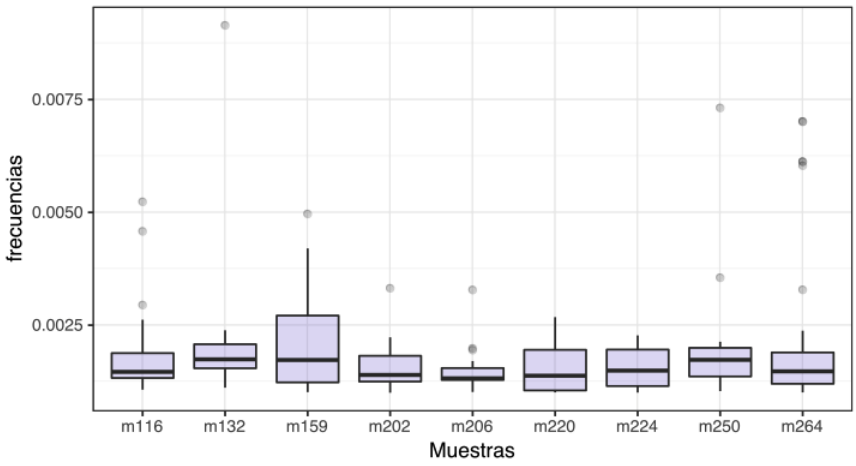
Figura 1. Resumen algoritmos de reconstrucción de cuasiespecies virales.
(A) y (B) representan el conteo de variantes reconstruídas para los genes HA y NA, respectivamente.

Estimación de frecuencias intra cuasiespecie:

Según la teoría, una cuasiespecie viral esta formada por miles de variantes genéticas relacionadas muy estrechamente, en donde cada mutante presenta una frecuencia determinada [Andino & Domingo, 2015; Eigen & Schuster, 1977; Eigen & Schuster, 1978a; Eigen & Schuster, 1978b]. Poder estimar la frecuencia de las variantes intra-cuasiespecie es fundamental para comprender la dinámica de estas poblaciones. En ese sentido, estimar la proporción de las variantes completas dentro de una población podría ayudar a entender cómo el virus evoluciona y se adapta a cambios durante el transcurso de una infección [Biebricher & Eigen, 2006]. Por ejemplo, podría ser muy útil evidenciar mutantes de escape a diferentes presiones selectivas como las impuestas por las drogas antivirales y el sistema inmune. Esto sería con el fin de mejorar las vacunas y terapias antivirales.

El resultado del análisis anterior mostró que los algoritmos de reconstrucción no logran reproducir el número de variantes que la teoría indica. Sin embargo, se cuenta con un número importante de variantes que deberían ser analizadas en mayor profundidad. Para ello, se estimó, mediante el algoritmo de reconstrucción QuRe, la frecuencia de cada variante. Se obtuvo para el gen HA frecuencias minoritarias principalmente en el rango de 0.0010046-0.0091451, mientras que para el gen NA éstas cayeron en su mayoría en el rango 0.001005-0.009873 (Figura 2). De igual forma, se encontraron, en menor proporción, variantes mayoritarias para HA en el rango de 0.011-1 y para NA en el rango 0.01063-0.9967 (Figura 3).

(A)



(B)

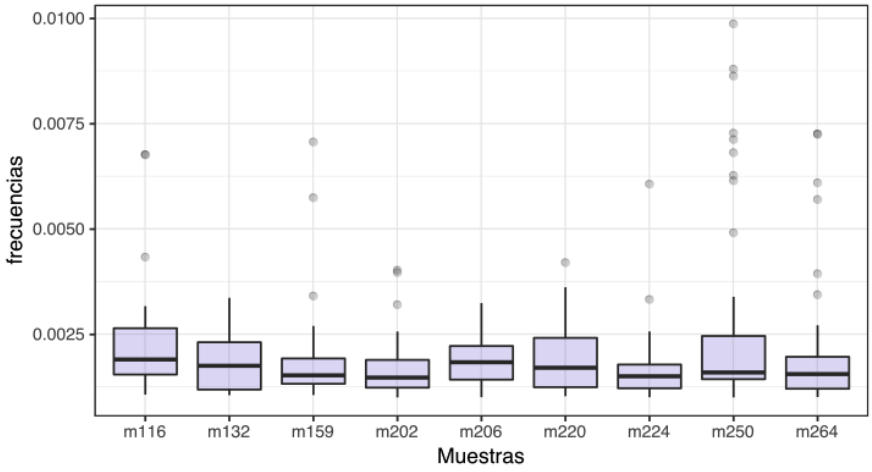
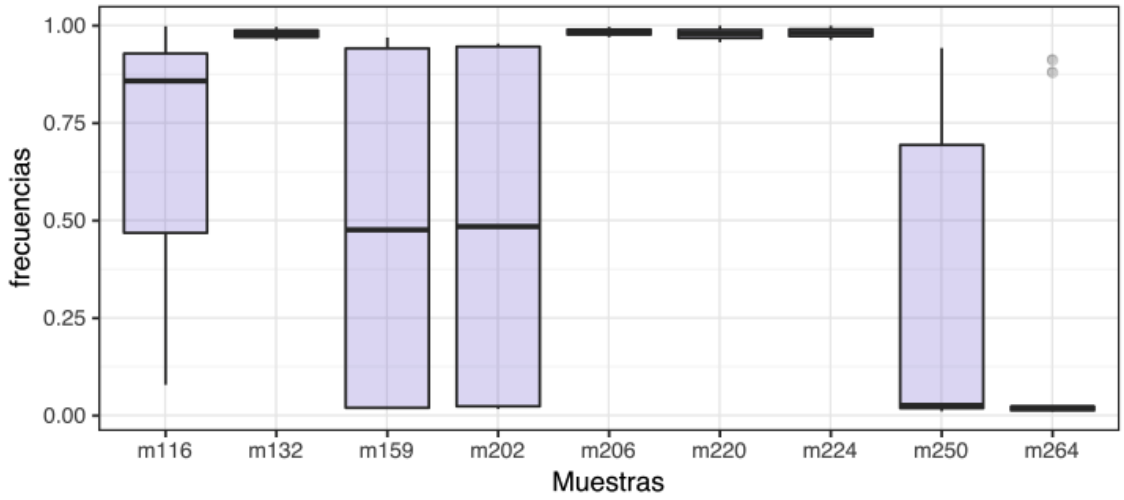


Figura 2. Estimación de frecuencias minoritarias.

(A) y (B) corresponden a la estimación de las proporciones de las variantes minoritarias en los genes HA y NA, respectivamente.

(A)



(B)

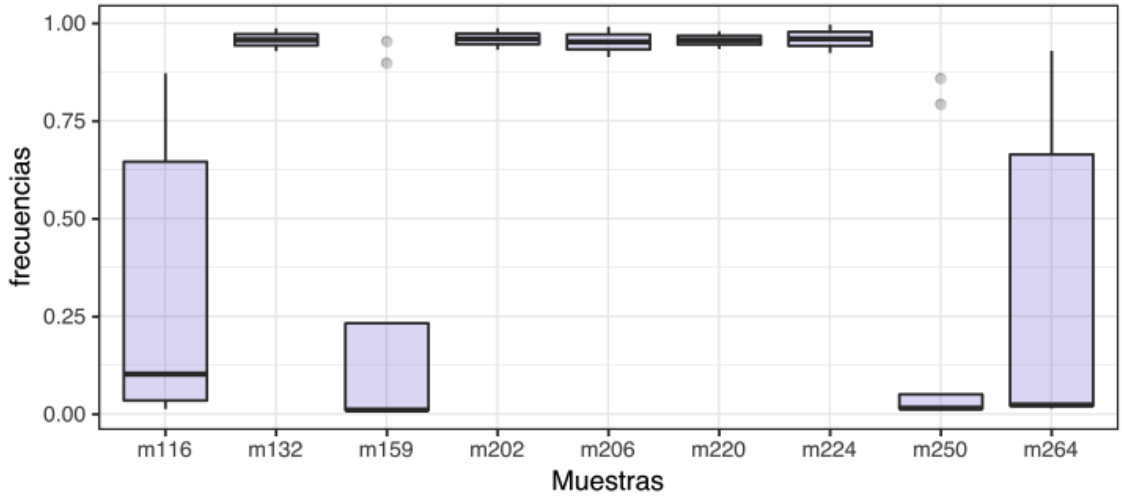
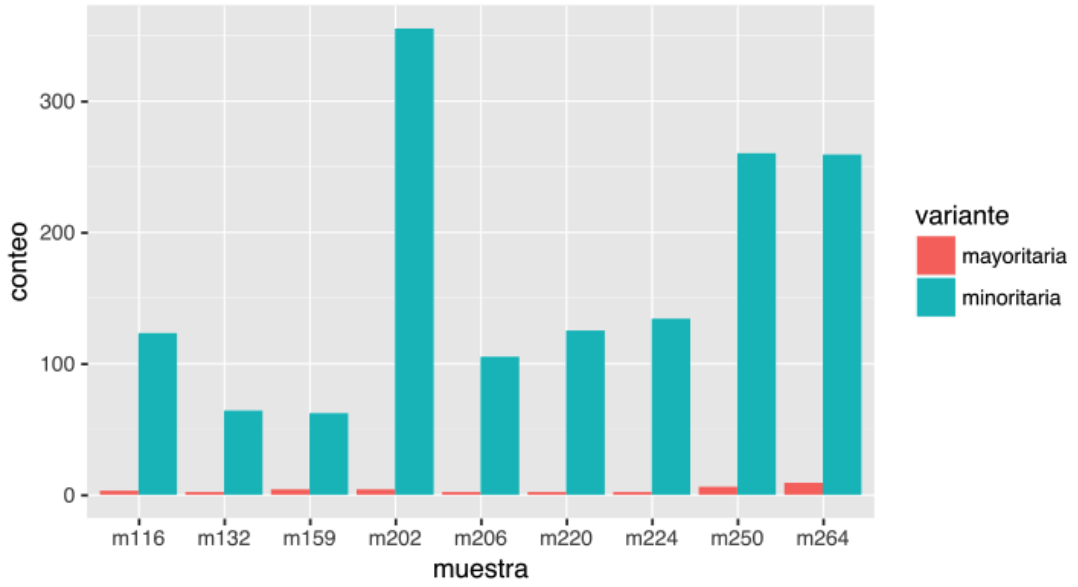


Figura 3. Estimación de frecuencias mayoritarias.
(A) y (B) corresponden a la estimación de las proporciones de las variantes mayoritarias en los genes HA y NA, respectivamente.

Asimismo, se estimó la cantidad de variantes minoritarias y mayoritarias para ambos genes. Este análisis mostró que el número de haplotipos minoritarios varía ampliamente entre las muestras y genes analizados. En el caso del gen HA se encontró cierta cantidad de variantes minoritarias en el rango de 62-355. La muestra que presentó menor cantidad de variantes minoritarias fue la 159, y la que mostró mayor cantidad fue la 202. Por otro lado, el conteo de haplotipos minoritarios para el

gen NA mostró variantes en el rango de 72-271. Se encontró que la muestra 264 presentó la menor cantidad de variantes minoritarias, mientras que la mayor cantidad se reportó en la muestra 206. Por otro lado, si se analiza desde la óptica de ambos genes y muestras procesadas, el análisis de las variantes mayoritarias resultaron en cantidades mayormente homogéneas. En general la cantidad de haplotipos mayoritarios para el gen HA cayó en el rango de 2-4 para la gran mayoría de las muestras, a excepción de las muestras 250 y 264 para las que se encontraron 6 y 9 haplotipos mayoritarios, respectivamente (Figura 4A). En esta misma linea, en gran parte de las muestras se vieron 2 haplotipos mayoritarios para NA, a excepción de las muestras 116, 159, 250 y 264, en que se hallaron 6, 8, 9 y 6 haplotipos mayoritarios, respectivamente (Figura 4B).

(A)



(B)

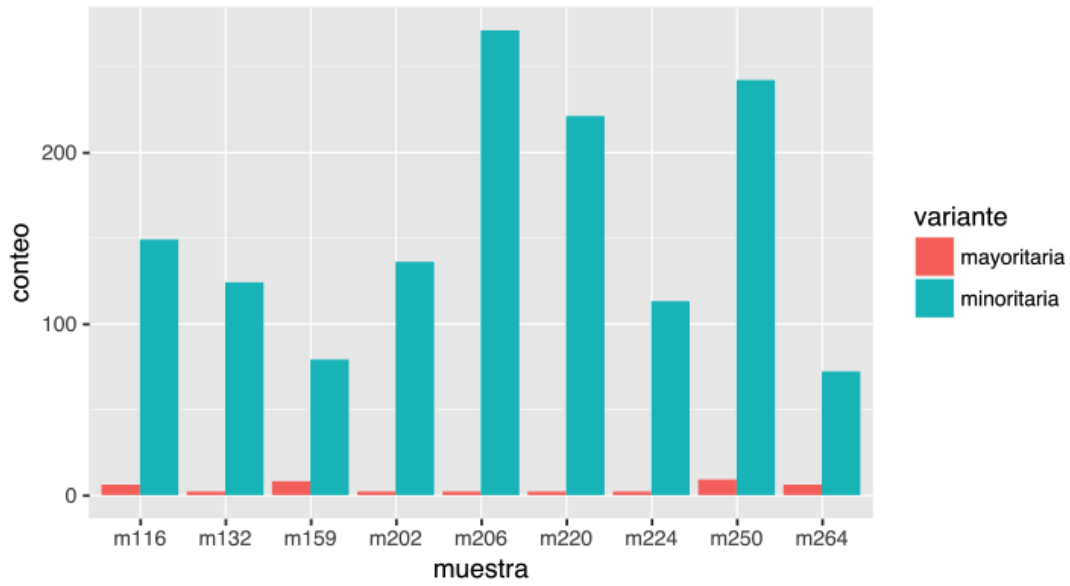


Figura 4. Conteo de tipos de variantes
(A) y (B) representan el conteo de variantes minoritarias y mayoritarias para los genes HA y NA, respectivamente.

Este análisis detectó la presencia de unas pocas variantes mayoritarias (~ de 2 a 9) rodeadas de varios haplotipos minoritarios. Estos resultados alinean, en parte, con la teoría de cuasiespecies que predice una o pocas variantes mayoritarias sumergidas en una constelación de variantes minoritarias [Domingo y cols., 2006; Vignuzzi y cols., 2006; Domingo & Gomez, 2007]. La estimación de frecuencias minoritarias mostró para ambos genes una distribución que varía entre muestras. Varias de las muestras alcanzaron valores entre 0.005 y 0.0075, incluso algunas lograron valores de 0.01. Sin embargo, se pudo ver que hay una clara concentración de haplotipos que caen entre el rango de frecuencias 0.0013-0.0020 para HA y 0.0012-0.0025 para el gen NA (Figura 2). Asimismo, también se vio que los valores de frecuencias mayoritarias varían ampliamente entre muestras. Los resultados para ambos genes evidenciaron que hay muestras (m132, m206, m220, m224) que presentan variantes mayoritarias principalmente con frecuencias cercanas a 1. Sin embargo, para el caso de la muestra 202, la distribución de frecuencias de haplotipos mayoritarios pareció ser más amplia para el gen HA, mientras que para NA las variantes presentaron frecuencias cercanas a 1. Lo opuesto ocurre con la muestra 264, aquí la distribución de variantes mayoritarias resulto más variable para el gen NA con dos variantes cercanas a

frecuencias de 0.8 y el resto cercanas a 0.012, mientras que para HA las frecuencias resultaron principalmente cercanas a 0.01. En esta misma linea, la muestra 250 mostró una distribución de frecuencias mayoritarias más amplia para el gen HA con frecuencias en el rango de 0.03-0.8, mientras que para el gen NA las variantes se restringieron a valores de frecuencias en un rango de 0.03-0.06 (Figura 3).
Este análisis mostró, por muestra y gen analizado, la presencia de determinado número de variantes mayoritarias rodeadas de un grupo bastante extenso de variantes minoritarias.

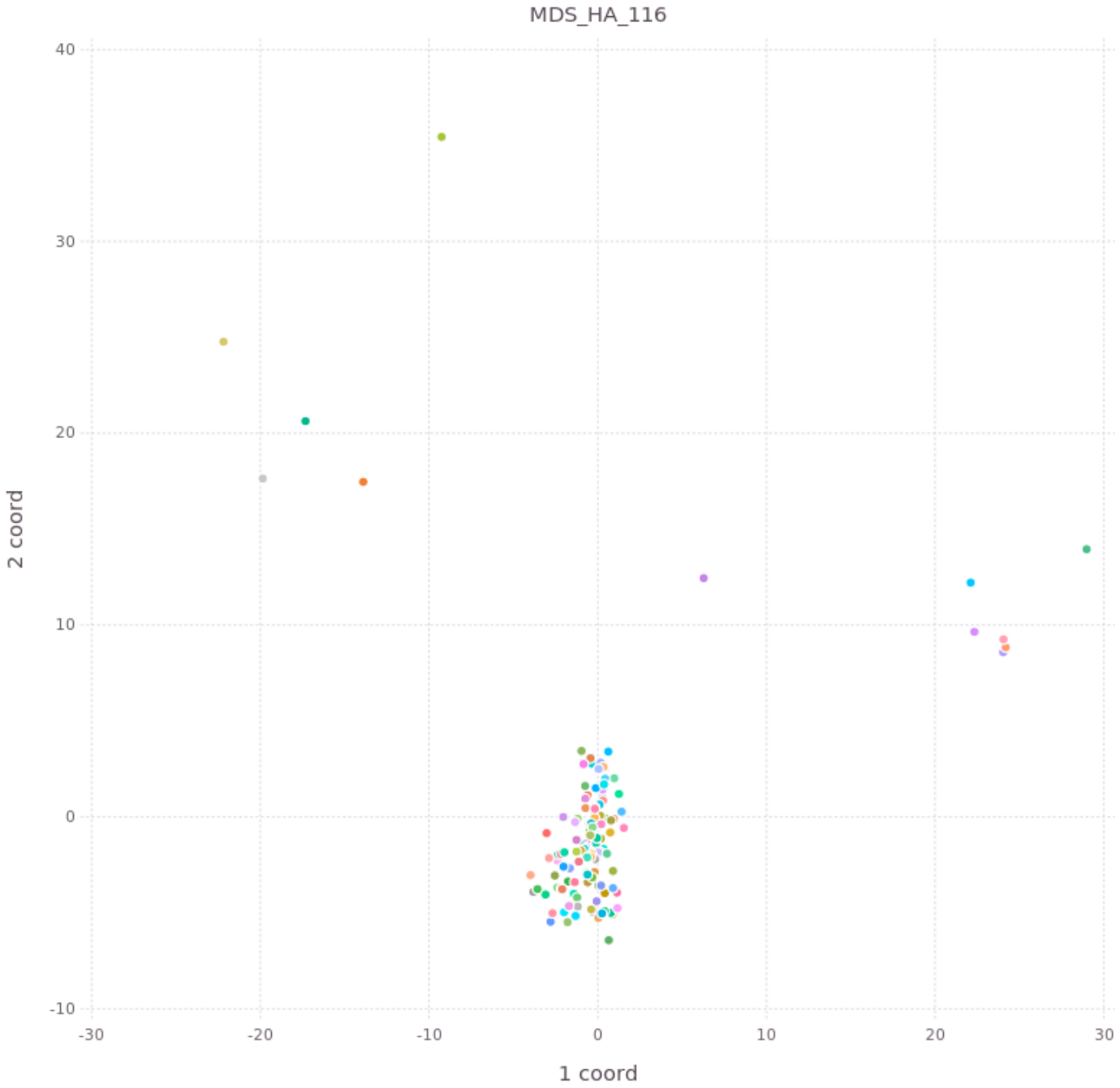
Relaciones entre haplotipos por muestra y entre muestras:

Según la teoría de las cuasiespecies, la cual se explaya en la parte introductoria de esta tesis, una cuasiespecie viral es, en resumen, una nube de mutantes constituída por miles de variantes genéticas estrechamente relacionadas que compiten y cooperan entre sí, en donde el objeto de la selección es la población como un todo [Andino & Domingo, 2015; Domingo E, 2000; Ruiz-Jarabo C M y cols., 2000]. Conocer las relaciones entre las variantes tanto a nivel intra-cuasiespecie como a nivel inter-cuasiespecie es fundamental para entender la dinámica poblacional y el modo de evolución de estas poblaciones virales [Domingo y cols., 2012]. Asimismo, conocer la distribución y probabilidad de mutantes es fundamental para evidenciar variantes de escape a diferentes presiones selectivas, como las drogas antivirales y las presiones impuestas por sistema inmune [Domingo y cols., 2012].

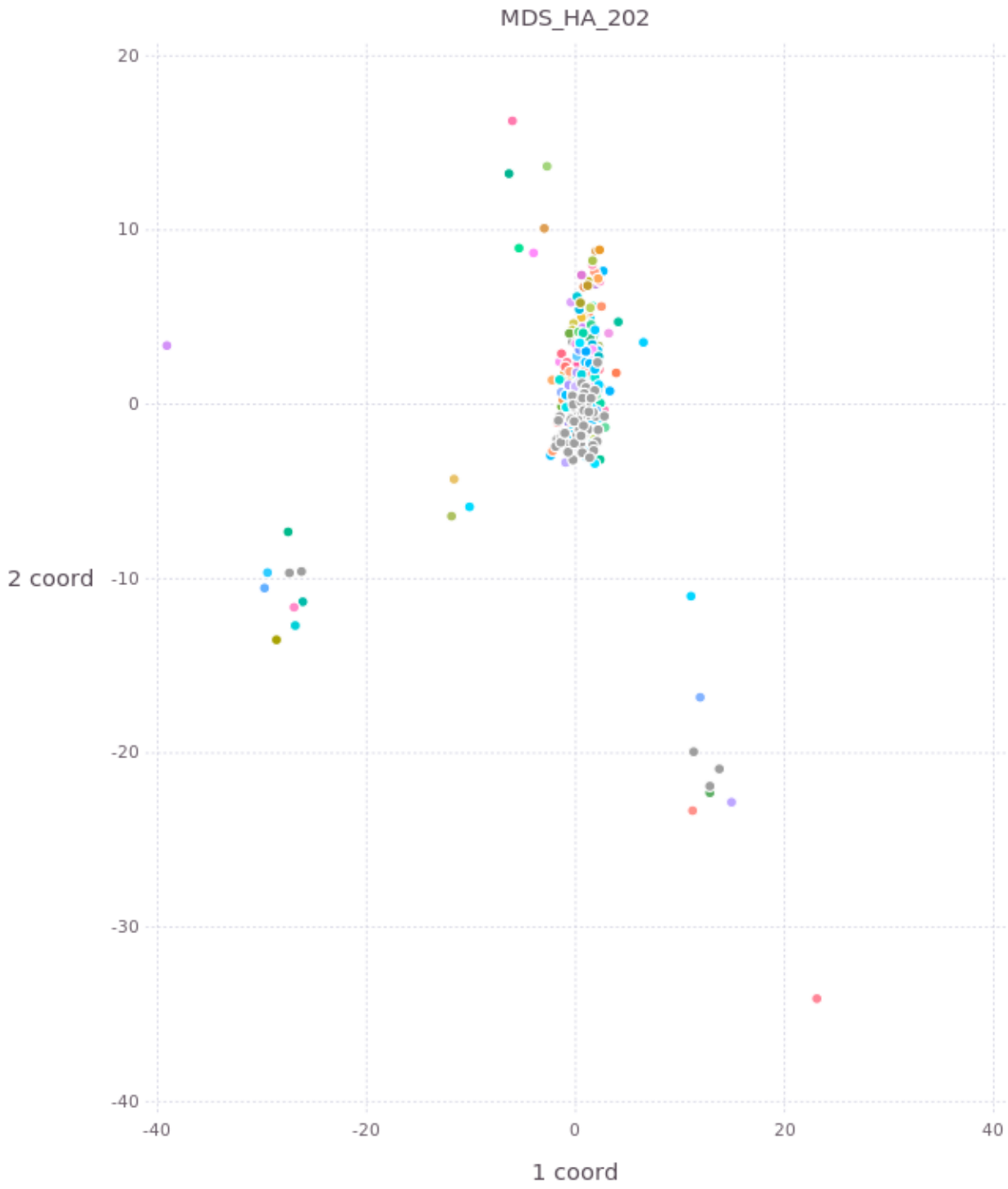
Análisis de Escalamiento Multidimensional:

Con el objetivo de evidenciar las relaciones intra-cuasiespecie se calculó una matriz de distancia para cada muestra y gen analizado. Estas fueron utilizadas para llevar a cabo análisis de escalado multidimensional (MDS). Los resultados para ambos genes muestran cierto conjunto de mutantes agrupados (aglomerados) en determinada región del MDS. Estos aglomerados de mutantes integran decenas y hasta centenas de variantes. Asimismo, se pueden ver, para todos los casos, outliers que representan variantes con una distancia genética mayor en relación a los aglomerados descriptos (Figura 5).

(A)



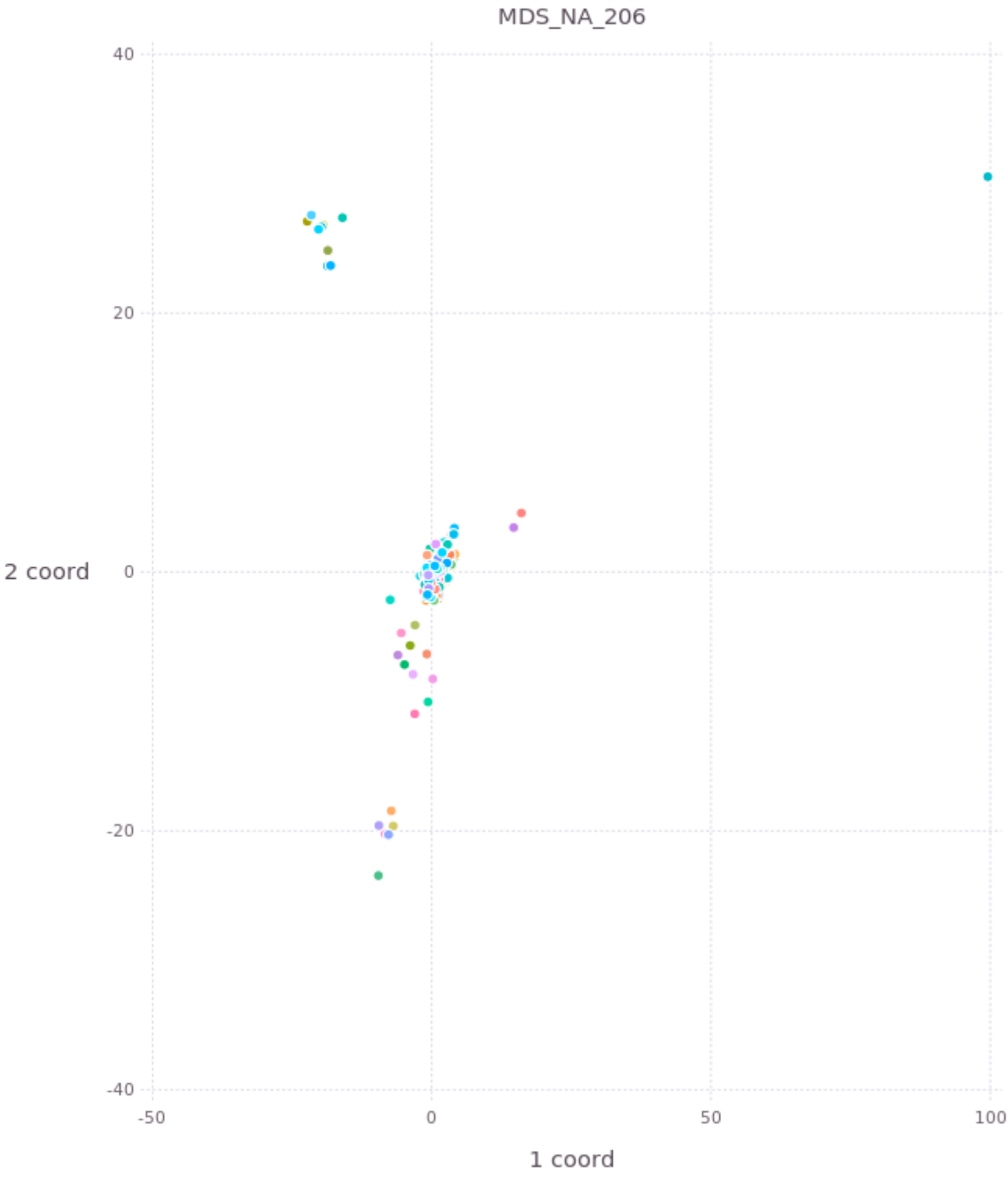
(B)



(C)

(D)

(E)



(F)

Figura 5. Análisis de escalamiento multidimensional.

Se observa una muestra representativa de cada año.

(A), (B) y (C) son los MDS para el gen HA de las muestras m116, m202 y m250, respectivamente.

(D), (E) y (F) son los MDS para el gen NA de las muestras m116, m206 y m250, respectivamente.

Se estima que el cúmulo principal de mutantes esta conformado por variantes muy relacionadas, en donde cada punto representa una distancia de pocos cambios nucleotídicos. Por otro lado, los grupos de puntos outliers involucrarían variantes genéticas con cambios mayores a nivel de secuencia, como indels o varias mutaciones puntuales. Los resultados de este análisis muestran una distribución de variantes mayoritarias heterogénea rodeada de una nube de variantes minoritarias. Sin embargo, como perspectiva a este punto se plantea dilucidar qué secuencia conforma cada grupo de haplotipos y evidenciar que los outliers corresponde a secuencias con indels.

Análisis de correlación mediante heatmaps:

Con el objetivo de profundizar el estudio de las relaciones intra-cuasiespecie y evidenciar las relaciones inter-cuasiespecie se construyó un alineamiento para cada gen. Dicho conjunto de datos incluyó una cantidad representativa de cada muestra original. Con cada alineamiento se calculó una matriz de distancia para llevar a cabo un análisis de correlación con mapas de calor. El conjunto de datos se analizó con la librería 'seqinr' de CRAN.R-project.org. Se utilizó la función 'read.alinment' para ingresar a los alineamientos desde la consola de RStudio y posteriormente se calculó la matriz de distancias con la función 'dist.alignment' la cual se procesó con la función 'heatmap', todos disponibles en el repositorio CRAN.R-project.org.

Se obtuvieron los heatmaps para las variantes de los genes HA y NA reconstruídas por el programa QuRe.

Este estudio mostró una fuerte relación entre las variantes de cada muestra. Esto se ve reflejado por la formación de 9 clusters principales (Figura 5 y Figura 6). De igual forma se puede ver relación entre muestras de un mismo año. Asimismo, se pudo ver cierta relación entre las variantes de distintos años, lo que se nota por la presencia de tonos rojos entre distintos clusters. Sin embargo, este patrón no se observa para todas las muestras. Se puede ver que la distancia genética, entre las variantes de HA ensambladas para las muestras del año 2013 (esquina superior izquierda del heatmap) y el resto de las variantes es mayor, lo que queda reflejado por los tonos claros del heatmap. La poca distancia genética entre las variantes HA del 2013 (dado por la intensidad de colores del heatmap) habla de una cuasiespecie más compacta, en términos de variantes. Este fenómeno puede ser atribuido al efecto cuello de botella gracias a la mejora en el componente vacunal. Como perspectiva de este análisis se plantea dilucidar qué secuencia conforma cada cluster de haplotipos y evidenciar con

qué otras secuencias se relacionan.

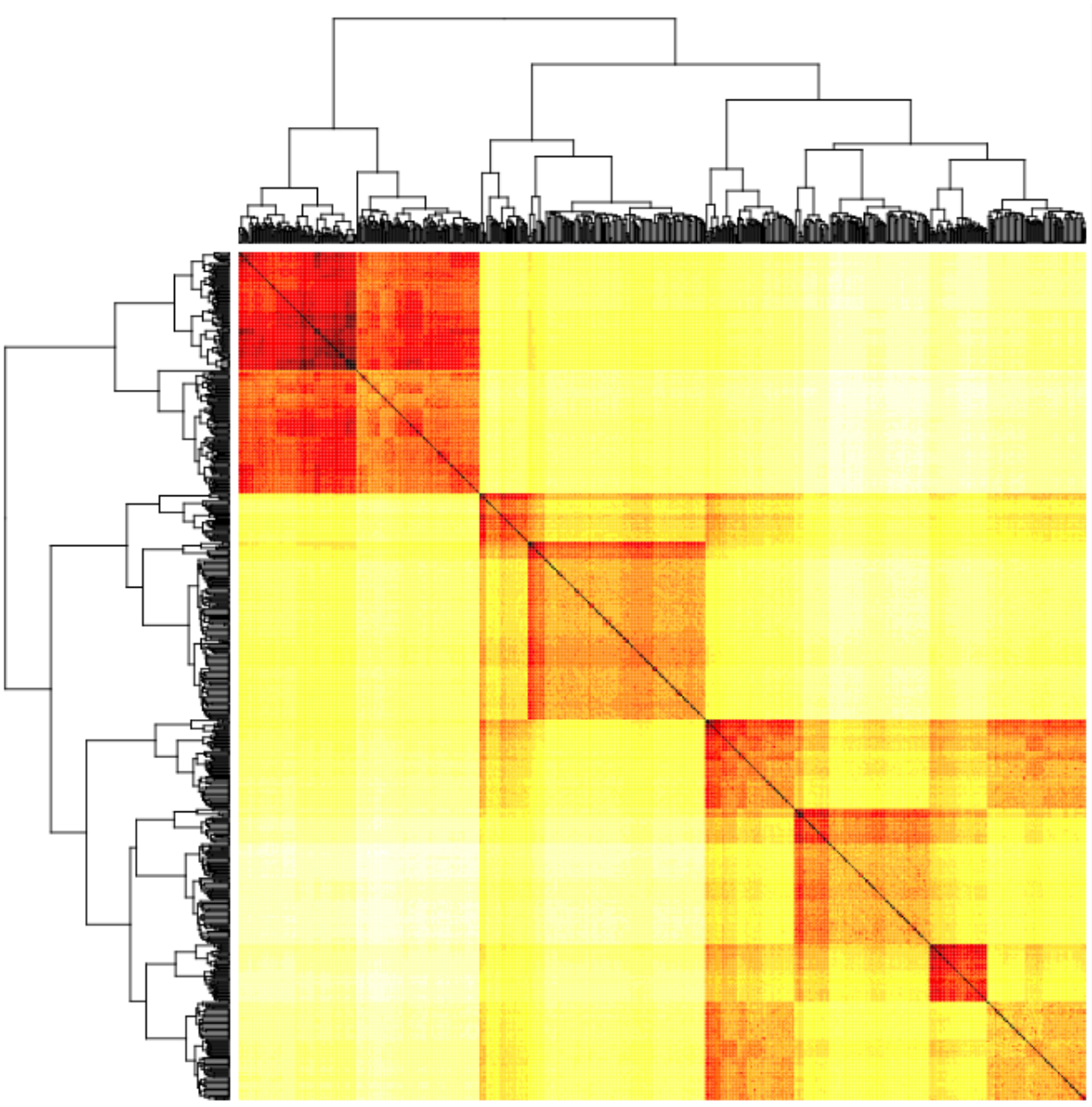


Figura 5. Correlación inter-cuasiespecie del gen HA.

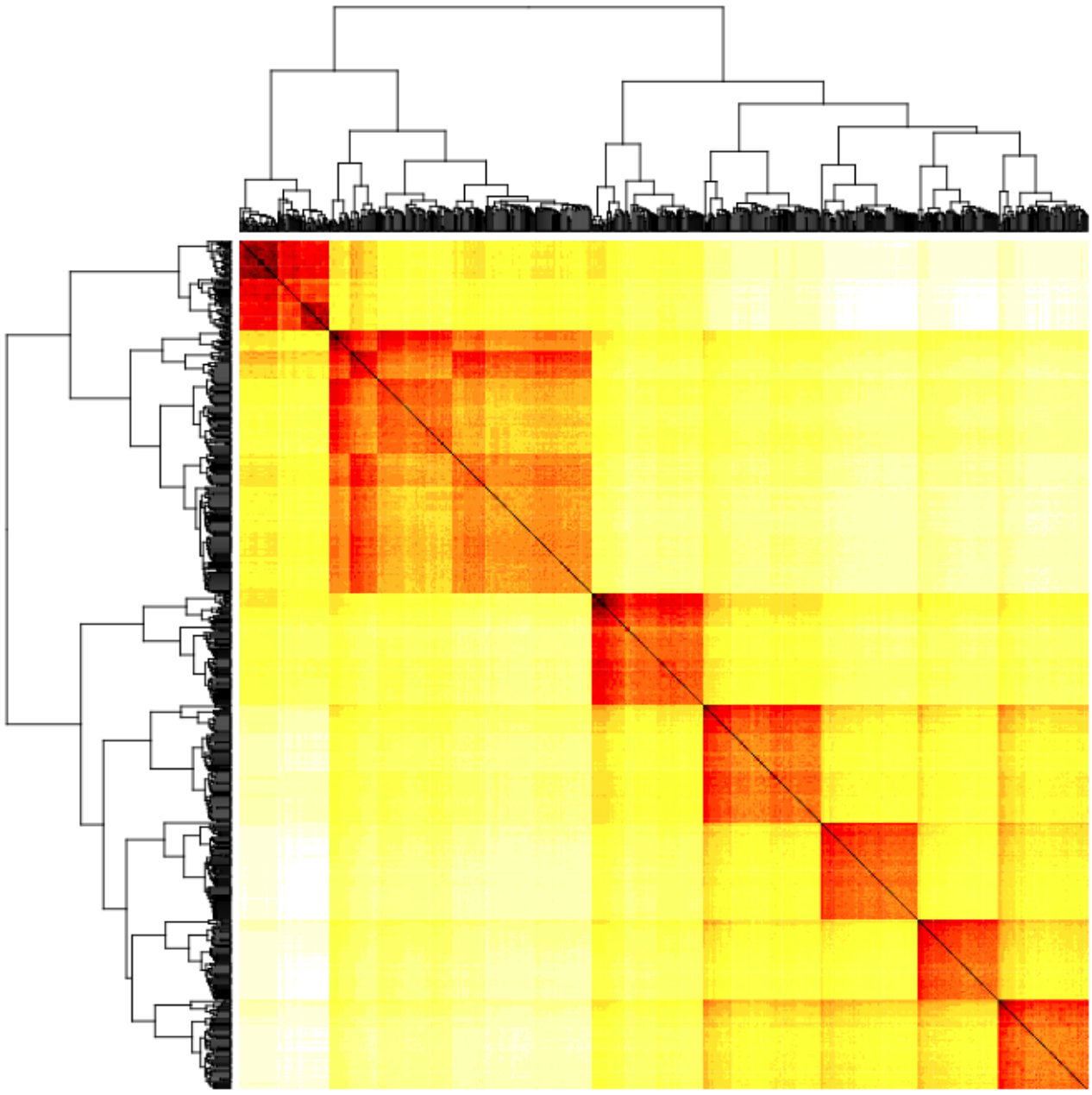


Figura 6. Correlación inter-cuasiespecie del gen NA.

CONCLUSIONES

Respecto a los análisis llevados a cabo en este estudio, se logró la implementación de tres algoritmos de reconstrucción de cuasiespecies virales. En relación a estas pruebas, a pesar de usar el mismo conjunto de datos inicial, se vio que la cantidad de haplotipos generados varía notoriamente entre los distintos algoritmos utilizados. El programa con mayor número de haplotipos por muestra fue QuRe. Se estimó la frecuencia de cada variante y se encontró en todas las muestras, para ambos genes, la presencia de variantes mayoritarias y minoritarias. Según los análisis de escalamiento multidimensional se encontraron, para cada muestra y gen analizado, ciertas acumulaciones de variantes destacando que en algunas el conglomerado es mayor. Esto deja ver la presencia de posibles subpoblaciones virales dentro de una cuasiespecie. Como perspectiva en este estudio se buscará identificar qué secuencias conforman cada grupo de haplotipos y dilucidar la información genética de los outliers. Finalmente, el análisis por mapas de calor evidenció relaciones intra-cuasiespecie. Asimismo, se encontró cierta relación entre las variantes de algunas muestras, lo que apoya la idea de relaciones inter-cuasiespecie. Como perspectiva se plantea conocer las secuencias de las variantes relacionadas y poder estimar la probabilidad de que las variantes circulantes en un año puedan emerger en años siguientes.

Entender cómo los virus se relacionan y evolucionan es de gran utilidad para poder afrontar futuras epidemias con mayor precisión. El conocimiento en profundidad de los componentes de una población viral intra paciente es fundamental para evidenciar posibles cepas resistentes a antivirales. En este sentido, conocer las variantes intra-cuasiespecie y sus relaciones, nos permiten brindar mayores herramientas a la población para poder dilucidar los problemas específicos de cada paciente y poder diseñar terapias antivirales paciente específicas, más conocidas como medicina personalizada.

Capítulo 3

Patrones evolutivos del virus de la enfermedad de Newcastle en la Antártida

RESUMEN

El virus de la enfermedad de Newcastle (NDV) plantea una grave amenaza para la industria avícola en todo el mundo. Recientemente, NDV ha sido aislado en la región Antártica. Estudios detallados sobre el modo de evolución de cepas de NDV aisladas en todo el mundo son relevantes para la comprensión de su historia evolutiva. Por esta razón hemos realizado análisis bayesianos de coalescencia con cepas de NDV aisladas en la Antártida y otras regiones geográficas, con el fin de investigar sus tasas evolutivas, dinámica poblacional y patrones de evolución. Los resultados son discutidos en términos del posible papel de la Antártida en el mantenimiento de poblaciones virales de NDV potencialmente emergentes o re-emergentes en todo el mundo. En base a análisis de secuencias del sitio de clivaje de su proteína F, se ha determinado que los aislados de NDV de la Antártida corresponden a cepas lentogénicas. Estas estirpes comparten con la cepa referencia del genotipo I, Ulster/67, un ancestro que circuló alrededor del año 1958. Asimismo, el tiempo del ancestro común más reciente (MRCA) para todos los virus de Clase II se estableció en el entorno del año 1883. Se estimó una tasa evolutiva de 1,78 10^{-3} substituciones por sitio por año (s/s/y) a partir del análisis de secuencias del gen F de diferentes cepas de NDV. El gráfico de Skyline Bayesiano indicó una disminución del tamaño poblacional de NDV en los últimos 25 años.

INTRODUCCIÓN

El virus de la enfermedad de Newcastle (NDV) es el agente causal de una de las enfermedades de mayor incidencia en la industria avícola en todo el mundo [Samal y cols., 2011] . NDV pertenece al género Avulavirus de la familia Paramyxoviridae. Su genoma está constituido por una molécula de ARN de sentido negativo de aproximadamente 15.186 nucleótidos (nt) de longitud [David & Howley, 2007] . Las cepas de NDV se agrupan según su fenotipo en lentogénicas, mesogénicas y velogénicas, en orden creciente de virulencia [Kim y cols., 2007] . Las cepas lentogénicas causan típicamente infecciones subclínicas o enfermedades respiratorias leves. Por su parte, las variantes mesogénicas suelen ser de virulencia intermedia, dando lugar generalmente a enfermedades respiratorias moderadas, con ocasionales afecciones neurológicas. Por último, los virus velogénicos son los más virulentos y pueden generar severas hemorragias, particularmente en el tracto gastrointestinal (variantes viscerotrópicas), y/o afecciones al sistema nervioso (cepas neurotrópicas) [Alexander, 2000].

La infección por NDV se inicia por la acción de dos glicoproteínas de la envoltura. Una de ellas media en la unión del virus a un receptor de la célula hospedera y se designa HN (Hemaglutinina-Neuraminidasa). La otra glicoproteína, designada como la proteína de fusión (F), es responsable de la penetración del virus en la célula hospedera y de la formación de sincicios [Toyoda y cols., 1988] . La proteína F desempeña además un papel clave en la virulencia y es un importante blanco para la respuesta inmune [Neyt y cols., 1989] . Se trata de una proteína integral de membrana trimérica del tipo I, que se sintetiza como un precursor inactivo, F0 (de 66 kDa), que es post-traduccionalmente procesada por proteasas celulares en dos subunidades unidas por puentes disulfuro, la F2 N-terminal (12,5 kDa) y la F1 C-terminal (55 kDa) [Nagai y cols., 1989; Samal y cols., 2012]. La secuencia del sitio de clivaje de la proteína F es un importante determinante de la patogenicidad de NDV. En particular, los sitios de clivaje de las cepas virulentas suelen contener múltiples residuos básicos, a diferencia de las variantes no virulentas que presentan menos residuos básicos [Panda y cols., 2004].

La secuencia consenso de este sitio en cepas velogénicas y mesogénicas es 112(R/K)RQ(R/K)RF117, mientras que la misma en variantes lentogénicas es 112(G/E)

(K/R)Q(G/E)RL117 [de Leeuw y cols., 2005]. Más específicamente, el motivo que presentan la mayoría de las cepas virulentas más recientemente reportadas es 112RRQKRF117 [Choi y cols., 2010; Pedersen y cols., 2004]. Siete epítopos neutralizantes han sido mapeados en la proteína F de NDV [Maminiaina y cols., 2010; Toyoda y cols., 1988; Yusoff y cols., 1989], con particular implicancia en sitios de neutralización de los aminoácidos 72, 74, 75, 78, 79, así como el tramo entre los residuos 157 y 171 [Maminiaina y cols., 2010; Mase y cols., 2011].

En base a análisis genéticos, las cepas de NDV se clasifican en dos clases: Clase I, formada por un único genotipo que circula en aves silvestres y son mayormente no virulentas [Czeglédi y cols., 2006]; y Clase II, compuesta por 18 o 19 genotipos (I-XIX), que son tanto virulentas como no virulentas y circulan en aves salvajes y domésticas [Diel y cols., 2012a; Fernandes y cols., 2014; Snoeck y cols., 2013]. Las cepas de los genotipos V, VI y VII de Clase II circulan actualmente en pollos alrededor del mundo [Xiao y cols., 2013].

La vacunación ha sido ampliamente utilizada para el control de NDV, que se reportó por primera vez en aves de corral en 1926 [Chong y cols., 2010]. Las vacunas vivas más comúnmente utilizadas son LaSota y Clone-30, que pertenecen al genotipo II [Rui y cols., 2010]. La caracterización genética de cepas de NDV es sumamente importante para evaluar sus modificaciones, anticipar nuevos brotes y desarrollar medidas de control adecuadas [Zhang y cols., 2011]. Sin embargo, aún falta mucha investigación acerca de la epidemiología y evolución de este virus, lo que limita el control de esta enfermedad [Afonso & Miller, 2013; Susta y cols., 2011].

Se han reportado tres grandes panzootias en el último siglo. La primera de ellas (1926-1960) fue causada por variantes de los genotipos II, III y IV, mientras que la segunda (1960-1973) y la tercera (1970-1980) fueron causadas por variantes de los genotipos V y VI [Maminiaina y cols., 2010]. En la década de 1990, se reportaron severos brotes en Europa Occidental y Meridional [Herczeg y cols., 1999; Lomniczi y cols., 1998], Sudáfrica [Abolnik y cols., 2004] y Taiwán [Yang y cols., 1999], causados por el genotipo VII, que actualmente circula en Asia, África y Europa [Maminiaina y cols., 2010]. Asimismo, una epidemia reciente en América del Sur (Venezuela), también ha sido atribuida a cepas de este genotipo, lo que sugiere su amplia diseminación geográfica [Diel y cols., 2012b; Perozo y cols., 2012].

En 2010 se confirmó la infección por variantes de NDV virulentas en 80 países, incluyendo aves silvestres en Canadá, Alemania, Israel, Italia, Kenia, Mongolia y los

EE.UU., e infecciones en aves domésticas de países de América del Norte y Sur, Europa, África y Asia [OIE, 2011]. Por otra parte, estudios recientes revelaron la circulación de NDV en pingüinos de la Isla King George en la región Antártica [Thomazelli y cols., 2010]. Análisis detallados sobre el modo de evolución de estas nuevas cepas son relevantes para inferir la historia evolutiva de NDV. Con el fin de profundizar en estas cuestiones, en este estudio se realizaron estudios bayesianos de coalescencia para investigar las tasas de evolución, la dinámica poblacional y los patrones evolutivos de NDV.

OBJETIVOS

Objetivos específicos

Evaluar las características de virulencia de las cepas de NDV.

Investigar los patrones evolutivos de las poblaciones de NDV.

Secuencias

Las secuencias nucleotídicas de las diferentes cepas de NDV utilizadas en este estudio fueron descargadas de la base de datos DDBJ (www.ddbj.nig.ac.jp). Los nombres de las cepas y sus números de acceso se resumen en la Tabla 1. Se utilizó el programa MUSCLE [Edgar, 2004] implementado en el programa MEGA 5 [Tamura y cols., 2011] para realizar el alineamiento de las secuencias nucleotídicas y su posterior traducción *in silico* a aminoácidos.

Tabla 1: Información de las secuencias nucleotídicas utilizadas en este estudio.

Número de Acceso	Genotipo*	Fecha	Nombre de la cepa
[DDBJ:AB524405]	Clase I	1991	Goose/Alaska/415/91
[DDBJ:AB524406]	Clase I	2009	9a5b-D5C1
[DDBJ:AY562991]	I	1967	chicken/N. Ireland/Ulster/67
[DDBJ:HM125898]	I	2004	WDK/JX/7793/2004
[DDBJ:JX401405]	I	2007	CBU2374
[DDBJ:JN653339]	I	2007	M4
[DDBJ:JX401404]	I	2007	CBU2179
[DDBJ:HM063424]	I	2005	*Rallus aquaticus*/China/R8/2005
[DDBJ:JX401403]	I	2007	CBU2249
[DDBJ:HM143848]	I	2006	02/Antarctic/Brazil/2006
[DDBJ:HM143849]	I	2006	39/Antarctic/Brazil/2006
[DDBJ:JX193083]	I	2010	duck/China/Guangxi22/2010
[DDBJ:AF077761]	II	1946	LaSota
[DDBJ:GQ994433]	II	2008	XD/Shandong/2008
[DDBJ:HM357251]	II	1987	NDV-4/chicken/Namakkal/Tamil Nadu/India
[DDBJ:AY845400]	II	2010	China/LaSota
[DDBJ:JX193082]	II	2010	duck/China/Guangxi21/2010
[DDBJ:GQ288378]	IIa	1987	northern pintail/US(OH)/87-486/1987
[DDBJ:GQ288391]	IIa	2001	mottled duck/US(TX)/01-130/2001
[DDBJ:GQ288380]	IIa	1986	mallard/US(OH)/86-233/1986
[DDBJ:GQ288377]	IIa	2004	mallard/US(OH)/04-411/2004
[DDBJ:GQ288379]	IIa	2003	mallard/US(MD)/03-632/2003
[DDBJ:GQ288389]	IIa	1999	mallard/US(MN)/99-376/1999
[DDBJ:GQ288392]	IIa	2000	mallard/US(MN)/MN00-39/2000
[DDBJ:FJ430159]	III	2005	JS/07/05/Ch
[DDBJ:FJ430160]	III	2003	JS/9/05/Go
[DDBJ:JF950509]	III	2010	Mukteswar
[DDBJ:EU293914]	IV	2008	Italien
[DDBJ:AY562886]	V	1993	anhinga/U.S.(Fl)/44083/1993
[DDBJ:GQ288388]	V	1992	cormorant/US(CA)/92-23071/1997
[DDBJ:GQ288387]	V	1992	cormorant/US(MN)/92-40140/1992
[DDBJ:FJ766531]	VI	2007	JS/07/03/Pi
[DDBJ:AY562988]	VI	1972	chicken/U.S. (CA)/1083(Fontana)/72
[DDBJ:FJ766529]	VI	1997	ZhJ-3/97
[DDBJ:GU564399]	VII	2006	FMW
[DDBJ:DQ485231]	VII	2003	chicken/China/Guangxi11/2003
[DDBJ:JN653340]	VII	2003	PX2/03
[DDBJ:JN400896]	VII	2011	Chicken/China/SDSG01/2011
[DDBJ:AF431744]	VII	2007	ZJ1
[DDBJ:JX867334]	VII	2008	YZCQ/Liaoning/08
[DDBJ:FJ872531]	VII	2002	duck/China(Fujian)/FP1/02
[DDBJ:JN599167]	VII	1999	BP01
[DDBJ:FJ754273]	VII	2000	WF00G
[DDBJ:FJ754272]	VII	2000	WF00D

[DDBJ:DQ485229]	VII	2002	chicken/China/Guangxi7/2002
[DDBJ:FJ751918]	VIII	1979	QH1
[DDBJ:FJ751919]	VIII	1985	QH4
[DDBJ:FJ436304]	IX	1985	FJ/1/85/Ch
[DDBJ:FJ436306]	IX	2002	JS/1/02/Du
[DDBJ:FJ436305]	IX	1997	JS/1/97/Ch
[DDBJ:HQ266604]	XI	2008	MG_MEOLA_08
[DDBJ:HQ266602]	XI	2008	MG_725_08
[DDBJ:HQ266603]	XI	1992	MG_1992
[DDBJ:KC152049]	XII	2010	GD1003/2010
[DDBJ:KC152048]	XII	2011	GD450/2011
[DDBJ:KM056348]	XIII	2013	ndv40/sarsa/04/2013
[DDBJ:KM056347]	XIII	2013	ndv32/vaherakhadi/04/2013
[DDBJ:KM056349]	XIII	2013	ndv42/gopalpura/04/2013
[DDBJ:KC568209]	XIV	2009	NG-720/KD.TW.03T
[DDBJ:JN872165]	XIV	2006	Chicken/Niger/VIR 1377-7/2006
[DDBJ:HF969180]	XVII	2009	avian/Cameroon/CAE08-318/2009
[DDBJ:KF442614]	XVII	2006	Nigeria/228-7/2006
[DDBJ:HF969185]	XVII	2007	chicken/Ivory Coast/CIV08-104/2007
[DDBJ:FJ772478]	XVII	2008	chicken-3490-149-Cameroon-2008
[DDBJ:HJ969209]	XVII	2011	chicken/Nigeria/NIE10-123/2011
[DDBJ:FJ772455]	XVIII	2006	avian-1532-14-Mauritania-2006
[DDBJ:HF969216]	XVIII	2011	chicken/Nigeria/NIE11-1286/2011
[DDBJ:JN942101]	XVIII	2008	Finch/Eastern Hemisphere/1409-12/2008
[DDBJ:HF969218]	XVIII	2007	chicken/Ivory Coast/CIV08-042/2007
[DDBJ:JX546248]	XVIII	2009	NDV/chicken/Togo/AKO18/2009
[DDBJ:KF792021]	XIX	2013	Chicken/BT-Israel/2013/120
[DDBJ:KF792020]	XIX	2012	parrot/Israel/2012/841
[DDBJ:KF792019]	XIX	2013	chicken/KY-Israel/2013/50
[DDBJ:KF792018]	XIX	2011	Chicken/Israel/2011/1115

*Las cepas de Clase II se indican de acuerdo a su genotipo.

Análisis de coalescencia

Con el objetivo de investigar la tasa evolutiva y el modo de evolución de NDV, se utilizó una aproximación bayesiana utilizando Cadenas de Markov mediante la simulación Monte Carlo (MCMC), implementada en el paquete BEAST v.1.7.5 [Drummond & Rambaut, 2007]. Las cepas incluidas en estos análisis se detallan en la Tabla 1. En primer lugar, se identificó el modelo evolutivo óptimo para el conjunto de datos a analizar mediante el servidor Datamonkey [Delport y cols., 2010], utilizando los criterios de información de Akaike y la prueba de la razón de verosimilitud jerárquica (AIC y HLRT). Utilizando este modelo evolutivo (HKY + Γ) y 50 millones de generaciones de MCMC, se testaron diferentes modelos poblacionales: tamaño poblacional constante; crecimientos poblacionales exponencial, expansional y logístico; y *Bayesian Skyline*. La incertidumbre estadística de los datos se refleja por la densidad de probabilidad mayor a 95% y la convergencia fue evaluada por valores de ESS (tamaño efectivo de muestreo) superiores a 200. Los resultados fueron examinados utilizando el programa TRACER v1.5 (disponible en

http://beast.bio.ed.ac.uk/Tracer). Este mismo programa fue utilizado para comparar los resultados obtenidos mediante diferentes modelos poblacionales, mediante el cálculo del factor de Bayes [Drummond y cols., 2006] (http://beast.bio.ed.ac.uk/Modelcomparison) Posteriormente se generó el árbol de máxima credibilidad de clados utilizando el programa TreeAnnotator v1.7.5 (del paquete BEAST) y se editó el mismo mediante el programa FigTree v1.4.1 (disponible en http://tree.bio.ed.ac.uk). Asimismo, se realizó un gráfico de *Skyline* Bayesiano (*Bayesian Skyline Plot*, BSP) con el objetivo de explorar los cambios en el tamaño poblacional efectivo a lo largo del tiempo [Drummond y cols., 2006; Suchard y cols., 2001].

Mapeo de las sustituciones aminoacídicas en la proteína F de cepas de NDV aisladas en la Antártida.

Estudios previos han identificado que las cepas de NDV aisladas en la Antártida pertenecen a la Clase II [Thomazelli y cols., 2010]. Con el objetivo de evaluar las características de virulencia de estas cepas, se alinearon secuencias parciales del gen F de cepas de NDV aisladas en la Antártida con secuencias comparables de diferentes miembros de distintos genotipos de cepas de Clase II, cuyo genoma completo está reportado. La región analizada corresponde a las posiciones 4502-4995, en relación a la cepa referencia de NDV, *LaSota* (número de acceso: AF077761). Los nombres, números de acceso y fechas de aislamiento de las cepas incluidas en este análisis se indican en la Tabla 1. Una vez alineadas, se realizó una traducción *in silico* a aminoácidos, mediante el programa MEGA5. Los resultados de las diferencias aminoacídicas encontradas se muestran en la Figura 1.

La secuencia del sitio de clivaje de la proteína F de las cepas aisladas en la Antártida es ^{112}GKQGRLI^{118}, lo que sugiere que se trata de cepas lentogénicas. Asimismo, no se observaron sustituciones de aminoácidos en los sitios 72, 74, 75, 78 y 79 de la proteína F2, que fueron previamente reportados de estar involucrados en la neutralización [Mase y cols., 2011]. También se observó la conservación de un potencial sitio aceptor de N-glicosilación (N-X-S/T, donde X corresponde a cualquier aminoácido excepto prolina o ácido aspártico), presente en las posiciones 85 a 87 de la proteína F2 [Paldurai y cols., 2010; Panda y cols., 2004], así como de los residuos de cisteína en las posiciones 25 y 76 [Seal, 2004].

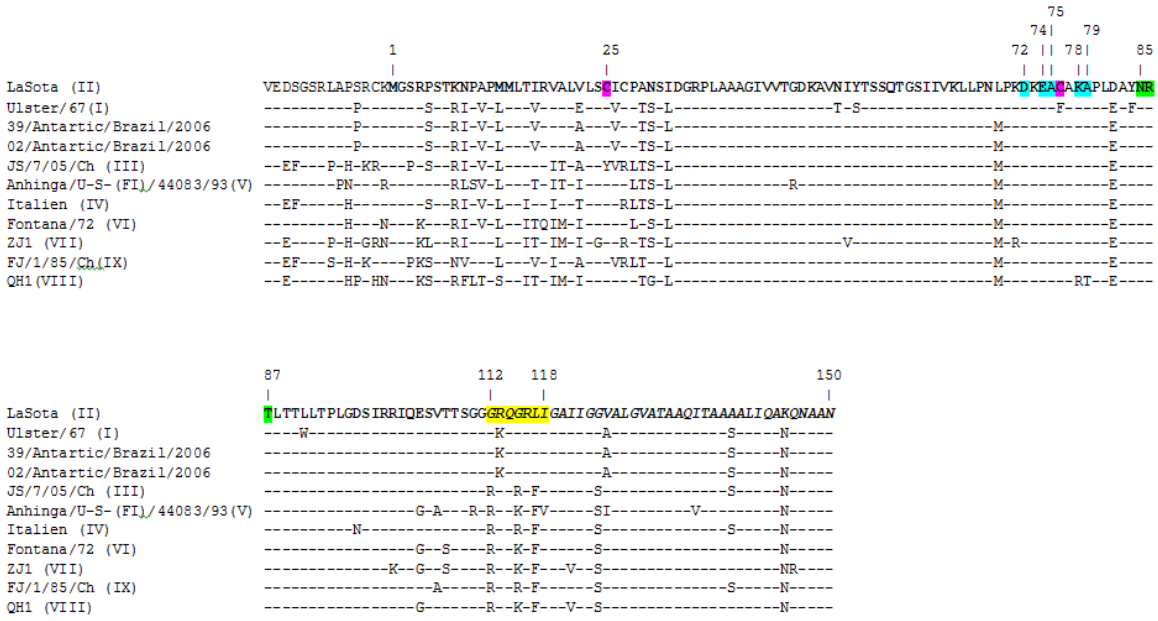


Figura 1. Alineamiento de secuencias aminoacídicas de la proteína F de cepas de NDV. Las cepas de Clase II se muestran sobre la izquierda según su nombre seguido de su genotipo entre paréntesis. La identidad con la cepa *LaSota* se indica por un guion. Las secuencias correspondientes a F2 se señalan en negrita, mientras que las de F1 se aprecian en negrita e itálica. Los valores indicados sobre el alineamiento hacen referencia a los sitios aminoacídicos de la proteína F. El sitio de clivaje de la proteína F está resaltado en amarillo. Las sustituciones de aminoácidos detectadas en sitios antigénicos de mutantes de escape a la neutralización se indican en turquesa [Neyt y cols., 1989; Toyoda y cols., 1988; Yusoff y cols., 1989]. Un potencial sitio aceptor de N-glicosilación en los residuos 85-87 se resalta en verde [Samal y cols., 2012]. Los residuos de cisteína en las posiciones 25 y 76 que se encuentran conservados en la mayoría de los aislados de NDV, se resaltan en color fucsia [Rui y cols., 2010].

Análisis bayesiano de coalescencia de cepas de NDV aisladas en la Antártida

Con el objetivo de investigar los patrones evolutivos de las poblaciones de NDV se utilizó una aproximación bayesiana basada en cadenas de Markov mediante la simulación Monte Carlo (MCMC) implementada en el programa BEAST [Drummond & Rambaut, 2007] . Se analizaron secuencias parciales del gen F de cepas de NDV aisladas en la Antártida, así como de 74 variantes correspondientes a la Clase I y a los genotipos I a XIX de la Clase II (ver Tabla 1). Los resultados del análisis realizado con 50 millones de generaciones de MCMC, utilizando el modelo de sustitución nucleotídica HKY + Γ, el modelo poblacional *Bayesian Skyline* y un reloj molecular relajado se muestran en la Tabla 2 [Drummond y cols., 2005] .

Tabla 2. Análisis Bayesiano de Coalescencia de NDV a partir de secuencias del gen F.

Grupo[a]	Parámetro	Valor[b]	HPD[c]	ESS[d]
Gen F	Log likelihood	-5576	-5595 to -5558	4010
	Posterior	-9101	-9055 a -9150	402
	Prior	-3525	-3573 a -3483	287
	Tasa promedio[e]	1.78×10^{-3}	9.22×10^{-4} a 2.56×10^{-3}	228
	Edad de raíz (años)	194	104 a 308	221
	MRCA[f]	1819	1705 to 1909	

[a]Ver Tabla 1 para apreciar las cepas incluidas en este análisis. [b]Se muestran los valores promedio para cada parámetro. [c]HPD: Valores de densidad de probabilidad superiores al 95% (*highest probability density*). [d]ESS: Tamaño efectivo de muestreo (effective sample size). [e]Tasa evolutiva promedio: Se expresa en sustituciones/sitio/año. [f]MRCA: Año del Ancestro Común Más Reciente

La tasa evolutiva promedio para las secuencias analizadas en el presente estudio se estimó en 1.78×10^{-3} sustituciones por sitio por año (s/s/a). El árbol de máxima credibilidad de clados reveló que todas las cepas de los distintos genotipos de Clase II evolucionaron a partir de un ancestro que circuló alrededor del año 1883 (130 años antes de los aislamientos más recientes incluidos en este estudio, ver Figura 2). Asimismo, las cepas de NDV de ambas Clases divergieron de un ancestro común que existió alrededor del año 1819 (Tabla 2 y Figura 2). Por su parte, las variantes aisladas en Antártida (genotipo I de Clase II) y la estirpe Ulster/67 (cepa referencia de este genotipo) evolucionaron a partir de un ancestro que circuló cerca del año 1958 (Figura 2). El gráfico BSP sugiere que el tamaño poblacional efectivo de NDV se mantuvo constante hasta fines de 1980 (Figura 3), donde se observa un marcado declive.

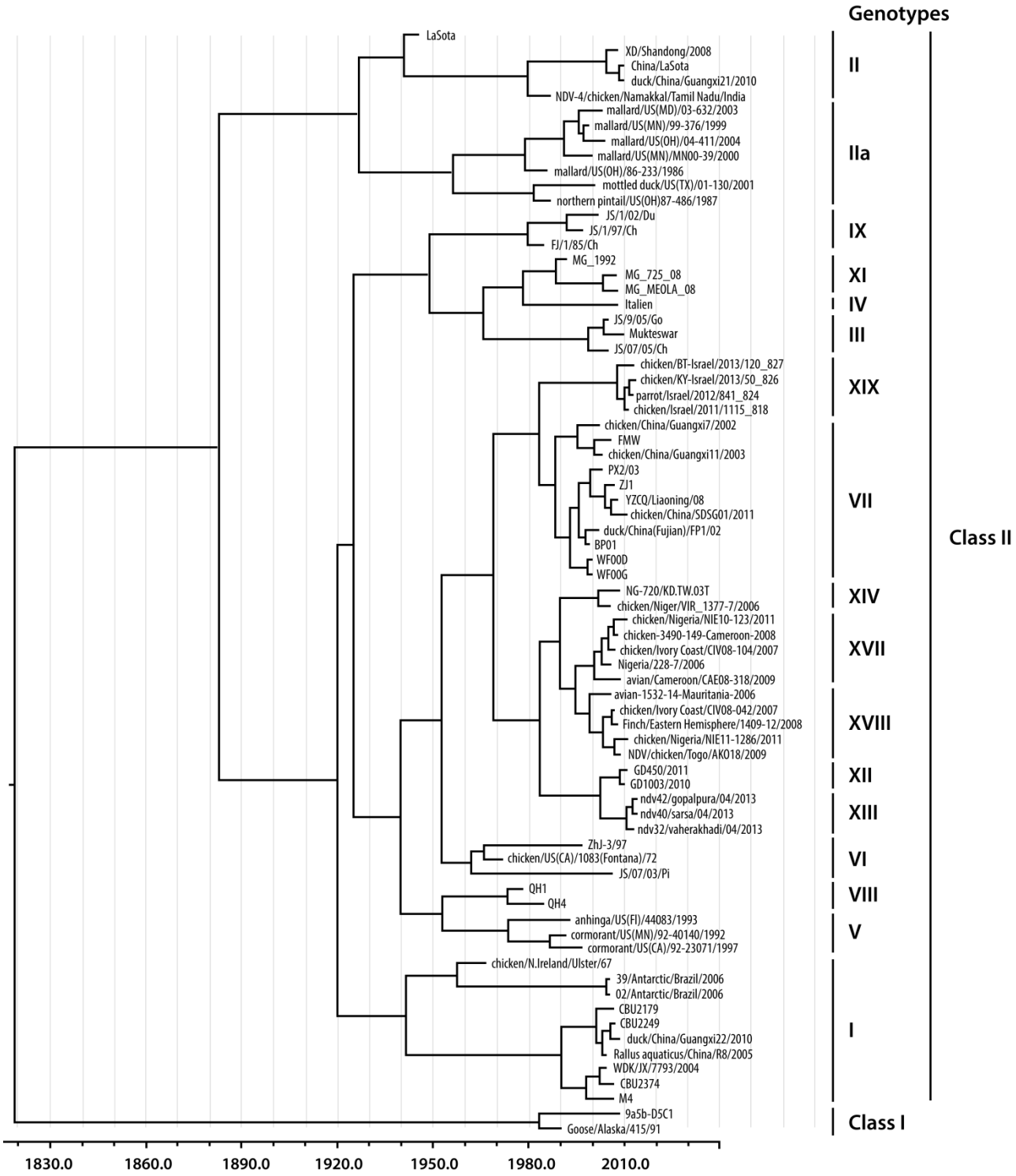


Figura 2. Árbol de máxima credibilidad de clados.

Las cepas se indican por su nombre y genotipo. Se incluyen secuencias correspondientes a la Clase I, así como de diferentes genotipos (I al XIX) de Clase II. Los años se muestran en el eje de las X.

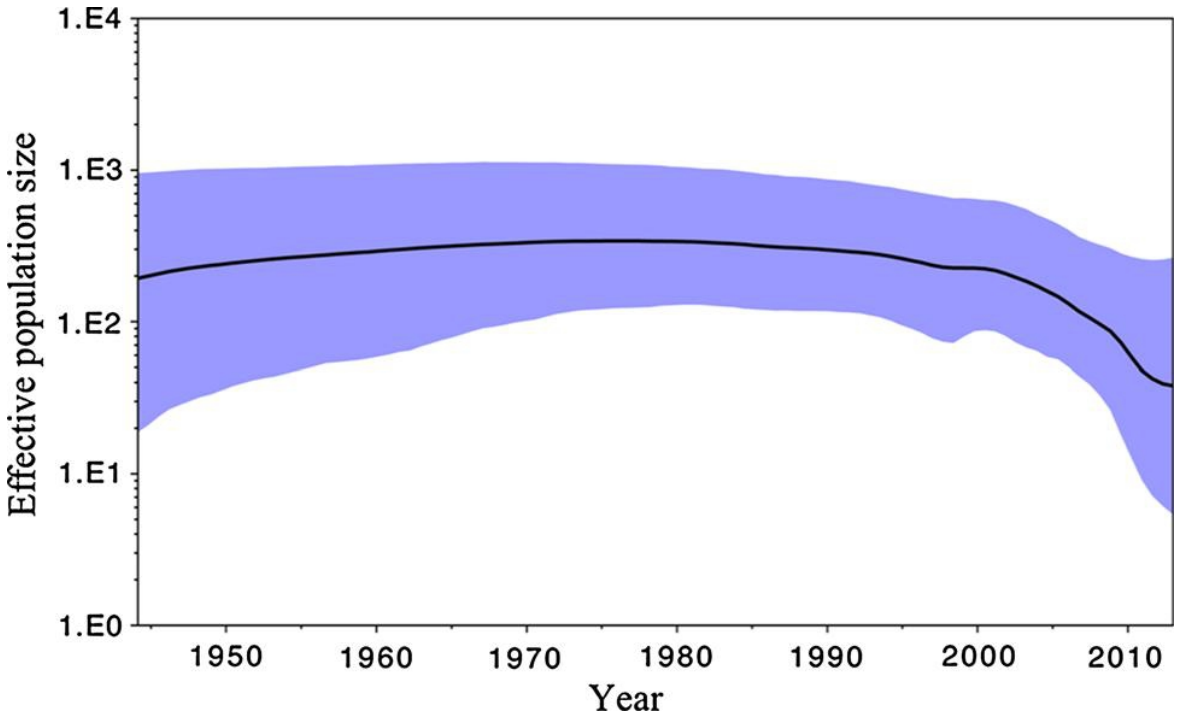


Figura 3. BSP de la historia poblacional de cepas de NDV. Se indican los años y el tamaño poblacional efectivo en los ejes X e Y, respectivamente. La línea sólida negra indica la media estimada mientras que el área celeste representa el 95% HPD.

DISCUSIÓN

En base a los análisis realizados en el presente estudio, las cepas de NDV aisladas de pingüinos en la Antártida fueron asignadas al genotipo I de Clase II (Figura 2), en concordancia con estudios previos [Thomazelli y cols., 2010]. Este resultado también se ajusta con estudios antigénicos realizados en cepas de NDV aisladas de pingüinos de la Antártida, que indicaban reactividad frente a un anticuerpo monoclonal dirigido contra la cepa Ulster/67 (referencia del genotipo I) [Alexander y cols., 1989]. Se han asociado variantes de este genotipo con brotes ocurridos en Australia entre 1998 y 2000 [Gould y cols., 2001]. Asimismo, se caracterizaron estas cepas epidémicas como

velogénicas, aunque reportes previos demostraron que el origen de estos virus se remonta a variantes de baja virulencia circulantes en aves acuáticas justo antes de dichos brotes [Kattenbelt y cols., 2006].

Se ha observado que cepas de NDV que circulan en una especie aviar particular pueden tener la capacidad de causar enfermedades en otras especies de aves. Por ejemplo, variantes de NDV que circulan en columbiformes (palomas) han sido reportadas de ser responsables de brotes en pollos [Alexander, 2000; Werner y cols., 1999]. Asimismo, prácticamente todas las especies de aves domésticas y silvestres son susceptibles a la infección por NDV [Alexander, 2000]. Por lo tanto, aunque parece poco probable la posibilidad de un contacto directo entre pingüinos y pollos, otras aves silvestres pueden actuar como portadoras de diferentes cepas de NDV y transmitirlas mediante rutas que hasta el momento no se conocen plenamente [Snoeck y cols., 2013]. La presencia de cepas de NDV en la Antártida, donde viven otras especies aviares, indica la importancia de la caracterización de variantes en todas las regiones del mundo. Los genotipos V, VI, y VII de Clase II circulan actualmente en pollos de todo el mundo [Xiao y cols., 2013]. El rol de la Antártida en el mantenimiento de otros genotipos de NDV refuerza la importancia de llevar a cabo estudios de vigilancia profundos.

La secuencia del sitio de clivaje de la proteína F ha demostrado ser un importante determinante de virulencia de NDV [Wakamatsu y cols., 2006]. El análisis de este motivo en las variantes aisladas en la Antártida indica que se tratarían de cepas no virulentas (Figura 1). Estas regiones son insensibles a proteasas intracelulares, por lo que dependen de proteasas extracelulares secretadas para su clivaje, lo que limita la replicación de estas cepas no virulentas en los tractos respiratorios y entéricos [de Leeuw y cols., 2005; Panda y cols., 2004; Samal y cols., 2012]. De todas formas, se requieren más estudios para confirmar el fenotipo no virulento (lentogénico) de las estirpes de NDV aisladas de pingüinos de la Antártida.

Los resultados de los análisis de coalescencia realizados revelan una tasa evolutiva de 1.78×10^{-3} s/s/y para las cepas de NDV analizadas (Tabla 2). Esta tasa es levemente superior a las estimaciones realizadas mediante el análisis de secuencias completas del gen F (1.35×10^{-3} s/s/y), aunque se incluye en el intervalo de confianza de este reporte ($0.71 - 1.98 \times 10^{-3}$ s/s/y) [Chong y cols., 2010]. Asimismo, esta tasa evolutiva es comparable con valores estimados para otros virus de ARN de rápida evolución, tales

como HIV-1 (2.4 x 10^{-3} s/s/y) [Korber y cols., 2000] y HCV (3.4 x 10^{-3} s/s/y) [Allain y cols., 2000], entre otros.

El MRCA para todos los virus de Clase II fue estimado alrededor del año 1883 (Figura 2). Esta estimación se ajusta a reportes previos que ubicaron este ancestro en el año 1885 [Chong y cols., 2010], así como a estudios realizados en 1956 que sugirieron a NDV como el agente causal de una epidemia reportada en aves domésticas en el Noroeste de Escocia entre 1897 y 1898 [Macpherson, 1956].

Estudios recientes han explorado la historia demográfica de variantes de NDV de Clase II (genotipos I al VII), sugiriendo el mantenimiento de un tamaño poblacional constante hasta fines de los 90', donde se aprecia una disminución repentina con posterior recuperación alrededor del 2000 [Chong y cols., 2010]. Un comportamiento poblacional similar fue sugerido por los análisis desarrollados en el presente estudio. Los mismos se resumen en un BSP que presenta un estrecho rango de valores con probabilidad posterior superior al 95% (Figura 3). Curiosamente, la dinámica poblacional observada en los últimos años de nuestro análisis sugiere un comportamiento diferente en comparación con reportes previos, ya que se observó una disminución continua persistente en el tamaño efectivo de la población. Esta diferencia puede ser explicada por el mayor número de genotipos de Clase II considerados en el presente análisis (I-XIX), así como al hecho de que distintos genotipos pueden presentar patrones poblacionales diferentes [Chong y cols., 2010]. Aunque las razones de la disminución observada actualmente se desconocen, se ha sugerido previamente al cambio climático y las medidas de control contra la gripe aviar como posibles factores [Chong y cols., 2010]. Más estudios deben llevarse a cabo con el fin de abordar estas cuestiones.

Algunos reportes sugieren que NDV parece estar evolucionando rápidamente hacia una mayor virulencia [Miller y cols., 2010]. Además, se ha reportado un aumento en la patogenicidad, así como en el rango de huésped, con la ocurrencia de brotes incluso en animales vacunados [Nakamura y cols., 2008; Wan y cols., 2004]. Estos hechos revelan la importancia de llevar a cabo estudios que promuevan la caracterización de nuevas cepas aisladas durante el curso de brotes epidémicos en todo el mundo, lo que permitirá investigar la forma en que NDV está evolucionando y sus patrones de dispersión.

CONCLUSIONES

Respecto al conjunto de análisis llevados a cabo en el presente trabajo se logró asignar a las cepas de NDV, aisladas de pingüinos de la Antártida, al genotipo I de Clase II. El análisis del sitio de clivaje de la proteína F indicó que estas variantes corresponderían a cepas no virulentas. Sin embargo estudios más exhaustivos son requeridos para confirmar el fenotipo lentogénico de las mismas. Los resultados del análisis Bayesiano de coalescencia revelaron una tasa evolutiva de 1.78×10^{-3} sustituciones/sitio/año para las cepas de NDV procesadas. Se pudo estimar el tiempo del MRCA para todos los virus de Clase II. Esta estimación ubica al ancestro cerca del año 1883. El análisis demográfico de las variantes de NDV sugiere un tamaño poblacional efectivo constante hasta fines de la década del 80' en dónde se observa una disminución persistente hacia el presente.

REFERENCIAS

Afonso C and Miller P. (2013). Newcastle disease: progress and gaps in the development of vaccines and diagnostic tools. *Dev. Biol. (Basel).* 135, 95–106. DOI:10.1159/000178459

Abolnik C, Horner R, Bisschop S, Parker M, Romito M and Vijoen G. (2004). A phylogenetic study of South African Newcastle disease virus strain isolated between 1990 and 2002 suggests epidemiological origins in the Raf East. Arch. Virol. 149, 603-19. doi:10.1007/s00705-003-0218-2.

Alexander D. (2000). Newcastle disease and other avian paramyxoviruses. *Rev. Sci. Tech.* 19, 443–62.

Alexander D, Manvell R, Collins M, Brockman S, Westbury H, Morgan I, et al. (1989). Characterization of paramyxoviruses isolated from penguins in Antarctica and sub-Antarctica during 1976-1979. *Arch. Virol.* 109, 135–43.

Allain J, Dong Y, Vandamme A, Moulton V and Salemi M. (2000). Evolutionary rate and genetic drift of hepatitis C virus are not correlated with the host immune response: studies of infected donor-recipient clusters. *J. Virol.* 74, 2541–9.

Anders S, Pyl PT, Huber W. (2014). HTSeq — A Python framework to work with high-throughput sequencing data Bioinformatics. DOI:10.1093/bioinformatics/btu638

Andino R, Domingo E. (2015) Viral quasispecies. *Virology*, DOI: 10.1016/j.virol.2015.03.022.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arias A, Ruiz-Jarabo C, Escarmis C, Domingo E. (2004). Fitness increase of memory genomes in a viral quasispecies. *J. Mol. Biol*. 339, 405–412.

Ariën K, Troyer R, Gali Y, Colebunders R, Arts E, Vanham G. (2005) Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time. *AIDS*. 19,1555–64.
Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mandoiu I, et al. (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics* 12 Suppl 6, S1.

Balzola F, Bernstein C, Ho G, Qin J, Li R, Raes J, Arumugam M, Burgdorf K, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende D, Li J, Xu J, Li S, Li D, Cao J, Bo Wang, Liang H, Zheng H, Xie Y, Tap T, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen H, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Shengting Li, Jian M, Zhou Y, Yingrui Li, Zhang X, Songgang Li, Nan Qin, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Consortium M, Bork P, Ehrlich S, Wang J. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65 DOI:10.1038/nature08821.

Bar, Li H, Chamberland A, Tremblay C, Routy J, Grayson T, Sun C, Wang S, Learn G, Morgan C, Schumacher J, Haynes B, Keele B, Hahn B, Shaw G. (2010) Wide variation in the multiplicity of HIV-1 infection among injection drug usersJ. *Virol*. 84, 6241-6247.
Barzon L, Lavezzo E, Militello V, Toppo S, Palù G. (2011). Applications of Next-Generation Sequencing Technologies to Diagnostic Virology. *Int. J. Mol. Sci*. 12,7861-7884. DOI:10.3390/ijms12117861.

Batten CA, Maan S, Shaw AE, Maan NS, Mertens PP. (2008) A European field strain of bluetongue virus derived from two parental vaccine strains by genome segment reassortment. *Virus Res.* 2008 Oct; 137(1):56-63.

Beerenwinkel N, Gunthard H, Roth V, Metzner K J. (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol* 3, 329.

Beerenwinkel N, Zagordi O. (2011). Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*. 1, 413–418.

Biebricher C, Eigen M. (2006). What is a quasispecies? *Curr Top Microbiol Immunol*, *299*, 1-31.

Borucki M, Chen-Harris H, Lao V, Vanier G, Wadford D, Messenger S, Allen J. (2013) Ultra-Deep Sequencing of Intra-host Rabies Virus Populations during Cross-species Transmission. *PLoS Negl Trop Dis*. 7(11): e2555.

Briones C, Domingo E. (2008). Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications. *AIDS Rev.* 10, 93–109.

Briones C, de Vicente A, Molina-Paris C, Domingo E. (2006). Minority memory genomes can influence the evolution of HIV-1 quasispecies in vivo. *Gene.* 384, 129–138. DOI: 10.1016/j.gene.2006.07.037.

Briones C, Domingo E, Molina-París C. (2003). Memory in retroviral quasispecies: experimental evidence and theoretical model for human immunodeficiency virus.
J.Mol.Biol. 331, 213–229.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data.Bioinformatics 30, 2114–2120. DOI: 10.1093/bioinformatics/btu170.

Boni MF, Zhou Y, Taubenberger JK, Holmes EC. (2008) Homologous recombination is very rare or absent in human influenza A virus. *J Virol*. 82:4807–11.
DOI: 10.1128/JVI.02683-07 PMID: 18353939

Boni MF, de Jong MD, van Doorn HR, van Doorn, Holmes EC. (2010) Guidelines for identifying homologous recombination events in influenza A virus. *PLoS One.* 5:e10434. DOI: 10.1371/journal.pone.0010434.

Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, Cameron B, Maher L, Dore GJ, White PA, Lloyd AR. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog*. 7:e1002243.

Burrows M, Wheeler D. (1994). A block sorting lossless data compression algorithm, Technical Report 124, Digital Equipment Corporation.

Choi K, Lee E, Jeon W and Kwon J. (2010). Antigenic and immunogenic investigation of the virulence motif of the Newcastle disease virus fusion protein. *J. Vet. Sci.* 11, 205–11.

Chong Y, Padhi A, Hudson P and Poss M. (2010). The effect of vaccination on the evolution

and population dynamics of avian paramyxovirus-1. *PLoS Pathog.* 6, e1000872. doi:10.1371/journal.ppat.1000872.

Coffin M. (1979) Structure, replication and recombination of retroviruses genomes: some unifying hypotheses. *J Gen Virol* 42, 1–26. DOI:10.1099/0022-1317-42-1-1.

Colina R, Cristina J. (2004) Evidence of intratypic recombination in natural populations of hepatitis C virus. *J Gen Virol.* 85, 31-37. DOI:10.1186/1743-422X-3-53.

Colman PM. (1994). Influenza virus neuraminidase: structure, antibodies, and inhibitors. Protein Sci. 3(10):1687–1696.

Czeglédi A, Ujvári D, Somogyi E, Wehmann E, Werner O and Lomniczi B. (2006). Third genome size category of avian paramyxovirus serotype 1 (Newcastle disease virus) and evolutionary implications. *Virus Res.* 120, 36–48. DOIi:10.1016/j.virusres.2005.11.009.

David M, & Howley P. (2007). *Fields virology*. 5th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

De Jong JC, Rimmelzwann GF, Fouchier RA, Osterhaus AD. (2000) Influenza virus: a master of metamorphosis. *J Infect* 40: 428H433.

de Leeuw O, Koch G, Hartog L, Ravenshorst N and Peeters B. (2005). Virulence of Newcastle disease virus is determined by the cleavage site of the fusion protein and by both the stem region and globular head of the haemagglutinin-neuraminidase protein. *J. Gen. Virol.* 86, 1759–69. doi:10.1099/vir.0.80822-0.

Delport W, Poon A, Frost S and Kosakovsky Pond, S. L. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26, 2455–7. doi:10.1093/bioinformatics/btq429.

Diel D, da Silva L, Liu H, Wang Z, Miller P and Afonso C (2012a). Genetic diversity of avian paramyxovirus type 1: proposal for a unified nomenclature and classification system of Newcastle disease virus genotypes. *Infect. Genet. Evol.* 12, 1770–9. doi:10.1016/j.meegid.2012.07.012.

Diel D, Susta L, Cardenas Garcia S, Killian M, Brown C, Miller P, et al. (2012b). Complete genome and clinicopathological characterization of a virulent Newcastle disease virus isolate from South America. *J. Clin. Microbiol.* 50, 378–87. doi:10.1128/JCM.06018-11.

Domingo E. (2000). Viruses at the edge of adaptation. *Virology*. 270, 251–253.

Domingo E. (2005). Viruses as Quasispecies: Biological Implications. *Quasispecies:* Concept and Implications. *Virology*. 51–82.

Domingo E, Gomez J. (2007). Quasispecies and its impact on viral hepatitis, *Virus Res*. 127, 131–150.

Domingo E, Martin V, Perales C, Dávila M. (2006). Viral fitness can influence the repertorie of virus variants selected by antibodies. *J Mol Biol*. 8, 44-54.

Domingo E, Sheldon J, Perales C. (2012). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews 76*, 159–216. DOI:10.1128/MMBR.05023-11.

Drummond A, Ho S, Phillips M and Rambaut A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4, e88. DOI:10.1371/journal.pbio.0040088.

Drummond AJ, and Rambaut A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214. DOI:10.1186/1471-2148-7-214.

Drummond A, Rambaut A, Shapiro B and Pybus O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–92. DOI:10.1093/molbev/msi103.

Edgar R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–7. doi:10.1093/nar/gkh340.

Eigen M. (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58, 465–523. DOI:10.1007/ BF00623322.

Eigen M, Schuster P. (1977). The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften*, 64, 541–565.

Eigen M, Schuster P. (1978b). The hypercycle, a principle of natural self-organization. Part C: the realistic hypercycle. *Naturwissenschaften*. 65, 341- 369. DOI: 10.1007/BF00439699.

Eigen M, Schuster P. (1978a). Hypercycle—principle of natural self- organization. Part B: the abstract hypercycle. *Naturwissenschaften* 65, 7–41.17.

Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N. (2008). Viral population estimation using pyrosequencing. *PLoS Comput Biol*. 9, 4 DOI: 10.1371/journal.pcbi.1000074.

Escarmís C, Dávila M, Domingo E. (1999). Multiple molecular pathways for fitness recovery of an RNA virus debilitated by operation of Muller's ratchet. *J. Mol. Biol.* 285,495–505.

Eshaghi A, Shalhoub S, Rosenfeld P, Li A, Higgins RR, Stogios PJ, Savchenko A, Bastien N, Li Y, Rotstein C, and Gubbay JB. (2014). Multiple Influenza A (H3N2) Mutations Conferring Resistance to Neuraminidase Inhibitors in a Bone Marrow Transplant Recipient. *Antimicrobial Agents and Chemotherapy*, 58(12), 7188–7197. DOI:10.1128/AAC.03667-14.

Esteller M. (2011). Non-coding RNAs in human disease. Nature Reviews Genetics. 12, 861–874. doi:10.1038/nrg3074.

Ewing B, Hillier L, Wendl MC, Green P. (1998a). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*. 8 (3): 175–185. DOI: 10.1101/gr.8.3.175.

Ewing B, Green P (1998b). Base-calling of automated sequencer traces using phred. Error probabilities. *Genome research*. 8 (3): 186–194. DOI:10.1101/gr.8.3.186

Ferguson NM, Gavani AP, Bush RM (2003). Ecological and immunological determinants of influenza evolution. *Nature* 40:218H228.

Fernandes C, Varani A, Lemos E, de Miranda V, Silva K, Fernando F, et al. (2014). Molecular and phylogenetic characterization based on the complete genome of a virulent pathotype of Newcastle disease virus isolated in the 1970s in Brazil. *Infect. Genet. Evol.* 26, 160–7. doi:10.1016/j.meegid.2014.05.014.

Ferragina P, Manzini G. (2000). Opportunistic data structures with applications. *In Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*, pages 390-398.

Ferragina P, Manzini G. (2001). An experimental study of a compressed index. Information Sciences: special issue on “Dictionary Based Compression”, 135:13-28.

Ferragina P, Manzini G. (2001). An experimental study of a compressed index. In Proc. 12th ACM-SIAM Symposium on Discrete Algorithms, pages 269-278.

Fisher R, Gert U, Zyl V, Simon A, Travers A, Sergei L, Kosakovsky Pond, Engelbrech S, Murrell B, Scheffler K , Smith D. (2012). Deep Sequencing Reveals Minor Protease Resistance Mutations in Patients Failing a Protease Inhibitor Regimen. *J Virol*. 86(11): 6231–6237. DOI: 10.11128/JVI.06541-11.

García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics.* 15;28(20):2678-9. DOI: 10.1093/bioinformatics/bts503.

Garten R, Davis C, Russell C. (2009). Antigenic and genetic characteristics of swine origin 2009 (H1N1) influenza viruses circulating in humans. *Science* 325 (5937):197-201.

Ghedin E, Fitch A, Boyne A, Griesemer S, DePasse J, Bera J, Zhang X, Halpin RA, Smit M, Jennings L, George KS, Holmes EC, and Spiro DJ. (2009). Mixed Infection and the Genesis of Influenza Virus Diversity . *Journal of Virology*, *83*(17), 8832–8841. DOI: 10.1128/JVI.00773-09

Ghedin E, Laplante J, DePasse J, Wentworth D, Santos R, Lepow M, Porter J, Stellrecht K, Lin X, Operario D, Griesemer S, Fitch A, Halpin R, Stockwell T, Spiro D, Holmes E, St. George K. (2011). Deep Sequencing Reveals Mixed Infection with 2009 Pandemic Influenza A (H1N1) Virus Strains and the Emergence of Oseltamivir Resistance. *The Journal of Infectious Diseases*. 203, 168–174.

Ghedin E, Holmes E, DePasse J, Pinilla L, Fitch A, Hamelin M-H, Papenburg J, Boivin G. (2012). Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. *Journal of infectious Diseases*. 206:10, 1504-1511.

Giallonardo FD, Töpfer A, Rey M, Prabhakaran S, Duport Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR, Patrignani A, Däumer M, Beisel C, Rusert P, Trkola A, Günthard HF, Roth V, Beerenwinkel N, and Metzner KJ. (2014). Full-length haplotype

reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Research*, Volume 42, Issue 14, e115, DOI:10.1093/nar/gku537.

Gilbert and Christopher L. Dupont. (2011). Microbial Metagenomics: Beyond the Genome. The Annual Review of Marine Science is online at marine.*annualreviews*. DOI: 10.1146/annurev-marine-120709-142811.

Gonzaga-Jauregui C, Lupski, JR, & Gibbs RA. (2012). Human Genome Sequencing in Health and Disease. *Annual Review of Medicine*, *63*, 35–61. DOI:10.1146/annurev-med-051010-162644.

González-Candelas F, López-Labrador FX, Bracho MA. (2011). Recombination in hepatitis C virus. *Viruses.* 3:2006-2024 DOI: 10.3390/v3102006.

Goymer P. (2007). "Synonymous mutations break their silence". *Nature Reviews Genetics*. 8,92. doi:10.1038/nrg2056.

Gould A, Kattenbelt J, Selleck P, Hansson E, Della-Porta A and Westbury HA. (2001). Virulent Newcastle disease in Australia: molecular epidemiological analysis of viruses isolated prior to and during the outbreaks of 1998-2000. *Virus Res.* 77, 51- 60.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinforma. Oxf. Engl. 29:1072–5.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes D, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, and Regev A. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols*, 8(8), DOI: 10.1038/nprot.2013.084.

Hajarizadeh B, Grebely J, Applegate T, Matthews GV, Amin J, Petoumenos K, Hellard M, Rawlinson W, Lloyd A, Kaldor J, Dore GJ. (2014) Dynamics of HCV RNA levels during acute hepatitis C virus infection. *J Med Virol*. 86,1722–1729. DOI:10.1002/jmv.24010.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, *10*(3), R32. DOI:10.1186/gb-2009-10-3-r32.

Henn M , Boutwell C, Charlebois P, Lennon N , Power K , Macalalad A, Berlin A , Malboeuf C, Ryan E, Gnerre S, Zody M, Erlich R, Green L, Berical A, Wang Y, Casali M, Streeck H, Bloom A, Dudek T, Tully D, Newman R, Axten K, Gladden A, Battis L, Kemper M, Zeng Q, Shea T, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Gunthard H, Brumme Z, Brumme C, Bazner S, Rychert J, Tinsley J , Mayer K , Rosenberg E, Pereyra F, Levin J, Young S, Jessen H, Altfeld M, BirrenB, Walker B, Allen T. (2012). Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathog* 8(3): e1002529. DOI:10.1371/journal.ppat.1002529.

Herczeg J, Wehmann E, Bragg R, Travassos Dias P, Hadjiev G, Werner O, et al. (1999). Two novel genetic groups (VIIb and VIII) responsible for recent Newcastle disease outbreaks in Southern Africa, one (VIIb) of which reached Southern Europe. *Arch. Virol.* 144, 2087–99.

Hillerman MR. (2002) Realities and enigmas of human virus influenza: pathogenesis, epidemiology and control. *Vaccine* 20:3068-87.

Holland J, Spindler k, Horodyski F, Grabau E, Nichol S, VandePol S. (1982). Rapid Evolution of RNA Genomes. *Science*. 215, 1577-85.

Homer N, Merriman B, Nelson SF. (2009). BFAST: an alignment tool for large scale genome resequencing. *PLoS One* 4(11):e7767

Horner DS, Pavesi G, Castrignanò T, De Meo PD, Liuni S, Sammeth M, Picardi E, Pesole G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform* 11:181–197.

Huang A, Kantor R, Delong A, Schreier L, Istrail S. (2012). QColors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. *In Silico Biol* 11, 193–201.

Humphrey W, Dalke A, Schulten K. (1996). VMD: Visual molecular dynamics. J Mol Graph. 14(1):33-8, 27-8.

Hurt AC, Besselaar TG, Daniels RS, et al. (2016). Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2014-2015. *Antiviral Res*. 132:178–85 DOI: 10.1016/j.antiviral.2016.06.001

Isakov O, Bordería AV, Golan D, Hamenahem A, Celniker G, Yoffe L, Blanc H, Vignuzzi M, and Shomron N. (2015). Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. *Bioinformatics*, 31;13(1):2141–2150, DOI: 10.1093/bioinformatics/btv101

Jacobi MN, Nordahl M. (2006) Quasispecies and recombination. *Theor. Popul. Biol.* 70:479-485.

Jayasundara D, Saeed I, Maheswararajah S, Chang BC, Tang SL, Halgamuge SK. (2014). ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics.* 15;31(6):886-96. DOI: 10.1093/bioinformatics/btu754.

Kim J, Lee I, Park S, Bae JY, Yoo K, Cheong HJ, Noh JY, Hong KW, Lemey P, Vrancken B, Kim J, Nam M, Yun S-H, Cho W-I, Song JY, Kim WJ, Park MS, Song J-W, Kee S-H, Song K-J, and Park M-S. (2017). Phylogenetic relationships of the HA and NA genes between vaccine and seasonal influenza A(H3N2) strains in Korea. *PLoS One.* 12(3): e0172059. Doi: 10.1371/journal.pone.0172059

Jojic V, Hertz T, & Jojic N. (2008). Population sequencing using short reads: HIV as a case study. *Pac Symp Biocomput.* 114–125.

Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. (Version 1.33)

Kalinina O, Norder H, Mukomolov S, Magnius LO. (2002) A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *J Virol*; 76:40344043.

Kampmann M, Fordyce S, Ávila-Arcos M, Rasmussen M, Willerslev E, Nielsen L, Gilbert T. (2011). A simple method for the parallel deep sequencing of full influenza A genomes. *Journal of Virological Methods.* 178, 243–248.

Kattenbelt J, Stevens M and Gould, A. R. (2006). Sequence variation in the Newcastle disease virus genome. *Virus Res.* 116, 168–84. doi:10.1016/j.virusres.2005.10.001.

Kawaoka Y, Krauss S, Webster RG. (1989). Avian to human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *J Virol* 63(11): 4603H8.

Kim L, King D, Curry P, Suarez D, Swayne D, Stallknecht D, et al. (2007). Phylogenetic diversity among low-virulence newcastle disease viruses from waterfowl and shorebirds and comparison of genotype distributions to those of poultry-origin isolates. *J. Virol.* 81, 12641–53. doi:10.1128/JVI.00843-07.

Kriesel J, Hobbs M, Jones B, Milash B, Nagra R, Fischer K. (2012). Deep Sequencing for the Detection of Virus-Like Sequences in the Brains of Patients with Multiple Sclerosis: Detection of GBV-C in Human Brain. *PLoS One*. 7(3): e31886. DOI: 10.1371/journal.pone.0031886.

Koelle K, Cobey S, Grenfell B, Pascual M. (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*. 314, 1898–1903.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, et al. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science* 288, 1789–96.

Lai M. (1992). RNA recombination in animal and plant viruses. *Microbiol Rev. 56,61- 79.*

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*10:R25.

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

Lassmann T, Hayashizaki Y, Daub CO. (2011). SAMStat: monitoring biases in next generation sequencing data. Bioinformatics. 27,130–131. DOI:10.1093/bioinformatics/btq614

Lau KA, Wong JJL. (2013) Current trends of HIV recombination worldwide. *Infectious Disease Reports.* 5(Suppl 1): e4. DOI: 10.4081/idr.2013.s1.e4.

Lauring A, Andino R. (2010). Quasispecies Theory and the Behavior of RNA Viruses, *Plos Pathog. 6*(7). DOI:10.1371/journal.ppat.1001005.

Lázaro Lázaro Ester, Homs C. (2002). Virus emergentes: la amenaza oculta. *Equipo Sirius.*

Lee C-Y, Chiu Y-C, Wang L-B, Kuo Y-L, Chuang E Y, Lai L-C, Tsai M-H. (2013). Common applications of next-generation sequencing technologies in genomic research. DOI: 10.3978/j.issn.2218-676X.2013.02.09. Transl Cancer Res. 2(1):33-45

Le Guillou-Guillemette, Vallet S, Gaudy-Graffin C, Payan C, Pivert A, Goudeau A, & Lunel-Fabiani F. (2007) Genetic diversity of the hepatitis C virus: Impact and issues in the antiviral therapy. *World Journal of Gastroenterology*. 13,2416-2426. DOI:10.3748/wjg.v13.i17.2416.

Li H, & Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14),1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, & Wang J. (2009a). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124–1132. DOI: 10.1101/gr.088013.108.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J, (2009b). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics.* 25:1966–1967. DOI: 10.1093/bioinformatics/btp336.

Liu D, Mewalal R, Hu R, Tuskan G A, and Yang X. (2017). New technologies accelerate the exploration of non-coding RNAs in horticultural plants. Horticulture Research 4, 17031; DOI:10.1038/hortres.2017.31

Lomniczi B, Wehmann E, Herczeg J, Ballagi-Pordány A, Kaleta E, Werner O, et al. (1998). Newcastle disease outbreaks in recent years in western Europe were caused by an old (VI) and a novel genotype (VII). *Arch. Virol.* 143, 49–64.

Lorenzo-Redondo R, Borderia AV, Lopez-Galindez C. (2011). Dynamics of in vitro fitness recovery of HIV-1. *J. Virol.* 85, 1861–1870.

Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, et al. (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol 8: e1002417.

Macpherson L. (1956). Some Observations On The Epizootiology Of NewCastle Disease. *Can. J. Comp. Med. Vet. Sci.* 20, 155–68.

Maminiaina O, Gil P, Briand F, Albina E, Keita D, Andriamanivo H, et al. (2010). Newcastle disease virus in Madagascar: identification of an original genotype possibly deriving from a died out ancestor of genotype IV. *PLoS One* 5, e13987. DOI:10.1371/journal.pone.0013987.

Mancuso N, Tork B, Skums P, Mandoiu I, Zelikovsky A. (2011). Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads. In Bioinformatics and Biomedicine Workshops (BIBMW), IEEE International Conference 94–101.

Mangul S, Wu NC, Mancuso N, Zelikovsky A, Sun R, Eskin E. (2014). Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics*, 30(12), i329-i337.

Mardis ER. (2013). Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif). 6,287-303. doi: 10.1146/annurev-anchem-062012-092628.

Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC. (2013) Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathog. 9(6):e1003421.* DOI: 10.1371/journal.ppat.1003421.

Martin M., (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal, 17(1):10-12.

Mase M, Murayama K, Karino A and Inoue T. (2011). Analysis of the fusion protein gene of Newcastle disease viruses isolated in Japan. *J. Vet. Med. Sci.* 73, 47–54.

Hussain M, Galvin HD, Haw TY, Nutsford AN, and Husain M. (2017). Drug resistance in influenza A virus: the epidemiology and management. *Infect Drug Resist*. 2017; 10: 121–134. doi: 10.2147/IDR.S105473.

Meijer A, Rebelo-de-Andrade H, Correia V, et al. (2014). Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2012-2013. *Antiviral Res*. 110:31–41.

MCElroy K, Thomas T, Luciani F. (2014). Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb Inform Exp.* 4,1. DOI:10.1186/2042-5783-4-1.

McHardy A, Adams B. (2009). The role of genomics in tracking the evolution of influenza A virus. *Plos Path* 5: e1000566.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, and DePristo MA. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. DOI:10.1101/gr.107524.110.

Medvedev P, Stanciu M, Brudno M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 6:S13–S20.

Meyerhans A, Vartanian J. (1999). The fidelity of cellular and viral polymerases and its manipulation for hypermutagenesis. *Academic Press.* 87-114. DOI: 10.1016/B978-012220360-2/50006-4.

Mielczarek M, Szyda J. (2016). Review of alignment and SNP calling algorithms for next-generation sequencing data. *J Appl Genet.* 57(1):71-9. DOI: 10.1007/s13353-015-0292-7.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD and Marshall D. (2013). Using Tablet for visual exploration of second-generation sequencing data. Briefings in

Bioinformatics 14(2), 193-202.

Miller P, Decanini E, and Afonso C. (2010). Newcastle disease: evolution of genotypes and the related diagnostic challenges. *Infect. Genet. Evol.* 10, 26–35. doi:10.1016/j.meegid.2009.09.012.

Morelli MJ, Wright CF, Knowles NJ, Juleff N, Paton DJ, King DP, and Haydon DT. (2013). Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Veterinary Research.* 44,12. DOI: 10.1186/1297-9716-44-12.

Moreno MP, Casane D, López L, Cristina J. (2006) Evidence of recombination in quasispecies populations of a Hepatitis C Virus patient undergoing anti-viral therapy. *Virol J.* 3:87. DOI: 10.1186/1743422X-387.

Moreno MP, Alvarez M, López L, Moratorio G, Casane D, Castells M, Castro S, Cristina J, Colina R. (2009) Evidence of recombination in Hepatitis C Virus populations infecting a hemophiliac patient. *Virol J.* 6: 203 DOI:10.1186/1743-422X-6-203.

Nagai Y, Hamaguchi M and Toyoda T. (1989). Molecular biology of Newcastle disease virus. *Prog. Vet. Microbiol. Immunol.* 5, 16–64.

Nakamura K, Ohtsu N, Nakamura T, Yamamoto Y, Yamada M, Mase M, et al. (2008). Pathologic and immunohistochemical studies of Newcastle disease (ND) in broiler chickens vaccinated with ND: severe nonpurulent encephalitis and necrotizing pancreatitis. *Vet. Pathol.* 45, 928–33. DOI:10.1354/vp.45-6-928.

Nayak DP, Balogun RA, Yamada H, Zhou ZH, Barman S. (2009). Influenza virus morphogenesis and budding. Virus Res. 143(2):147–161.

Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3):443-53.

Nelson M, Simonsen L, Viboud C, Miller M, Holmes E. (2007). Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog*. 3, 1220–1228. DOI:10.1371/journal.ppat.0030131.

Nelson M, Viboud C, Simonsen L, Bennett R, Griesemer S, St Geore K, Taylor J, Spiro D, Sengamalay N, Ghedin E, Tauberger J, Holmes E. (2008) Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog*. 4 : e 1 0 0 0 0 1 2 . DOI:10.1371/journal.ppat.1000012.

Neumann G, Noda T, Kawaoka Y. (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 459: 931–9. doi: 10.1038/nature08157.

Neyt C, Geliebter J, Slaoui M, Morales D, Meulemans G and Burny A. (1989). Mutations located on both F1 and F2 subunits of the Newcastle disease virus fusion protein confer resistance to neutralization with monoclonal antibodies. *J. Virol.* 63, 952–4.

Nielsen R, Paul JS, Albrechtsen A, & Song YS. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451. DOI: 10.1038/nrg2986

Nishijima N, Marusawa H, Ueda Y, Takahashi K, Nasu A, Osaki Y, Kou T, Yazumi S, Fujiwara S, Tsuchiya S, Shimizu K, Uemoto S, Chiba T. (2012). Dynamics of Hepatitis B Virus Quasispecies in Association with Nucleos(t)ide Analogue Treatment Determined by Ultra-Deep Sequencing. *Plos One* 7(4): e35052. DOI:10.1371/journal.pone.0035052.

Nobusawa, E. and Sato, K. (2006). Comparison of the Mutation Rates of Human Influenza A and B Viruses. *Journal of Virology*. 80, 3675–3678. DOI:10.1128/JVI.80.7.3675–3678.2006.

Novella IS, et al. (1995). Exponential increases of RNA virus fitness during large population transmissions. *Proc. Natl. Acad. Sci*. 92,5841–5844

OIE (2011). World Organisation for Animal Health.

Okonechnikov K, Conesa A, & García-Alcalde F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2), 292–294. DOI: 10.1093/bioinformatics/btv566

O'Neil ST, & Emrich SJ. (2012). Haplotype and minimum-chimerism consesus determination using short sequence data. *BMC Genomics* 13, S4.

Page K, Nowak M, (2002). Unifying evolutionary dynamics. *J. Theor. Biol*. 219, 93–98.

Paldurai A, Kumar S, Nayak B and Samal S. (2010). Complete genome sequence of highly virulent neurotropic Newcastle disease virus strain Texas GB. *Virus Genes* 41, 67–72. doi:10.1007/s11262-010-0486-3.

Palmer BA, Moreau I, Levis J, Harty C, Crosbie O, Kenny-Walsh E, Fanning LJ. (2012) Insertion and recombination events at hypervariable region 1 over 9.6 years of hepatitis C virus chronic infection. *J Gen Virol.* 93:2614-2624 DOI:10.1099/vir.0.045344-0.

Perales C, Lorenzo-Redondo R, López-Galíndez C, Martínez M, Domingo E. (2010).Mutant spectra in virus behavior. *Future Virol*. 5, 679–698.

Panda A, Huang Z, Elankumaran S, Rockemann D and Samal, S. K. (2004). Role of fusion protein cleavage site in the virulence of Newcastle disease virus. *Microb. Pathog.* 36, 1–10.

Pandey RV, Nolte V, Boenigk J, Schlo¨tterer C. (2011). CANGS DB: a stand-alone web-based database tool for processing, managing and analyzing 454 data in biodiversity studies. BMC Res. Notes 4, 227– 237. doi:10.1186/1756-0500-4-227.

Patel RV, Jain M. (2012). NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE 7, e30619. doi:10.1371/journal.pone. 0030619

Pedersen J, Senne D, Woolcock P, Kinde H, King D, Wise M, et al. (2004). Phylogenetic

relationships among virulent Newcastle disease virus isolates from the 2002-2003 outbreak in California and other recent outbreaks in North America. *J. Clin. Microbiol.* 42, 2329–34. DOI:10.1128/jcm.42.5.2329-2334.2004.

Perozo F, Marcano R and Afonso, C. L. (2012). Biological and phylogenetic characterization of a genotype VII Newcastle disease virus from Venezuela: efficacy of field vaccination. *J. Clin. Microbiol.* 50, 1204–8. doi:10.1128/JCM.06506-11.

Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V. (2014). HIV haplotype inference using a propagating Dirichlet process mixture model. IEEE/ACM *Transactions on Computational Biology and Bioinformatics* (TCBB), 11(1), 182-191.

Prosperi MC. et al. (2011). Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics* 12, 5.

Prosperi MC, & Salemi M. (2012). Qure: software for viral quasispecies reconstruction from nextgeneration sequencing data. *Bioinformatics* 28, 132– 133.

Prosperi MC, Yin L, Nolan DJ, Lowe AD, Goodenow MM, et al. (2013) Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. Scientific reports 3, 2837.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12, 38.

Quinones-Mateu M, Arts E. (2006) Virus fitness: concept, quantification, and application to HIV population dynamics. *Curr Top Microbiol Immunol.* 299, 83–140.

Rambaut A, Pybus O, Nelson M, Viboud C, Tauberger J, Holmes E. (2008) The genomic and epidemiological dynamics of human influenza A virus. *Nature.* 453:615–619. DOI:10.1038/nature06945.

Richman D. (2006). Antiviral drug resistance. *Antiviral Res*. 71, 117–121. DOI: 10.1016/j.antiviral.2006.03.004.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. (2011). Integrative Genomics Viewer. *Nature Biotechnology* 29, 24–26.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192.

Rolland M, Brander C, Nickle DC, Herbeck JT, Gottlieb GS, et al. (2007) HIV-1 over time: fitness loss or robustness gain?. *Nat Rev Microbiol*. 5,1–2.

Ruiz-Jarabo C, Arias A, Baranowski E, Escarmís C, Domingo E. (2000). Memory in viral quasispecies. *J. Virol.* 74, 3543–3547.

Ruiz-Jarabo C, Arias A, París C, Briones C, Baranowski E, Escarmís C, Domingo E. (2002). Duration and fitness dependence of quasispecies memory. *J.Mol. Biol*. 315,285–296.

Ruiz-Jarabo C, Miller E, Gómez-Mariano G, Domingo E. (2003). Synchronous loss of quasispecies memory in parallel viral lineages: a deterministic feature of viral quasispecies.

J.Mol. Biol. 333, 553–563.

Rui Z, Juan P, Jingliang S, Jixun Z, Xiaoting W, Shouping Z, et al. (2010). Phylogenetic characterization of Newcastle disease virus isolated in the mainland of China during 2001-2009. *Vet. Microbiol.* 141, 246–57. doi:10.1016/j.vetmic.2009.09.020.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 22, 557-567

Samal S, Anderson D and Wang L. (2011). The Biology of Paramyxoviruses.

Samal S, Khattar S, Kumar S, Collins P and Samal, S. K. (2012). Coordinate deletion of N-glycans from the heptad repeats of the fusion F protein of Newcastle disease virus yields a hyperfusogenic virus with increased replication, virulence, and immunogenicity. *J. Virol.* 86, 2501–11. doi:10.1128/JVI.06380-11.

Sanjuán R, Nebot M, Chirico N, Mansky L and Belshaw R. (2010). Viral Mutation Rates. *Joural of Virology.* 84, 9733–9748. DOI:10.1128/JVI.00694-10.

Seal B. (2004). Nucleotide and predicted amino acid sequence analysis of the fusion protein and hemagglutinin-neuraminidase protein genes among Newcastle disease virus isolates. Phylogenetic relationships among the Paramyxovirinae based on attachment glycoprotein sequenc. *Funct. Integr. Genomics* 4, 246–57. doi:10.1007/s10142-004-0113-2.

Schaefer C, & Rost B. (2012). Predict impact of single amino acid change upon protein structure. *BMC Genomics.* 13(Suppl 4):S4 DOI:10.1186/1471-2164-13-s4

Schadt EE, Linderman MD, Sorenson J, Lee L, & Nolan, GP. (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews. Genetics*, *11*(9), 647–657. DOI:10.1038/nrg2857.

Schirmer M, Sloan WT, Quince C. (2014). Benchmarking of viral haplotype reconstruction programmed: an overview capacities and limitations of currently available programmes. *Brief Bioinform.* 15(3):431-42. DOI: 10.1093/bib/bbs081.

Schmieder R, Edwards R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864. doi:10.1093/ bioinformatics/btr026

Scholtissek C, von Hoyningen V, Rott R. (1978). Genetic relatedness between the new 1977 epidemic strains (H1N1) of influenza and human influenza strains isolated between 1847 and 1957 (H1N1). *Virol* 89 (2): 613H17.

Schönherz A, Lorenzen N, Guldbrandtsen B, Buitenhuis B, Einer-Jensen K. (2016). Ultra-deep sequencing of VHSV isolates contributes to understanding the role of viral quasispecies. *Veterinary Research*. *47*,10. DOI:10.1186/s13567-015-0298-5.

Sentandreu V, Jiménez-Hernández N, Torres-Puente M, Bracho MA, Valero A, Gosalbes MJ, Ortega E, Moya A, González-Candelas F. (2008). Evidence of recombination in intrapatient populations of hepatitis C virus. *PLoS One*. 3: e3239. DOI:10.1371/journal.pone.0003239.

Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler

DA, Gibbs RA, Yu F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research*, *20*(2), 273–280. DOI:10.1101/gr.096388.109.

Shtyrya YA, Mochalova LV, Bovin NV. (2009). Influenza virus neuraminidase: structure and function. Acta Naturae. 1(2):26–32.

Simon-Loriere E, Holmes E. (2011) Why do RNA viruses recombine?. *Nat Rev Microbiol.* 9,617–26. DOI: 10.1038/nrmicro2614 PMID: 21725337.

Skalsky R, Cullen B. (2010) Viruses, microRNAs, and Host Interactions. *Annu Rev Microbiol*. 64, 123–141. DOI:10.1146/annurev.micro.112408.134243.

Skums P, et al. (2012). Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* 13 Suppl 10, S6.

Smith GJD, Bahl J, Vijaykrishna D, Zhang J, Poon LLM, Chen H, Robert G. Webster RG, Peiris M, Yi Guan. (2009b) Dating the emergence of pandemic influenza viruses. *Natl Acad Sci*. 106: 11709-12. DOI: 10.1073/pnas.0904991106.

Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma S, Cheung CL, Raghwani J, Bhatt S, Peris JSM, Guan Y & Rambaut A. (2009a) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459, 1122–5. DOI:10.1038/nature08182.

Smith TF, Waterman MS (1981). Identification of common molecular subsequences. *J Mol Biol.* 147 (1): 195-7.

Simonsen L, Viboud C, Grenfell B, Dushoff J, Jennings L, Smit M, Macken C, Hata M, Gog J, Miller M, Holmes E. (2007). The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. *Mol Biol Evol*. 24:1811–1820.

Sim S, Aw P, Wilm A, Teoh G, Thi Hue K, Nguyen N, Nagarajan N, Simmons C, Hibberd M. (2015). Tracking Dengue Virus Intra-host Genetic Diversity during Human-to-Mosquito Transmission. *PLoS Negl Trop Dis* 9(9): e0004052.

Singh K, Kaur R, Qiu W. (2012). New virus discovery by deep sequencing of small RNAs. *Methods Mol Biol*. 883, 177-91. DOI: 10.1007/978-1-61779-839-9_14.

Snoeck C, Owoade A, Couacy-Hymann E, Alkali B, Okwen M, Adeyanju A, et al. (2013). High genetic diversity of Newcastle disease virus in poultry in West and Central Africa: cocirculation of genotype XIV and newly defined genotypes XVII and XVIII. *J. Clin. Microbiol.* 51, 2250–60. doi:10.1128/JCM.00684-13.

Snyder M, Du J, & Gerstein M. (2010). Personal genome sequencing: current approaches and challenges. *Genes & Development*, *24*(5), 423–431. DOI:10.1101/gab.1864110.

Solmone M, Vincenti D, Prosperi M, Bruselles A, Ippolito G, Capobianchi M. (2009) Use of Massively Parallel Ultradeep Pyrosequencing To Characterize the Genetic Diversity of Hepatitis B Virus in Drug-Resistant and Drug-Naive Patients and To Detect Minor Variants in Reverse Transcriptase and Hepatitis B S Antigen. Journal of Virology. 1718–1726. DOI:10.1128/JVI.02011-08.

Soyeon Ahn & Haris Vikalo. (2017). aBayesQR: A Bayesian method for reconstruction of viral populations characterized by low diversity. *BioRxiv*. DOI: 10.1101/103630.

Suchard M, Weiss R and Sinsheimer, J. S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18, 1001–13.

Susta L, Miller P, Afonso C and Brown, C. C. (2011). Clinicopathological characterization in poultry of three strains of Newcastle disease virus isolated from recent outbreaks. *Vet. Pathol.* 48, 349–60. doi:10.1177/0300985810375806.

Svarovskaia E, Martin R, McHutchison J, Miller M, Mo H. (2012). Abundant Drug-Resistant NS3 Mutants Detected by Deep Sequencing in Hepatitis C Virus-Infected Patients Undergoing NS3 Protease Inhibitor Monotherapy. *J Clin Microbiol*. 50(10):3267–3274. DOI:10.1128/JCM.00838-12

Takashita E, Meijer A, Lackenby A, et al. (2015). Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2013-2014. *Antiviral Res*. 117:27–38

Tamura K., Nei M., and Kumar S. (2004). Prospects for inferring very large phylogenies by using neighbor-joining method. *Proceedings of the National Academy of Sciences.* 101, 11030-11035.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–9. doi:10.1093/molbev/msr121.

Taubenberger JK. (2006). Influenza hemaglutinin attachment to target cell: 'birds to do it, we do it'. *Future Virol* 1:415H418.

Thomazelli L, Araujo J, Oliveira D, Sanfilippo L, Ferreira C, Brentano L, et al. (2010). Newcastle disease virus in penguins from King George Island on the Antarctic region. *Vet. Microbiol.* 146, 155–60. DOI: 10.1016/j.vetmic.2010.05.006.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and Pachter L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, *28*(5), 511–515. DOI: 10.1038/nbt.1621.

Triverdi UH, Cézard T, Bridgett S, Montazam A, Nichols J, Blaxter M, & Gharbi K. (2014). Quality control of next-generation sequencing data without a reference. *Frontiers in Genetics*, *5*, 111. DOI: 10.3389/fgene.2014.00111

Tong S, Zhu X, Li Y, Shi M., Zhang J, Bourgeois M, Yang H, Chen X, Recuenco S, Gomez J, Chen LM, Johnson A, Tao Y, Dreyfus C, Yu W, McBride R, Carney PJ, Gilbert AT, Chang J, Guo Z, Davis CT, Paulson JC, Stevens J, Rupprecht CE, Holmes EC, Wilson IA, Donis RO. (2013) New world bats harbor diverse influenza a viruses. PLoS Pathog. 9:e1003657. DOI: 10.1371/journal.ppat.1003657.

Töpfer A, Marschall T, Bull RA, Luciani F, Schnhuth A, & Beerenwinkel N. (2014). Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput Biol*, 10(3), e1003515.

Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, Beerenwinkel N. (2013). Probabilistic inference of viral quasispecies subject to recombination. *Journal of Computational Biology*, 20(2), 113-123.

Toyoda T, Gotoh B, Sakaguchi T, Kida H and Nagai Y. (1988). Identification of amino acids relevant to three antigenic determinants on the fusion protein of Newcastle disease virus that are involved in fusion inhibition and neutralization. *J. Virol.* 62, 4427–30.

Varble A, Albrecht R, Backes S, Crumiller M, Bouvier N, Sachs D, García-Sastre A, tenOever B. (2014). Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe*. 12, 691–700. DOI:10.1016/j.chom.2014.09.020.

Verbinnen T, Van Marck H, Vandenbroucke I, Vijgen L, Claes M, Lin T-I, Simmen K, Neyts J, Fanning G, Lenz O. (2010). Tracking the Evolution of Multiple *In Vitro* Hepatitis C Virus Replicon Variants under Protease Inhibitor Selection Pressure by 454 Deep Sequencing. *Journal of Virology*, *84*(21), 11124–11133. DOI: 10.1128/JVI.01217-10

Vignuzzi M, Stone J, Andino R. (2005) Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications. *Virus Research* 107, 173-181.

Vignuzzi M, Stone J, Arnold J, Cameron C, Andino R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature.* 439, 344-8.

Wakamatsu N, King D, Seal B, Peeters B and Brown C. (2006). The effect on pathogenesis of Newcastle disease virus LaSota strain from a mutation of the fusion cleavage site to a virulent sequence. *Avian Dis.* 50, 483–8. doi:10.1637/7515-020706R.1.

Wan H, Chen L, Wu L and Liu X. (2004). Newcastle disease in geese: natural occurrence and experimental infection. *Avian Pathol.* 33, 216–21. doi:10.1080/0307945042000195803.

Wang ZO and Pollock DD. (2005). Context Dependence and Coevolution Among Amino Acid Residues in Proteins. *Methods Enzymol*. 395: 779–790.

Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, & Shafer RW. (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome research* 17, 1195–1201.

Wang S, Li W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012 Aug 15;28(16):2184-5. DOI:10.1093/bioinformatics/bts356.

Watson SJ, Welkers MRA, Depledge DP, Coulter E, Breuer JM, de Jong MD, Kellam P. (2013). Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1614), 20120205. DOI:10.1098/rstb.2012.0205.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. (1992) Evolution and ecology of influenza-A viruses. *Microbiol Rev* 56:152–179.

Wei Z, Wang W, Hu P, Lyon G J, & Hakonarson H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*, *39*(19), e132. DOI:10.1093/nar/gkr599

Weikard R, Hadlich F, & Kuehn C. (2013). Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing.*BMC Genomics*,*14*, 789. DOI:10.1186/1471-2164-14-789.
Werner O, Römer-Oberdörfer A, Köllner B, Manvell R and Alexander D (1999). Characterization of avian paramyxovirus type 1 strains isolated in Germany during 1992 to 1996. *Avian Pathol.* 28, 79–88. doi:10.1080/03079459995082.

Westbrooks K.et al. (2008) HCV quasispecies assembly using network flows. Lect N Bioinformat 4983, 159–170.

Westgeest KB, Russell CA, Lin X, Spronken MI, Bestebroer TM, Bahl J, et al. (2014). Genomewide analysis of reassortment and evolution of human influenza A(H3N2) viruses circulating between 1968 and 2011. J Virol. 88(5):2844–57. doi:10.1128/JVI.02163-13.

Wilke C. (2005). Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology*, *5*, 44. DOI:10.1186/1471-2148-5-44.

Wilk E, Pandey A, Leist S, Hatesuer B, Preusse M, Pommerenke C, Wang J, Schughart K. (2015). RNAseq expression analysis of resistant and susceptible mice after influenza A virus infection identifies novel genes associated with virus replication and important for host resistance to infection. *BMC Genomics*, *16*, 655. DOI:10.1186/s12864-015-1867-8.

Worobey M, Holmes E. (1999). Evolutionary aspects of recombination in RNA viruses. *J Gen Virol*. 10, 2553-43. DOI:10.99/0022-1317-80-10-2535.

Wright C, Morelli M, Thébaud G, Knowles N, Herzyk P, Paton D, Haydon D, King D. (2011). Beyond the Consensus: Dissecting Within-Host Viral Population Diversity of Foot and-Mouth Disease Virus by Using Next-Generation Genome Sequencing. *Journal of Virology*. 2266-2275.

Wright S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution.

The Sixth International Congress of Genetics, I, 356-66.

Wu L, Zhang X, Zhao Z, Wang L, Li B, Li G, Dean M, Yu Q, Wang Y, Lin X, Rao W, Mei Z, Li Y, Jiang R, Yang H, Li F, Xie G, Xu L, Wu K, Zhang J, Chen J, Wang T, Kristiansen K, Zhang X, Li Yingrui, Wang J, Hou Y. (2015). Full-length single-cell RNA-seq applied to a viral human cancer: applications to HPV expression and splicing analysis in HeLa S3 cells. *GigaScience*, *4*, 51. DOI:10.1186/s13742-015-0091-4.

Xiao S, Paldurai A, Nayak B, Mirande A, Collins P and Samal SK. (2013). Complete genome sequence of a highly virulent newcastle disease virus currently circulating in Mexico. *Genome Announc.* 1. doi:10.1128/genomeA.00177-12.

Yang C, Shieh H, Lin Y and Chang P. (1999). Newcastle disease virus isolated from recent outbreaks in Taiwan phylogenetically related to viruses (genotype VII) from recent outbreaks in western Europe. *Avian Dis.* 43, 125–30.

Yen H-L, Hoffmann E, Taylor G, Scholtissek C, Monto AS, Webster RG, & Govorkova EA. (2006). Importance of Neuraminidase Active-Site Residues to the Neuraminidase Inhibitor Resistance of Influenza Viruses. *Journal of Virology,* 80(17), 8787–8795. DOI:10.1128/JVI.00477-06.

Yun Z, Lara C, Johansson B, Lorenzana de Rivera I, Sönnerborg A. (1996) Discrepancy of hepatitis C virus genotypes as determined by phylogenetic analysis of partial NS5 and core sequences. *J Med Virol.* 49: 155-160. DOI:10.1002/(SICI)1096-9071(199607)49:3<155::AID-JMV1>3.0.CO;2-3

Yusoff K, Nesbit M, McCartney H, Meulemans G, Alexander D, Collins M, et al. (1989). Location of neutralizing epitopes on the fusion protein of Newcastle disease virus strain Beaudette C. *J. Gen. Virol.*, 3105–9. doi:10.1099/0022-1317-70-11-3105.

Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12, 119.

Zagordi O, Däumer M, Beisel C, Beerenwinkel N. (2012). Read length versus depth of coverage for viral quasispecies reconstruction. PLOS ONE 7: e47046.

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. (2010a). Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 17, 417–428.

Zagordi O, Klein R, Daumer M, & Beerenwinkel N. (2010b). Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 38, 7400–7409.

Zerbino DR, & Birney E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. DOI: 10.1101/gr.074492.107

Zhang S, Wang X, Zhao C, Liu D, Hu Y, Zhao J, et al. (2011). Phylogenetic and pathotypical analysis of two virulent Newcastle disease viruses isolated from domestic ducks in China. *PLoS One* 6, e25000. doi:10.1371/journal.pone.0025000.

Zhou Q, Su X, Wang A, Xu J, Ning K. (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. PLoS ONE:e60234. DOI: 10.1371/journal.pone.0060234

Anexo I

Contenido:

Tabla 1. Características de las plataformas de secuenciación de nueva generación.

Plataforma	Rendimiento Máximo Mb/corrida	Longitud Media (nt)	Tasa de error	Aplicaciones	Principal fuente de error
454 FLX	700	~800 (experimentos shotgun) ~400 (experimentos con amplicones)	10^{-3}-10^{-4}	Secuenciado de novo y resecuenciado, resecuenciado de blancos, genotipificación, metagenómica.	Corte de intensidad, homopolímeros, interferencia de señal cruzada entre vecinos, amplificación, beads mixtos.
Illumina	6000	~100	10^{-2}-10^{-3}	Resecuenciado de genomas, transcriptómica cuantitativa, genotipificación, metagenómica.	Interferencia de señal en clusters vecinos, phasing, homopolímeros, etiquetado de nucleótidos, amplificación, baja cobertura de regiones ricas en AT.
SOLiD	20000	~50	10^{-2}-10^{-3}	Resecuenciado de genomas, transcriptómica cuantitativa, genotipificación.	Interferencia de señal en clusters vecinos, eliminación gradual, marcaje de nucleótidos, degradación de la señal, beads mixtos, baja cobertura de regiones ricas en AT.
Helicos	21000-35000	~35	10^{-2}	Muestras no amplificabes, libre de PCR y análisis cuantitativos no sesgados.	Polimerada empleada, pérdida de moléculas, intensidades bajas.
Ion Torrent PGM	1000	~200	$3x10^{-2}$	Secuenciamiento de genomas de novo, resecuenciamiento de target, genotipificación, RNA-seq de transcriptomas de baja complejidad, metagenómica.	Homopolimeros, amplificación.
GS Junior	~35	~400	10^{-3}-10^{-4}	Resecuenciamiento de target (amplicones), genotipado.	Corte de intensidad, homopolímeros, interferencia de señal cruzada entre vecinos, amplificación, beads mixtos.

* Tasa de error considerando únicamente sustituciones, no indels.
Tomado de Barzon y cols., 2011 (Applications of Nex-Generation Sequencing Technologies to Diagnostic Virology Barzon et al. 2011).

Tabla 2. Características clínicas y moleculares de los pacientes analizados. EPOC: Enfermedad Pulmonar Obstructiva Crónica. IR: Insuficiencia Respiratoria. IRA: Insuficiencia Respiratoria Aguda. IRAG: Insuficiencia Respiratoria Aguda Grave. NAC: Neumonía Aguda Comunitaria. S/D: Sin dato.

Id.	Toma de muestra	Localidad	EDAD	SEXO	M2	H1N1	DATO CLÍNICO	MUESTRA
37	07/04/11	Mdeo	57a	M	Positivo	Positivo	Probable neumonia H1N1	Hisopo nasal
45	07/08/11	Mdeo	RN	F	Positivo	Positivo	Recien Nacido	Aspirado Nasofaringeo
48	07/11/11	Canelones	28a	F	Positivo	Positivo	S/D	Hisopo nasal
56	07/13/11	Mdeo	49a	M	Positivo	Positivo	S/D	Hisopo nasal
81	07/25/11	Mdeo	51a	F	Positivo	Negativo	IR	Hisopo nasal
88	07/28/11	Canelones	38a	M	Positivo	Positivo	Broncoespasmo, fiebre, tos, mialgia, odinofagia	Hisopo nasal
90	07/29/11	Mdeo	70a	M	Positivo	Positivo	S/D	Hisopo nasal
92	07/29/11	Mdeo	23a	F	Positivo	Positivo	Neumonia	Hisopo nasal
96	07/29/11	Mdeo	42a	F	Positivo	Positivo	S/D	Hisopo nasal
98	08/01/11	Mdeo	80a	M	Positivo	Positivo	EPOC, bronqueoectasias, fiebre, IR	Hisopo nasal
103	08/04/11	Mdeo	28a	F	Positivo	Positivo	IRA, probable H1N1	Aspirado Nasofaringeo
108	08/05/11	Maldonado	44a	F	Positivo	Positivo	Sind. Dawn, fiebre, tos, IR, compromiso hemodinamico	Aspirado Nasofaringeo
112	08/08/11	Mdeo	63a	M	Positivo	Negativo	Diabetes, obesidad, tos, mialgia, odinofagia, Cefalea, IR	Hisopo nasal
113	08/08/11	Mdeo	5m	M	Positivo	Positivo	Tos, IR, compromiso hemodinamico	Aspirado Nasofaringeo
116	08/09/11	Mdeo	75a	M	Positivo	Negativo	NAC	Hisopo nasal
117	08/09/11	Mdeo	35a	F	Positivo	Positivo	S/D	Hisopo nasal
118	08/09/11	Mdeo	28a	F	Positivo	Positivo	Asma, tos	Hisopo nasal
132	08/15/11	Mdeo	70a	M	Positivo	Negativo	Mialgia, cefalea, IRA	Hisopo nasal
141	08/17/11	Mdeo	62a	M	Positivo	Positivo	S/D	Hisopo nasal
154	08/24/11	Mdeo	62a	F	Positivo	Positivo	IR	Hisopo nasal
159	08/26/11	Mdeo	55a	F	Positivo	Negativo	Obesidad, inmunodepresion, IR, fiebre, tos, mialgias	Hisopo nasal
202	06/07/12	Mdeo	78a	M	Positivo	Negativo	Tos, IRA, fiebre	Hisopo nasal
206	10/07/12	Canelones	1a	S/D	Positivo	Negativo	Descartar Influenza	Aspirado Nasofaringeo
216	23/07/12	Mdeo	82a	M	Positivo	Negativo	NAC severa, tto	Hisopo nasal
220	25/07/12	Mdeo	79a	S/D	Positivo	Negativo	NAC	Hisopo nasal
224	31/07/12	Rocha	67a	F	Positivo	Negativo	Cuadro de gripe, diabetes, neumonitis	Hisopo nasal
250	05/28/13	S/D	S/D	M	Positivo	Negativo	S/D	Hisopo nasal
264	07/10/13	S/D	S/D	M	Positivo	Negativo	S/D	Hisopo nasal

Tabla 3. Conteo de reads crudos y trimados. Resumen de los resultados de la secuenciación, cantidad de lecturas crudas y procesadas. El rendimiento indica el número de reads filtrados por calidad sobre el número de reads crudos. La cantidad de lecturas es indicado en pares de reads.

Muestra	Reads_crudos	trimm_adapt	trimm_primer	trimm_qual	Rendimiento
m116	1602212	1602212	1600372	1109016	69.22
m132	2119546	2119546	2115780	1672112	78.89
m159	2273642	2273642	2268220	1656173	72.84
m202	2643182	2643182	2637242	1979200	74.88
m206	2452824	2452824	2448632	1928386	78.62
m220	6061352	6061352	6049928	4583605	75.62
m224	2409792	2409792	2406036	1925255	79.89
m250	6952326	6952326	6937540	5072990	72.97
m264	2049174	2049174	2044410	1569365	76.59

Tabla 4A. Resultado Alineadores.

Muestra	Gen	Método	total_reads	paired	single	reads_aln	aln_rate
m116	HA	Bowtie2	626706	482310	144396	384421	61.34
	NA					210198	33.54
	HA	Bowtie				322495	51.46
	NA					176529	28.17
	HA	BWA	1109853	965380	144473	710569	64.02
	NA		1109222	964788	144434	380138	34.27
m132	HA	Bowtie2	898078	774034	124044	560310	62.39
	NA					297803	33.16
	HA	Bowtie				476386	53.05
	NA					248594	27.68
	HA	BWA	1674351	1550152	124199	1097770	65.56
	NA		1672584	1548503	124081	572334	34.22
m159	HA	Bowtie2	921396	734777	186619	531553	57.69
	NA					337230	36.60
	HA	Bowtie				412736	44.79
	NA					264581	28.72
	HA	BWA	1657389	1470634	186755	1030437	62.17
	NA		1656779	1470047	186732	623390	37.63

m202	HA	Bowtie2	1077406	901794	175612	727033	67.48
	NA					285835	26.53
	HA	Bowtie				579994	53.83
	NA					176028	16.34
	HA	BWA	1980243	1804511	175732	1431611	72.29
	NA		1980127	1804404	175723	546836	27.62
m206	HA	Bowtie2	1030737	897649	133088	676988	65.68
	NA					295100	28.63
	HA	Bowtie				561292	54.46
	NA					190282	18.46
	HA	BWA	1929476	1796258	133218	1349732	69.95
	NA		1929133	1795957	133176	574281	29.77
m220	HA	Bowtie2	2447112	2136493	310619	1404887	57.41
	NA					901516	36.84
	HA	Bowtie				1004108	41.03
	NA					600129	24.52
	HA	BWA	4585338	4274484	310854	2818728	61.47
	NA		4586405	4275533	310872	1764912	38.48

m224	HA	Bowtie2	1023941	901314	122627	552211	53.93
	NA					416129	40.64
	HA	Bowtie				457287	44.66
	NA					321408	31.39
	HA	BWA	1926579	1803842	122737	1105768	57.39
	NA		1927117	1804305	122812	819961	42.55
m250	HA	Bowtie2	2739013	2333977	405036	1606705	58.66
	NA					979471	35.76
	HA	Bowtie				1163698	42.49
	NA					730707	26.68
	HA	BWA	5075978	4670742	405236	3060184	60.29
	NA		5074826	4669614	405212	1895035	37.34
m264	HA	Bowtie2	834019	735346	98673	580310	69.58
	NA					208336	25.07
	HA	Bowtie				453873	54.42
	NA					154113	18.48
	HA	BWA	1570301	1570301	98744	1132874	72.14
	NA		1569711	1471020	98691	404163	25.75

Tabla 4B. Resumen general Alineadores. Se puede ver el porcentaje de lecturas totales que alinean contra cada gen de referencia.

Gen	Alineadores	Reads_de_calidad	Aln_count	Aln_%
HA	Bowtie	11598408	5431869	46.83
	Bowtie2	11598408	7024418	60.56
	BWA	21509508	13737673	63.87
NA	Bowtie	11598408	2862371	24.68
	Bowtie2	11598408	3931618	33.89
	BWA	21505904	7581050	35.25

Tabla 5. Resultado Ensambladores.

Muestra	Gen	Método	#Contig	contig_length	N50	%GC	Ref_len	Contig/Ref
m116	HA	Trinity	3	1775	1767	42.12	1584	1.1206
	NA		1	1377	1377	43.28	1262	1.0911
	HA	SPAdes	2	866	866	42.14	1584	0.5467
	NA		1	1415	1415	43.18	1262	1.1212
m132	HA	Trinity	1	1781	1781	42.06	1566	1.1373
	NA		1	1396	1396	43.19	1256	1.1115
	HA	SPAdes	2	806	806	42.16	1566	0.5147
	NA		2	677	677	43.37	1256	0.5390
m159	HA	Trinity	1	1780	1780	42.08	1566	1.1367
	NA		1	1409	1409	43.65	1256	1.1218
	HA	SPAdes	2	826	826	42.57	1566	0.5275
	NA		2	610	610	43.66	1256	0.4857
m202	HA	Trinity	1	1780	1780	41.91	1566	1.1367
	NA		1	1361	1361	43.58	1233	1.1038
	HA	SPAdes	1	874	874	41.88	1566	0.5581
	NA		2	847	847	43.59	1233	0.6869

m206	HA	Trinity	1	1780	1780	41.91	1566	1.1367
	NA		1	1660	1660	43.80	1233	1.3463
	HA	SPAdes	2	857	857	42.09	1566	0.5473
	NA		1	1361	1361	43.50	1233	1.1038
m220	HA	Trinity	2	1772	1772	43.20	1566	1.1315
	NA		1	1383	1383	43.60	1233	1.1217
	HA	SPAdes	2	911	911	41.89	1566	0.5817
	NA		2	774	774	43.43	1233	0.6277
m224	HA	Trinity	1	1770	1770	41.65	1566	1.1303
	NA		1	1379	1379	43.36	1233	1.1184
	HA	SPAdes	2	849	849	41.60	1566	0.5421
	NA		2	758	758	43.14	1233	0.6148
m250	HA	Trinity	2	1447	1447	41.91	1611	0.8982
	NA		1	1353	1353	43.61	1239	1.0920
	HA	SPAdes	2	624	624	41.47	1611	0.3873
	NA		2	568	568	43.43	1239	0.4584
m264	HA	Trinity	1	1779	1779	41.77	1602	1.1105
	NA		1	1340	1340	43.06	1264	1.0601
	HA	SPAdes	2	986	986	41.63	1602	0.6155
	NA		1	610	610	43.28	1264	0.4826

Tabla 6. Cambios AAs. SNPs, frecuencias, cambios sinónimos y no sinónimos. Para HA y NA.
Variantes NA:

Muestra 116						Muestra 132						Muestra 159					
Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%	Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%	Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%
'ATG'	23	0	0	0													
'CAA'	24	0	0	0													
'ATT'	25	0	0	0		'ATT'	25	0	0	0		'ATT'	25	0	0	0	
'GCC'	26	0	0	0		'GCC'	26	0	0	0		'GCC'	26	0	0	0	
'ATC'	27	0	0	0		'ATC'	27	0	0	0		'ATC'	27	0	0	0	
'TTG'	28	0	0	0		'TTG'	28	0	0	0		'TTG'	28	0	0	0	
'ATA'	29	0	0	0		'ATA'	29	0	0	0		'ATA'	29	0	0	0	
'ACT'	30	0	0	0		'ACT'	30	0	0	0		'ACT'	30	0	0	0	
'ACT'	31	0	0	0		'ACT'	31	0	0	0		'ACT'	31	0	0	0	
'GTA'	32	0	0	0		'GTA'	32	GTG	V	Sin	0.78	'GTA'	32	0	0	0	
'ACA'	33	0	0	0		'ACA'	33	0	0	0		'ACA'	33	0	0	0	
'TTG'	34	CTG	L	Sin	0.57	'TTG'	34	CTG	L	Sin	0.87	'TTG'	34	0	0	0	
'CAT'	35	0	0	0		'CAT'	35	0	0	0		'CAT'	35	0	0	0	
'TTC'	36	0	0	0		'TTC'	36	0	0	0		'TTC'	36	0	0	0	

'AAG'	37	0	0	0		'AAG'	37	GAG	E	NoSin	0.53	'AAG'	37	0	0	0	
'CAA'	38	0	0	0		'CAA'	38	GAA	E	NoSin	0.5	'CAA'	38	0	0	0	
'TAT'	39	0	0	0		'TAT'	39	0	0	0		'TAT'	39	0	0	0	
'GAA'	40	0	0	0		'GAA'	40	0	0	0		'GAA'	40	0	0	0	
'TTC'	41	0	0	0		'TTC'	41	0	0	0		'TTC'	41	0	0	0	
'AAC'	42	0	0	0		'AAC'	42	GAC/AGC	D/S	NoSin	0.42/0.43	'AAC'	42	0	0	0	
'TCC'	43	CCC	P	NoSin	0.39	'TCC'	43	TCT	S	Sin	0.75	'TCC'	43	0	0	0	
'CCC'	44	0	0	0		'CCC'	44	0	0	0		'CCC'	44	0	0	0	
'CCA'	45	0	0	0		'CCA'	45	0	0	0		'CCA'	45	0	0	0	
'AAC'	46	0	0	0		'AAC'	46	0	0	0		'AAC'	46	0	0	0	
'AAC'	47	AGC	S	NoSin	0.36	'AAC'	47	0	0	0		'AAC'	47	0	0	0	
'CAA'	48	0	0	0		'CAA'	48	0	0	0		'CAA'	48	0	0	0	
'GTG'	49	0	0	0		'GTG'	49	GCG	A	NoSin	0.38	'GTG'	49	0	0	0	
'ATG'	50	ACG	T	NoSin	0.35	'ATG'	50	GTG	V	NoSin	0.35	'ATG'	50	0	0	0	
'CTG'	51	CCG	P	NoSin	0.47	'CTG'	51	0	0	0		'CTG'	51	CCG	P	NoSin	0.5
'TGT'	52	0	0	0		'TGT'	52	TGC	C	Sin	0.37	'TGT'	52	CGT	R	NoSin	0.36
'GAA'	53	GGA/GAG	G/E	NoSin/Sin	0.41/0.33	'GAA'	53	GGA/GAG	G/E	NoSin/Sin	0.38/0.32	'GAA'	53	0	0	0	
'CCA'	54	CCG	P	Sin	0.37	'CCA'	54	CTA	L	NoSin	0.36	'CCA'	54	0	0	0	
'ACA'	55	0	0	0		'ACA'	55	0	0	0		'ACA'	55	0	0	0	
'ATA'	56	0	0	0		'ATA'	56	0	0	0		'ATA'	56	0	0	0	
'ATA'	57	GTA	V	NoSin	0.29	'ATA'	57	0	0	0		'ATA'	57	0	0	0	

'GAA'	58	0	0	0		'GAA'	58	0	0	0		'GAA'	58	0	0	0	
'AGA'	59	AGG	R	Sin	0.33	'AGA'	59	0	0	0		'AGA'	59	0	0	0	
'AAC'	60	AGC	S	NoSin	0.41	'AAC'	60	AGC	S	NoSin	0.7	'AAC'	60	AGC	S	NoSin	0.43
'ATA'	61	0	0	0		'ATA'	61	0	0	0		'ATA'	61	0	0	0	
'ACA'	62	GCA	A	NoSin	0.32	'ACA'	62	GCA	A	NoSin	0.39	'ACA'	62	GCA	A	NoSin	0.36
'GAG'	63	GGG	G	NoSin	0.29	'GAG'	63	0	0	0		'GAG'	63	GGG	G	NoSin	0.33
'ATA'	64	0	0	0		'ATA'	64	GTA	V	NoSin	0.32	'ATA'	64	GTA	V	NoSin	0.35
'GTG'	65	0	0	0		'GTG'	65	0	0	0		'GTG'	65	0	0	0	
'TAT'	66	0	0	0		'TAT'	66	CAT	H	NoSin	0.27	'TAT'	66	0	0	0	
'CTG'	67	0	0	0		'CTG'	67	0	0	0		'CTG'	67	CCG	P	NoSin	0.28
'ACC'	68	0	0	0		'ACC'	68	GCC	A	NoSin	0.26	'ACC'	68	GCC	A	NoSin	0.31
'AAC'	69	0	0	0		'AAC'	69	0	0	0		'AAC'	69	AGC	S	NoSin	0.26
'ACC'	70	GCC	A	NoSin	0.32	'ACC'	70	GCC	A	NoSin	0.25	'ACC'	70	GCC	A	NoSin	0.29
'ACC'	71	0	0	0		'ACC'	71	0	0	0		'ACC'	71	0	0	0	
'ATA'	72	0	0	0		'ATA'	72	0	0	0		'ATA'	72	ACA	T	NoSin	0.58
'GAG'	73	0	0	0		'GAG'	73	0	0	0		'GAG'	73	0	0	0	
'AAG'	74	AGG	R	NoSin	0.26	'AAG'	74	0	0	0		'AAG'	74	AGG	R	NoSin	0.29
'GAA'	75	0	0	0		'GAA'	75	0	0	0		'GAA'	75	0	0	0	
'ATA'	76	0	0	0		'ATA'	76	GTA/ AAA	V/K	NoSin	0.63/0. 35	'ATA'	76	0	0	0	
'TGC'	77	0	0	0		'TGC'	77	TGA	STOP	0	0.26	'TGC'	77	0	0	0	
'CCC'	78	0	0	0		'CCC'	78	0	0	0		'CCC'	78	0	0	0	
'AAA'	79	0	0	0		'AAA'	79	0	0	0		'AAA'	79	0	0	0	

'CTA'	80	0	0	0		'CTA'	80	CCA	P	NoSin	0.4	'CTA'	80	0	0	0	
'GCA'	81	0	0	0		'GCA'	81	GCG	A	Sin	0.24	'GCA'	81	0	0	0	
'GAA'	82	0	0	0		'GAA'	82	0	0	0		'GAA'	82	0	0	0	
'TAC'	83	TGC	C	NoSin	0.29	'TAC'	83	0	0	0		'TAC'	83	0	0	0	
'AGA'	84	GGA	G	NoSin	0.48	'AGA'	84	GGA	G	NoSin	0.25	'AGA'	84	GGA	G	NoSin	0.32
'AAT'	85	0	0	0		'AAT'	85	0	0	0		'AAT'	85	0	0	0	
'TGG'	86	0	0	0		'TGG'	86	0	0	0		'TGG'	86	0	0	0	
'TCA'	87	CCA	P	NoSin	0.44	'TCA'	87	CCA/ TCG	P/S	NoSin /Sin	0.36/0. 39	'TCA'	87	CCA	P	NoSin	0.37
'AAG'	88	0	0	0		'AAG'	88	AGG	R	NoSin	0.24	'AAG'	88	AGG	R	NoSin	0.43
'CCG'	89	0	0	0		'CCG'	89	0	0	0		'CCG'	89	0	0	0	
'CAA'	90	0	0	0		'CAA'	90	0	0	0		'CAA'	90	0	0	0	
'TGT'	91	TGC	C	Sin	0.4	'TGT'	91	TGC	C	Sin	0.31	'TGT'	91	TGC	C	Sin	0.37
'GAC'	92	GGC	G	NoSin	0.33	'GAC'	92	GGC	G	NoSin	0.56	'GAC'	92	GGC	G	NoSin	0.43
'ATT'	93	0	0	0		'ATT'	93	0	0	0		'ATT'	93	0	0	0	
'ACA'	94	ACG	T	Sin	0.39	'ACA'	94	ACG	T	Sin	0.46	'ACA'	94	ACG	T	Sin	0.31
'GGA'	95	0	0	0		'GGA'	95	0	0	0		'GGA'	95	0	0	0	
'TTT'	96	0	0	0		'TTT'	96	0	0	0		'TTT'	96	0	0	0	
'GCA'	97	0	0	0		'GCA'	97	GCG	A	Sin	0.25	'GCA'	97	0	0	0	
'CCT'	98	CCC	P	Sin	0.39	'CCT'	98	CCC	P	Sin	0.41	'CCT'	98	CCC	P	Sin	0.42
'TTT'	99	0	0	0		'TTT'	99	0	0	0		'TTT'	99	0	0	0	
'TCT'	100	0	0	0		'TCT'	100	0	0	0		'TCT'	100	0	0	0	
'AAG'	101	0	0	0		'AAG'	101	AGG	R	NoSin	0.45	'AAG'	101	AGG	R	NoSin	0.32

'GAC'	102	GGC	G	NoSin	0.82	'GAC'	102	GGC	G	NoSin	0.34	'GAC'	102	GGC	G	NoSin	0.3
'AAT'	103	AGT	S	NoSin	0.84	'AAT'	103	0	0	0		'AAT'	103	0	0	0	
'TCG'	104	0	0	0		'TCG'	104	0	0	0		'TCG'	104	0	0	0	
'ATT'	105	0	0	0		'ATT'	105	0	0	0		'ATT'	105	0	0	0	
'AGG'	106	0	0	0		'AGG'	106	0	0	0		'AGG'	106	0	0	0	
'CTT'	107	0	0	0		'CTT'	107	0	0	0		'CTT'	107	0	0	0	
'TCC'	108	0	0	0		'TCC'	108	0	0	0		'TCC'	108	0	0	0	
'GCT'	109	0	0	0		'GCT'	109	0	0	0		'GCT'	109	GCC	A	Sin	0.27
'GGT'	110	0	0	0		'GGT'	110	0	0	0		'GGT'	110	0	0	0	
'GGG'	111	0	0	0		'GGG'	111	0	0	0		'GGG'	111	0	0	0	
'GAC'	112	GGC	G	NoSin	0.31	'GAC'	112	AAC/ GGC	N/G	NoSin	0.55/ 0.34	'GAC'	112	GGC	G	NoSin	0.31
'ATC'	113	0	0	0		'ATC'	113	GTC/ ACC	V/T	NoSin	0.25/ 0.34	'ATC'	113	GTC	V	NoSin	0.25
'TGG'	114	0	0	0		'TGG'	114	0	0	0		'TGG'	114	0	0	0	
'GTG'	115	0	0	0		'GTG'	115	GCG	A	NoSin	0.27	'GTG'	115	GCG	A	NoSin	0.26
'ACA'	116	GCA	A	NoSin	0.44	'ACA'	116	GCA	A	NoSin	0.35	'ACA'	116	GCA	A	NoSin	0.26
'AGA'	117	AGG	R	Sin	0.26	'AGA'	117	0	0	0		'AGA'	117	AGG	R	Sin	0.26
'GAA'	118	GGA/ GAG	G/E	NoSin /Sin	0.26/ 1.7	'GAA'	118	GGA/ GAG	G/E	NoSin /Sin	0.47/ 0.48	'GAA'	118	GGA/ GAG	G/E	NoSin /Sin	0.45/ 0.33
'CCT'	119	0	0	0		'CCT'	119	0	0	0		'CCT'	119	CCC	P	Sin	0.25
'TAT'	120	0	0	0		'TAT'	120	0	0	0		'TAT'	120	0	0	0	
'GTG'	121	GCG	A	NoSin	0.47	'GTG'	121	GCG	A	NoSin	0.56	'GTG'	121	GCG	A	NoSin	0.5
'TCA'	122	CCA/	P/S	NoSin	0.34/0.	'TCA'	122	CCA	P	NoSin	0.38	'TCA'	122	CCA/	P/S	NoSin	0.4/

		TCG		/Sin	25								TCG		/Sin	0.29
'TGC'	123	0	0	0		'TGC'	123	0	0	0		'TGC'	123	0	0	0	
'GAT'	124	0	0	0		'GAT'	124	GGT	G	NoSin	0.4	'GAT'	124	GGT	G	NoSin	0.24
'CCT'	125	CCC	P	Sin	0.33	'CCT'	125	CCC	P	Sin	0.33	'CCT'	125	CCC	P	Sin	0.27
'GAC'	126	GGC	G	NoSin	0.42	'GAC'	126	GGC	G	NoSin	0.24	'GAC'	126	GGC	G	NoSin	0.29
'AAG'	127	AGG	R	NoSin	0.35	'AAG'	127	AGG	R	NoSin	0.25	'AAG'	127	0	0	0	
'TGT'	128	0	0	0		'TGT'	128	0	0	0		'TGT'	128	0	0	0	
'TAT'	129	0	0	0		'TAT'	129	0	0	0		'TAT'	129	0	0	0	
'CAA'	130	0	0	0		'CAA'	130	0	0	0		'CAA'	130	0	0	0	
'TTT'	131	0	0	0		'TTT'	131	0	0	0		'TTT'	131	0	0	0	
'GCC'	132	0	0	0		'GCC'	132	0	0	0		'GCC'	132	0	0	0	
'CTT'	133	0	0	0		'CTT'	133	CCT	P	NoSin	0.24	'CTT'	133	0	0	0	
'GGA'	134	GGG	G	Sin	0.34	'GGA'	134	GGG	G	Sin	0.37	'GGA'	134	GGG	G	Sin	0.26
'CAG'	135	CGG	R	NoSin	0.49	'CAG'	135	CGG	R	NoSin	0.39	'CAG'	135	CGG	R	NoSin	0.3
'GGA'	136	GGG	G	Sin	0.3	'GGA'	136	GGG	G	Sin	0.32	'GGA'	136	GGG	G	Sin	0.31
'ACA'	137	GCA	A	NoSin	0.31	'ACA'	137	GCA	A	NoSin	0.28	'ACA'	137	GCA	A	NoSin	0.27
'ACA'	138	ACG	T	Sin	0.32	'ACA'	138	GCA ACG	A/T	NoSin /Sin	0.28/0. 29	'ACA'	138	GCA	A	NoSin	0.28
'CTA'	139	0	0	0		'CTA'	139	0	0	0		'CTA'	139	0	0	0	
'AAC'	140	0	0	0		'AAC'	140	AGC	S	NoSin	0.27	'AAC'	140	AGC	S	NoSin	0.3
'AAC'	141	0	0	0		'AAC'	141	AGC	S	NoSin	0.23	'AAC'	141	0	0	0	
'GTG'	142	0	0	0		'GTG'	142	0	0	0		'GTG'	142	GCG	A	NoSin	0.26
'CAT'	143	CGT	R	NoSin	2.39	'CAT'	143	0	0	0		'CAT'	143	0	0	0	

'TCA'	144	0	0	0		'TCA'	144	0	0	0		'TCA'	144	0	0	0	
'AAT'	145	0	0	0		'AAT'	145	0	0	0		'AAT'	145	0	0	0	
'AAC'	146	0	0	0		'AAC'	146	0	0	0		'AAC'	146	0	0	0	
'ACA'	147	GCA	A	NoSin	0	'ACA'	147	GCA	A	NoSin	0.28	'ACA'	147	GCA	A	NoSin	0.4
'GTA'	148	0	0	0		'GTA'	148	GCA	A	NoSin	0.45	'GTA'	148	0	0	0	
'CGT'	149	0	0	0		'CGT'	149	0	0	0		'CGT'	149	0	0	0	
'GAT'	150	0	0	0		'GAT'	150	0	0	0		'GAT'	150	0	0	0	
'AGG'	151	GGG	G	NoSin	0.25	'AGG'	151	GGG	G	NoSin	0.41	'AGG'	151	GGG	G	NoSin	0.24
'ACC'	152	GCC	A	NoSin	0.29	'ACC'	152	GCC	A	NoSin	0.28	'ACC'	152	GCC	A	NoSin	0.35
'CCT'	153	0	0	0		'CCT'	153	0	0	0		'CCT'	153	0	0	0	
'TAT'	154	0	0	0		'TAT'	154	0	0	0		'TAT'	154	0	0	0	
'CGG'	155	0	0	0		'CGG'	155	0	0	0		'CGG'	155	0	0	0	
'ACC'	156	0	0	0		'ACC'	156	GCC	A	NoSin	0.25	'ACC'	156	GCC	A	NoSin	0.24
'CTA'	157	0	0	0		'CTA'	157	0	0	0		'CTA'	157	0	0	0	
'TTG'	158	0	0	0		'TTG'	158	0	0	0		'TTG'	158	0	0	0	
'ATG'	159	0	0	0		'ATG'	159	0	0	0		'ATG'	159	0	0	0	
'AAT'	160	GAT/ AGT	D/S	NoSin	0.24/0. 25	'AAT'	160	AGT	S	NoSin	0.66	'AAT'	160	0	0	0	
'GAG'	161	0	0	0		'GAG'	161	GGG	G	NoSin	0.29	'GAG'	161	0	0	0	
'TTA'	162	0	0	0		'TTA'	162	TTG	L	Sin	0.23	'TTA'	162	0	0	0	
'GGT'	163	GGC	G	Sin	0.37	'GGT'	163	GGC	G	Sin	0.29	'GGT'	163	0	0	0	
'GTT'	164	GCT	A	NoSin	0.34	'GTT'	164	GAT/ GCT	D/A	NoSin	0.34/0. 3	'GTT'	164	GCT/ GTC	A/V	NoSin /Sin	0.26/ 0.27

'CCT'	165	0	0	0		'CCT'	165	CCC	P	Sin	0.35	'CCT'	165	0	0	0	
'TTT'	166	0	0	0		'TTT'	166	0	0	0		'TTT'	166	0	0	0	
'CAT'	167	CGT	R	NoSin	0.35	'CAT'	167	TAT	Y	NoSin	0.3	'CAT'	167	0	0	0	
'CTG'	168	0	0	0		'CTG'	168	0	0	0		'CTG'	168	0	0	0	
'GGG'	169	0	0	0		'GGG'	169	0	0	0		'GGG'	169	AGG	R	NoSin	0.46
'ACC'	170	GCC	A	NoSin	0.28	'ACC'	170	0	0	0		'ACC'	170	GCC	A	NoSin	0.36
'AAG'	171	GAG	E	NoSin	0.27	'AAG'	171	AGG	R	NoSin	0.82	'AAG'	171	AGG	R	NoSin	0.27
'CAA'	172	CAG	Q	Sin	0.28	'CAA'	172	CAG	Q	Sin	0.31	'CAA'	172	CAG	Q	Sin	0.3
'GTG'	173	0	0	0		'GTG'	173	0	0	0		'GTG'	173	0	0	0	
'TGC'	174	CGC	R	NoSin	0.37	'TGC'	174	CGC/ TGG	R/W	NoSin	0.25/0. 3	'TGC'	174	CGC	R	NoSin	0.28
'ATA'	175	0	0	0		'ATA'	175	0	0	0		'ATA'	175	0	0	0	
'GCA'	176	0	0	0		'GCA'	176	0	0	0		'GCA'	176	0	0	0	
'TGG'	177	0	0	0		'TGG'	177	0	0	0		'TGG'	177	0	0	0	
'TCC'	178	CCC	P	NoSin	0.4	'TCC'	178	CCC	P	NoSin	0.42	'TCC'	178	CCC	P	NoSin	0.28
'AGC'	179	0	0	0		'AGC'	179	0	0	0		'AGC'	179	0	0	0	
'TCA'	180	CCA	P	NoSin	0.43	'TCA'	180	CCA	P	NoSin	0.29	'TCA'	180	CCA	P	NoSin	0.31
'AGT'	181	0	0	0		'AGT'	181	0	0	0		'AGT'	181	0	0	0	
'TGT'	182	0	0	0		'TGT'	182	0	0	0		'TGT'	182	0	0	0	
'CAC'	183	CGC	R	NoSin	0.27	'CAC'	183	CGC	R	NoSin	0.26	'CAC'	183	0	0	0	
'GAT'	184	0	0	0		'GAT'	184	0	0	0		'GAT'	184	0	0	0	
'GGA'	185	0	0	0		'GGA'	185	0	0	0		'GGA'	185	0	0	0	
'AAA'	186	AAG	K	Sin	0.53	'AAA'	186	AAG	K	Sin	0.45	'AAA'	186	AAG	K	Sin	0.49

'GCA'	187	0	0	0		'GCA'	187	0	0	0		'GCA'	187	0	0	0	
'TGG'	188	0	0	0		'TGG'	188	0	0	0		'TGG'	188	0	0	0	
'CTG'	189	CCG/ CTA	P/L	NoSin /Sin	0.3/ 0.81	'CTG'	189	CCG	P	NoSin	0.45	'CTG'	189	CCG/ CTA	P/L	NoSin /Sin	0.43/ 0.4
'CAT'	190	0	0	0		'CAT'	190	CGT/ CAC	R/H	NoSin /Sin	0.43/0. 47	'CAT'	190	CGT	R	NoSin	0.4
'GTT'	191	0	0	0		'GTT'	191	0	0	0		'GTT'	191	GCT/ GTC	A/V	NoSin /Sin	0.34/ 0.35
'TGT'	192	TGC	C	Sin	0.48	'TGT'	192	TGC	C	Sin	0.45	'TGT'	192	TGC	C	Sin	0.51
'ATA'	193	0	0	0		'ATA'	193	0	0	0		'ATA'	193	0	0	0	
'ACG'	194	GCG	A	NoSin	0.37	'ACG'	194	GCG	A	NoSin	0.46	'ACG'	194	GCG	A	NoSin	0.44
'GGG'	195	0	0	0		'GGG'	195	0	0	0		'GGG'	195	0	0	0	
'GAT'	196	GGT	G	NoSin	0.32	'GAT'	196	GGT	G	NoSin	0.3	'GAT'	196	GGT	G	NoSin	0.32
'GAT'	197	GGT	G	NoSin	0.29	'GAT'	197	GGT	G	NoSin	0.43	'GAT'	197	GGT	G	NoSin	0.37
'AAA'	198	0	0	0		'AAA'	198	0	0	0		'AAA'	198	0	0	0	
'AAT'	199	0	0	0		'AAT'	199	0	0	0		'AAT'	199	0	0	0	
'GCA'	200	GCG	A	Sin	0.27	'GCA'	200	GCG	A	Sin	0.25	'GCA'	200	0	0	0	
'ACT'	201	GCT	A	NoSin	0.39	'ACT'	201	GCT/ ATT	A/I	NoSin	0.33/0. 35	'ACT'	201	0	0	0	
'GCT'	202	0	0	0		'GCT'	202	ACT	T	NoSin	0.43	'GCT'	202	0	0	0	
'AGC'	203	GGC	G	NoSin	0.29	'AGC'	203	GGC	G	NoSin	0.25	'AGC'	203	GGC	G	NoSin	0.4
'TTC'	204	CTC	L	NoSin	0.27	'TTC'	204	CTC	L	NoSin	0.31	'TTC'	204	CTC	L	NoSin	0.34
'ATT'	205	0	0	0		'ATT'	205	0	0	0		'ATT'	205	0	0	0	
'TAC'	206	0	0	0		'TAC'	206	0	0	0		'TAC'	206	0	0	0	

'AAT'	207	AGT	S	NoSin	0.31	'AAT'	207	AGT	S	NoSin	0.28	'AAT'	207	AGT	S	NoSin	0.36
'GGG'	208	0	0	0		'GGG'	208	0	0	0		'GGG'	208	0	0	0	
'AGG'	209	GGG	G	NoSin	0.38	'AGG'	209	GGG	G	NoSin	0.47	'AGG'	209	GGG	G	NoSin	0.39
'CTT'	210	0	0	0		'CTT'	210	0	0	0		'CTT'	210	0	0	0	
'GTA'	211	GTG	V	Sin	0.31	'GTA'	211	GTG	V	Sin	0.25	'GTA'	211	GTG	V	Sin	0.26
'GAT'	212	GGT	G	NoSin	0.32	'GAT'	212	GGT	G	NoSin	0.32	'GAT'	212	0	0	0	
'AGT'	213	AGC	S	Sin	0.32	'AGT'	213	AGC	S	Sin	0.32	'AGT'	213	AGC	S	Sin	0.32
'GTT'	214	0	0	0		'GTT'	214	0	0	0		'GTT'	214	0	0	0	
'GTT'	215	GCT	A	NoSin	0.32	'GTT'	215	GCT	A	NoSin	0.35	'GTT'	215	GCT	A	NoSin	0.38
'TCA'	216	TCG	S	Sin	0.34	'TCA'	216	TCG	S	Sin	0.27	'TCA'	216	0	0	0	
'TGG'	217	0	0	0		'TGG'	217	0	0	0		'TGG'	217	0	0	0	
'TCC'	218	0	0	0		'TCC'	218	CCC	P	NoSin	0.22	'TCC'	218	0	0	0	
'AAA'	219	0	0	0		'AAA'	219	AAG	K	Sin	0.22	'AAA'	219	AAG	K	Sin	0.28
'GAA'	220	0	0	0		'GAA'	220	0	0	0		'GAA'	220	0	0	0	
'ATC'	221	0	0	0		'ATC'	221	0	0	0		'ATC'	221	0	0	0	
'CTC'	222	0	0	0		'CTC'	222	0	0	0		'CTC'	222	0	0	0	
'AGG'	223	GGG	G	NoSin	0.24	'AGG'	223	GGG	G	NoSin	0.23	'AGG'	223	GGG	G	NoSin	0.28
'ACC'	224	GCC	A	NoSin	0.35	'ACC'	224	GCC	A	NoSin	0.28	'ACC'	224	0	0	0	
'CAG'	225	CGG	R	NoSin	0.25	'CAG'	225	0	0	0		'CAG'	225	CGG	R	NoSin	0.25
'GAG'	226	GGG	G	NoSin	0.29	'GAG'	226	GGG	G	NoSin	0.29	'GAG'	226	0	0	0	
'TCA'	227	CCA/ TCG	P/S	NoSin /Sin	0.31/0. 38	'TCA'	227	CCA/ TCG	P/S	NoSin /Sin	0.24/0. 25	'TCA'	227	CCA/ TCG	P/S	NoSin /Sin	0.29/ 0.27
'GAA'	228	0	0	0		'GAA'	228	0	0	0		'GAA'	228	0	0	0	

'TGC'	229	0	0	0		'TGC'	229	0	0	0		'TGC'	229	0	0	0	
'GTT'	230	0	0	0		'GTT'	230	GCT	A	NoSin	0.24	'GTT'	230	0	0	0	
'TGT'	231	0	0	0		'TGT'	231	TGC	C	Sin	0.36	'TGT'	231	TGC	C	Sin	0.28
'ATC'	232	0	0	0		'ATC'	232	0	0	0		'ATC'	232	0	0	0	
'AAT'	233	0	0	0		'AAT'	233	0	0	0		'AAT'	233	0	0	0	
'GGA'	234	GGG	G	Sin	0.25	'GGA'	234	GGG	G	Sin	0.29	'GGA'	234	0	0	0	
'ACT'	235	GCT	A	NoSin	0.36	'ACT'	235	GCT	A	NoSin	0.29	'ACT'	235	GCT	A	NoSin	0.33
'TGT'	236	0	0	0		'TGT'	236	TGC	C	Sin	0.31	'TGT'	236	0	0	0	
'ACA'	237	GCA	A	NoSin	0.29	'ACA'	237	GCA	A	NoSin	0.25	'ACA'	237	GCA	A	NoSin	0.24
'GTA'	238	0	0	0		'GTA'	238	0	0	0		'GTA'	238	0	0	0	
'GTA'	239	0	0	0		'GTA'	239	0	0	0		'GTA'	239	0	0	0	
'ATG'	240	0	0	0		'ATG'	240	0	0	0		'ATG'	240	0	0	0	
'ACT'	241	0	0	0		'ACT'	241	0	0	0		'ACT'	241	0	0	0	
'GAT'	242	GGT	G	NoSin	0.34	'GAT'	242	0	0	0		'GAT'	242	GGT	G	NoSin	0.25
'GGG'	243	0	0	0		'GGG'	243	0	0	0		'GGG'	243	0	0	0	
'AGT'	244	GGT	G	NoSin	0.27	'AGT'	244	GGT	G	NoSin	0.24	'AGT'	244	GGT	G	NoSin	0.34
'GCT'	245	0	0	0		'GCT'	245	0	0	0		'GCT'	245	0	0	0	
'TCA'	246	TCG	S	Sin	0.32	'TCA'	246	TCG	S	NoSin	0.22	'TCA'	246	0	0	0	
'GGA'	247	0	0	0		'GGA'	247	0	0	0		'GGA'	247	0	0	0	
'AAA'	248	AAG	K	Sin	0.54	'AAA'	248	AAG	K	Sin	0.53	'AAA'	248	AAG	K	Sin	0.52
'GCT'	249	0	0	0		'GCT'	249	0	0	0		'GCT'	249	0	0	0	
'GAT'	250	0	0	0		'GAT'	250	0	0	0		'GAT'	250	0	0	0	
'ACT'	251	0	0	0		'ACT'	251	0	0	0		'ACT'	251	0	0	0	

'AAA'	252	0	0	0		'AAA'	252	0	0	0		'AAA'	252	0	0	0	
'ATA'	253	0	0	0		'ATA'	253	0	0	0		'ATA'	253	0	0	0	
'CTA'	254	0	0	0		'CTA'	254	0	0	0		'CTA'	254	0	0	0	
'TTC'	255	CTC	L	NoSin	0.24	'TTC'	255	CTC	L	NoSin	0.23	'TTC'	255	CTC	L	NoSin	0.33
'ATT'	256	0	0	0		'ATT'	256	0	0	0		'ATT'	256	0	0	0	
'GAG'	257	GGG	G	NoSin	**0.41**	'GAG'	257	GGG/ GAA	G/E	NoSin /Sin	0.29/0. 47	'GAG'	257	GGG	G	NoSin	0.28
'GAG'	258	0	0	0		'GAG'	258	0	0	0		'GAG'	258	0	0	0	
'GGG'	259	0	0	0		'GGG'	259	0	0	0		'GGG'	259	0	0	0	
'AAA'	260	0	0	0		'AAA'	260	0	0	0		'AAA'	260	0	0	0	
'ATC'	261	0	0	0		'ATC'	261	0	0	0		'ATC'	261	ATT	I	Sin	0.28
'GTT'	262	0	0	0		'GTT'	262	0	0	0		'GTT'	262	0	0	0	
'CAT'	263	0	0	0		'CAT'	263	AAT	N	NoSin	0.5	'CAT'	263	0	0	0	
'ACT'	264	0	0	0		'ACT'	264	0	0	0		'ACT'	264	0	0	0	
'AGC'	265	0	0	0		'AGC'	265	0	0	0		'AGC'	265	0	0	0	
'ACA'	266	0	0	0		'ACA'	266	0	0	0		'ACA'	266	0	0	0	
'TTG'	267	0	0	0		'TTG'	267	TCG	S	NoSin	0.28	'TTG'	267	0	0	0	
'TCA'	268	TCG	C	NoSin	0.65	'TCA'	268	CCA/ TCG	P/S	NoSin /Sin	0.23/0. 29	'TCA'	268	TCG	S	Sin	0.27
'GGA'	269	0	0	0		'GGA'	269	0	0	0		'GGA'	269	0	0	0	
'AGT'	270	AGC	S	Sin	0.32	'AGT'	270	0	0	0		'AGT'	270	0	0	0	
'GCT'	271	0	0	0		'GCT'	271	0	0	0		'GCT'	271	0	0	0	
'CAG'	272	CGG	R	NoSin	0.51	'CAG'	272	0	0	0		'CAG'	272	CGG	R	NoSin	0.65

'CAT'	273	CGT/CAC	R/H	NoSin/Sin	0.68/0.32	'CAT'	273	CGT	R	NoSin	0.21	'CAT'	273	CGT/CAC	R/H	NoSin/Sin	0.61/0.39
'GTC'	274	GCC	A	NoSin	0.55	'GTT'	274	GTC	V	Sin	0.3	'GTC'	274	GCC	A	NoSin	0.43
'GAG'	275	0	0	0		'GAG'	275	0	0	0		'GAG'	275	0	0	0	
'GAG'	276	GGG	G	NoSin	0.65	'GAG'	276	GGG	G	NoSin	0.37	'GAG'	276	GGG	G	NoSin	0.74
'TGC'	277	0	0	0		'TGC'	277	TAC	Y	NoSin	0.26	'TGC'	277	0	0	0	
'TCT'	278	0	0	0		'TCT'	278	0	0	0		'TCT'	278	CCT	P	NoSin	0.28
'TGC'	279	0	0	0		'TGC'	279	0	0	0		'TGC'	279	0	0	0	
'TAT'	280	CAT	H	NoSin	0.29	'TAT'	280	CAT	H	NoSin	0.29	'TAT'	280	CAT	H	NoSin	0.27
'CCT'	281	0	0	0		'CCT'	281	0	0	0		'CCT'	281	0	0	0	
'CGG'	282	0	0	0		'CGA'	282	0	0	0		'CGA'	282	0	0	0	
'TAT'	283	0	0	0		'TAT'	283	0	0	0		'TAT'	283	0	0	0	
'CCT'	284	CCC	P	Sin	0.32	'CCT'	284	CCC	P	Sin	0.21	'CCT'	284	CCC	P	Sin	0.25
'GGT'	285	GGC	G	Sin	1.08	'GGT'	285	GGC	G	Sin	0.34	'GGT'	285	GGC	G	Sin	0.35
'GTC'	286	0	0	0		'GTC'	286	GCC	A	NoSin	0.56	'GTC'	286	GCC	A	NoSin	0.26
'AGA'	287	0	0	0		'AGA'	287	0	0	0		'AGA'	287	0	0	0	
'TGT'	288	0	0	0		'TGT'	288	0	0	0		'TGT'	288	0	0	0	
'GTC'	289	GCC	A	NoSin	0.34	'GCC'	289	0	0	0		'GTC'	289	GCC	A	NoSin	0.37
'TGC'	290	CGC	R	NoSin	0.3	'TGC'	290	CGC	R	NoSin	0.43	'TGC'	290	CGC	R	NoSin	0.28
'AGA'	291	AGG	R	Sin	0.31	'AGA'	291	GGA	G	NoSin	0.28	'AGA'	291	GGA/AGG	G/R	NoSin/Sin	0.24/0.23
'GAC'	292	GGC	G	NoSin	0.26	'GAC'	292	GGC	G	NoSin	0.25	'GAC'	292	GGC	G	NoSin	0.23
'AAC'	293	0	0	0		'AAC'	293	GAC	D	NoSin	0.32	'AAC'	293	0	0	0	

'TGG'	294	0	0	0		'TGG'	294	0	0	0		'TGG'	294	0	0	0	
'AAA'	295	GAA/AAG	E/K	NoSin/Sin	0.38/0.43	'AAA'	295	AAG	K	Sin	0.37	'AAA'	295	AAG	K	Sin	0.45
'GGA'	296	0	0	0		'GGA'	296	GGG	G	Sin	0.22	'GGA'	296	GGG	G	Sin	0.31
'TCC'	297	0	0	0		'TCC'	297	0	0	0		'TCC'	297	0	0	0	
'AAT'	298	0	0	0		'AAT'	298	0	0	0		'AAT'	298	0	0	0	
'AGG'	299	0	0	0		'AGG'	299	0	0	0		'AGG'	299	0	0	0	
'CCC'	300	0	0	0		'CCC'	300	0	0	0		'CCC'	300	0	0	0	
'ATC'	301	0	0	0		'ATC'	301	ACC	T	NoSin	0.31	'ATC'	301	ACC	T	NoSin	0.29
'GTA'	302	0	0	0		'GTA'	302	0	0	0		'GTA'	302	0	0	0	
'GAT'	303	0	0	0		'GAT'	303	0	0	0		'GAT'	303	0	0	0	
'ATA'	304	0	0	0		'ATA'	304	0	0	0		'ATA'	304	0	0	0	
'AAC'	305	AAT	N	Sin	0.37	'AAC'	305	0	0	0		'AAC'	305	0	0	0	
'ATA'	306	0	0	0		'ATA'	306	0	0	0		'ATA'	306	0	0	0	
'AAG'	307	0	0	0		'AAG'	307	0	0	0		'AAG'	307	0	0	0	
'GAT'	308	0	0	0		'GAT'	308	0	0	0		'GAT'	308	0	0	0	
'CAT'	309	0	0	0		'CAT'	309	CGT	R	NoSin	0.3	'CAT'	309	CGT	R	NoSin	0.23
'AGC'	310	GGC	G	NoSin	0.28	'AGC'	310	GGC	G	NoSin	0.23	'AGC'	310	GGC	G	NoSin	0.24
'ATT'	311	0	0	0		'ATT'	311	0	0	0		'ATT'	311	0	0	0	
'GTT'	312	0	0	0		'GTT'	312	0	0	0		'GTT'	312	0	0	0	
'TCC'	313	0	0	0		'TCC'	313	0	0	0		'TCC'	313	0	0	0	
'AGG'	314	GGG	G	NoSin	0.29	'AGG'	314	GGG	G	NoSin	0.22	'AGG'	314	GGG	G	NoSin	0.26
'TAT'	315	0	0	0		'TAT'	315	0	0	0		'TAT'	315	0	0	0	

'GTG'	316	0	0	0		'GTG'	316	GCG	A	NoSin	0.25	'GTG'	316	GCG	A	NoSin	0.22
'TGT'	317	0	0	0		'TGT'	317	CGT	R	NoSin	0.3	'TGT'	317	0	0	0	
'TCA'	318	TCG	S	Sin	0.38	'TCA'	318	TGA/ TCG	STOP/ S	NoSin /Sin	0.48/0. 27	'TCA'	318	TCG	S	Sin	0.31
'GGA'	319	AGA	R	NoSin	1.17	'GGA'	319	0	0	0		'GGA'	319	0	0	0	
'CTT'	320	0	0	0		'CTT'	320	0	0	0		'CTT'	320	0	0	0	
'GTT'	321	0	0	0		'GTT'	321	GCT	A	NoSin	0.21	'GTT'	321	0	0	0	
'GGA'	322	GGG	G	Sin	0.5	'GGA'	322	GGG	G	Sin	0.42	'GGA'	322	GGG	G	Sin	0.36
'GAC'	323	GGC	G	NoSin	0.39	'GAC'	323	GGC	G	NoSin	0.38	'GAC'	323	GGC	G	NoSin	0.46
'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.34/0. 53	'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.3/ 0.42	'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.28/ 0.35
'CCC'	325	0	0	0		'CCC'	325	0	0	0		'CCC'	325	0	0	0	
'AGA'	326	0	0	0		'AGA'	326	0	0	0		'AGA'	326	0	0	0	
'AAA'	327	0	0	0		'AAA'	327	0	0	0		'AAA'	327	0	0	0	
'AAC'	328	0	0	0		'AAC'	328	0	0	0		'AAC'	328	0	0	0	
'GAC'	329	GGC	G	NoSin	0.28	'GAC'	329	GGC/ GAA	G/E	NoSin	0.24/0. 39	'GAC'	329	GGC	G	NoSin	0.28
'AGC'	330	GGC	G	NoSin	0.38	'AGC'	330	GGC	G	NoSin	0.32	'AGC'	330	GGC	G	NoSin	0.3
'TCC'	331	0	0	0		'TCC'	331	0	0	0		'TCC'	331	0	0	0	
'AGC'	332	GGC	G	NoSin	0.26	'AGC'	332	0	0	0		'AGC'	332	GGC	G	NoSin	0.24
'AGT'	333	0	0	0		'AGT'	333	0	0	0		'AGT'	333	0	0	0	
'AGC'	334	GGC	G	NoSin	0.31	'AGC'	334	0	0	0		'AGC'	334	0	0	0	
'CAT'	335	0	0	0		'CAT'	335	0	0	0		'CAT'	335	0	0	0	
'TGT'	336	0	0	0		'TGT'	336	TGC	C	Sin	0.42	'TGT'	336	TGC	C	NoSin	0.22

'TTG'	337	0	0	0		'TTG'	337	0	0	0		'TTG'	337	0	0	0	
'GAT'	338	0	0	0		'GAT'	338	0	0	0		'GAT'	338	0	0	0	
'CCT'	339	0	0	0		'CCT'	339	0	0	0		'CCT'	339	0	0	0	
'AAC'	340	0	0	0		'AAC'	340	0	0	0		'AAC'	340	0	0	0	
'AAT'	341	0	0	0		'AAT'	341	0	0	0		'AAT'	341	0	0	0	
'GAA'	342	GAG	E	Sin	0.28	'GAA'	342	GAG	E	Sin	0.35	'GAA'	342	GAG	E	Sin	0.33
'GAA'	343	GAG	E	Sin	0.26	'GAA'	343	GAG	E	Sin	0.31	'GAA'	343	GAG	E	Sin	0.26
'GGT'	344	0	0	0		'GGT'	344	0	0	0		'GGT'	344	0	0	0	
'GGT'	345	0	0	0		'GGT'	345	GGC	G	Sin	0.29	'GGT'	345	GGC	G	Sin	0.35
'CAT'	346	0	0	0		'CAT'	346	0	0	0		'CAT'	346	0	0	0	
'GGA'	347	GGG	G	Sin	1.61	'GGA'	347	GGG	G	Sin	0.35	'GGA'	347	GGG	G	Sin	0.36
'GTG'	348	0	0	0		'GTG'	348	0	0	0		'GTG'	348	0	0	0	
'AAA'	349	AAG	K	Sin	0.51	'AAA'	349	AAG	K	Sin	0.4	'AAA'	349	AAG	K	Sin	0.4
'GGC'	350	0	0	0		'GGC'	350	0	0	0		'GGC'	350	0	0	0	
'TGG'	351	CGG	R	NoSin	0.26	'TGG'	351	CGG	R	NoSin	0.26	'TGG'	351	CGG	R	NoSin	0.32
'GCC'	352	0	0	0		'GCC'	352	0	0	0		'GCC'	352	0	0	0	
'TTT'	353	CTT	L	NoSin	0.3	'TTT'	353	CTT	L	NoSin	0.29	'TTT'	353	CTT	L	NoSin	0.32
'GAT'	354	GGT	G	NoSin	0.25	'GAT'	354	0	0	0		'GAT'	354	GGT	G	NoSin	0.24
'GAT'	355	0	0	0		'GAT'	355	0	0	0		'GAT'	355	GGT	G	NoSin	0.35
'GGA'	356	0	0	0		'GGA'	356	0	0	0		'GGA'	356	0	0	0	
'AAT'	357	AAC	N	Sin	3.07	'AAT'	357	AGT	S	NoSin	0.29	'AAT'	357	0	0	0	
'GAC'	358	GGC	G	NoSin	0.4	'GAC'	358	GGC	G	NoSin	0.36	'GAC'	358	GGC	G	NoSin	0.32
'GTG'	359	0	0	0		'GTG'	359	0	0	0		'GTG'	359	0	0	0	

'TGG'	360	0	0	0		'TGG'	360	0	0	0		'TGG'	360	0	0	0	
'ATG'	361	0	0	0		'ATG'	361	GTG	V	NoSin	0.23	'ATG'	361	0	0	0	
'GGA'	362	0	0	0		'GGA'	362	0	0	0		'GGA'	362	0	0	0	
'AGA'	363	0	0	0		'AGA'	363	0	0	0		'AGA'	363	0	0	0	
'ACG'	364	0	0	0		'ACG'	364	0	0	0		'ACG'	364	GCG	A	NoSin	0.24
'ATC'	365	0	0	0		'ATC'	365	GTC	V	NoSin	0.23	'ATC'	365	0	0	0	
'AAC'	366	AGC	S	NoSin	0.28	'AAC'	366	0	0	0		'AAC'	366	AGC	S	NoSin	0.29
'GAG'	367	0	0	0		'GAG'	367	GGG	G	NoSin	0.26	'GAG'	367	GGG	G	NoSin	0.23
'ACG'	368	GCG	A	NoSin	0.36	'ACG'	368	GCG	A	NoSin	0.47	'ACG'	368	GCG	A	NoSin	0.33
'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.52/0. 32	'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.37/0. 24	'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.36/ 0.24
'CGT'	370	0	0	0		'CGT'	370	0	0	0		'CGT'	370	0	0	0	
'TTA'	371	0	0	0		'TTA'	371	CTA	L	NoSin	0.23	'TTA'	371	0	0	0	
'GGG'	372	0	0	0		'GGG'	372	0	0	0		'GGG'	372	0	0	0	
'TAT'	373	0	0	0		'TAT'	373	0	0	0		'TAT'	373	0	0	0	
'GAA'	374	0	0	0		'GAA'	374	0	0	0		'GAA'	374	0	0	0	
'ACC'	375	GCC	A	NoSin	0.98	'ACC'	375	GCC	A	NoSin	0.33	'ACC'	375	GCC	A	NoSin	0.39
'TTC'	376	0	0	0		'TTC'	376	0	0	0		'TTC'	376	0	0	0	
'AAA'	377	0	0	0		'AAA'	377	0	0	0		'AAA'	377	0	0	0	
'GTC'	378	0	0	0		'GTC'	378	GCC	A	NoSin	0.23	'GTC'	378	GCC	A	NoSin	0.23
'ATT'	379	0	0	0		'ATT'	379	0	0	0		'ATT'	379	0	0	0	
'GAA'	380	0	0	0		'GAA'	380	GAG	E	NoSin	0.27	'GAA'	380	0	0	0	
'GGC'	381	0	0	0		'GGC'	381	0	0	0		'GGC'	381	0	0	0	

'TGG'	382	0	0	0		'TGG'	382	0	0	0		'TGG'	382	CGG	R	NoSin	0.24
'TCC'	383	CCC	P	NoSin	0.29	'TCC'	383	CCC	P	NoSin	0.24	'TCC'	383	CCC	P	NoSin	0.25
'AAC'	384	0	0	0		'AAC'	384	0	0	0		'AAC'	384	0	0	0	
'CCT'	385	0	0	0		'CCT'	385	0	0	0		'CCT'	385	CAT	H	NoSin	0.29
'AAG'	386	0	0	0		'AAG'	386	0	0	0		'AAG'	386	0	0	0	
'TCC'	387	0	0	0		'TCC'	387	CCC	P	NoSin	0.23	'TCC'	387	CCC	P	NoSin	0.26
'AAA'	388	0	0	0		'AAA'	388	0	0	0		'AAA'	388	0	0	0	
'TTG'	389	0	0	0		'TTG'	389	CTG	L	Sin	0.27	'TTG'	389	0	0	0	
'CAG'	390	CGG	R	NoSin	0.34	'CAG'	390	CGG	R	NoSin	0.42	'CAG'	390	CGG	R	NoSin	0.43
'ATA'	391	0	0	0		'ATA'	391	0	0	0		'ATA'	391	0	0	0	
'AAT'	392	0	0	0		'AAT'	392	0	0	0		'AAT'	392	0	0	0	
'AGG'	393	0	0	0		'AGG'	393	0	0	0		'AGG'	393	GGG	G	NoSin	0.24
'CAA'	394	CAG	Q	Sin	0.3	'CAA'	394	0	0	0		'CAA'	394	CAG	Q	Sin	0.37
'GTC'	395	GCC	A	NoSin	0.39	'GTC'	395	GCC	A	NoSin	0.33	'GTC'	395	GCC	A	NoSin	0.25
'ATA'	396	ATG	M	NoSin	0.29	'ATA'	396	CTA	L	NoSin	0.26	'ATA'	396	0	0	0	
'GTT'	397	0	0	0		'GTT'	397	0	0	0		'GTT'	397	0	0	0	
'GAC'	398	GGC	G	NoSin	0.37	'GAC'	398	GGC	G	NoSin	0.49	'GAC'	398	GGC	G	NoSin	0.43
'AGA'	399	AGG	R	Sin	0.42	'AGA'	399	AGG	R	Sin	0.37	'AGA'	399	AGG	R	Sin	0.36
'GGT'	400	0	0	0		'GGT'	400	0	0	0		'GGT'	400	0	0	0	
'AAT'	401	0	0	0		'AAT'	401	0	0	0		'AAT'	401	0	0	0	
'AGG'	402	0	0	0		'AGG'	402	GGG	G	NoSin	0.43	'AGG'	402	GGG	G	NoSin	0.27
'TCC'	403	0	0	0		'TCC'	403	0	0	0		'TCC'	403	0	0	0	
'GGT'	404	0	0	0		'GGT'	404	GGC	G	NoSin	0.34	'GGT'	404	0	0	0	

'TAT'	405	0	0	0		'TAT'	405	0	0	0		'TAT'	405	0	0	0	
'TCT'	406	0	0	0		'TCT'	406	0	0	0		'TCT'	406	0	0	0	
'GGT'	407	0	0	0		'GGT'	407	0	0	0		'GGT'	407	0	0	0	
'ATT'	408	0	0	0		'ATT'	408	0	0	0		'ATT'	408	0	0	0	
'TTT'	409	0	0	0		'TTT'	409	0	0	0		'TTT'	409	0	0	0	
'TCT'	410	0	0	0		'TCT'	410	0	0	0		'TCT'	410	0	0	0	
'GTT'	411	0	0	0		'GTT'	411	0	0	0		'GTT'	411	GCT	A	NoSin	0.34
'GAA'	412	0	0	0		'GAA'	412	GAG	E	Sin	0.56	'GAA'	412	GAG	E	Sin	0.37
'GGC'	413	0	0	0		'GGC'	413	0	0	0		'GGC'	413	0	0	0	
'AAA'	414	AGA	R	NoSin	0.51	'AAA'	414	AGA	R	NoSin	0.54	'AAA'	414	AGA	R	NoSin	0.38
'AGC'	415	GGC	G	NoSin	0.5	'AGC'	415	GGC	G	NoSin	0.74	'AGC'	415	GGC	G	NoSin	0.46
'TGC'	416	0	0	0		'TGC'	416	0	0	0		'TGC'	416	CGC	R	NoSin	0.36
'ATC'	417	GTC	V	NoSin	0.9	'ATC'	417	GTC	V	NoSin	0.48	'ATC'	417	0	0	0	
'AAT'	418	0	0	0		'AAT'	418	0	0	0		'AAT'	418	0	0	0	
'CGG'	419	0	0	0		'CGG'	419	0	0	0		'CGG'	419	0	0	0	
'TGC'	420	0	0	0		'TGC'	420	0	0	0		'TGC'	420	0	0	0	
'TTT'	421	0	0	0		'TTT'	421	CTT	L	NoSin	0.43	'TTT'	421	0	0	0	
'TAT'	422	0	0	0		'TAT'	422	0	0	0		'TAT'	422	0	0	0	
'GTG'	423	GCG	A	NoSin	0.41	'GTG'	423	0	0	0		'GTG'	423	GCG	A	NoSin	0.48
'GAG'	424	0	0	0		'GAG'	424	0	0	0		'GAG'	424	0	0	0	
'TTG'	425	0	0	0		'TTG'	425	0	0	0		'TTG'	425	0	0	0	
'ATT'	426	0	0	0		'ATT'	426	GTT	V	NoSin	0.45	'ATT'	426	0	0	0	
'AGG'	427	0	0	0		'AGG'	427	0	0	0		'AGG'	427	0	0	0	

'GGA'	428	0	0	0		'GGA'	428	0	0	0		'GGA'	428	0	0	0	
'AGA'	429	0	0	0		'AGA'	429	0	0	0		'AGA'	429	0	0	0	
'AAA'	430	0	0	0		'AAA'	430	AAG	K	Sin	0.95	'AAA'	430	AAG	K	Sin	0.65
'GAG'	431	0	0	0		'GAG'	431	0	0	0		'GAG'	431	0	0	0	
'GAA'	432	0	0	0		'GAA'	432	0	0	0		'GAA'	432	GAG	E	Sin	0.59
'ACT'	433	0	0	0		'ACT'	433	0	0	0		'ACT'	433	GCT	A	NoSin	0.59
'GAA'	434	0	0	0		'GAA'	434	GAG	E	Sin	0.6	'GAA'	434	GGA	G	NoSin	0.65
'GTC'	435	0	0	0		'GTC'	435	0	0	0		'GTC'	435	0	0	0	
'TTG'	436	CTG	L	Sin	0.73	'TTG'	436	0	0	0		'TTG'	436	0	0	0	
'TGG'	437	0	0	0		'TGG'	437	0	0	0		'TGG'	437	0	0	0	
'ACC'	438	0	0	0		'ACC'	438	0	0	0		'ACC'	438	0	0	0	
'TCA'	439	0	0	0		'TCA'	439	0	0	0		'TCA'	439	0	0	0	
'AAC'	440	0	0	0		'AAC'	440	AGC	S	NoSin	0.95	'AAC'	440	AGC	S	NoSin	1.09
'AGT'	441	0	0	0		'AGT'	441	0	0	0		'AGT'	441	0	0	0	
'ATT'	442	0	0	0		'ATT'	442	0	0	0		'ATT'	442	0	0	0	

Muestra 202						Muestra 206						Muestra 220					
Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%	Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%	Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%
'ATG'	23	0	0	0													
'CAA'	24	0	0	0													
'ATT'	25	0	0	0													
'GCC'	26	0	0	0													
'ATC'	27	0	0	0													
'TTG'	28	0	0	0													
'ATA'	29	0	0	0													
'ACT'	30	0	0	0													
'ACT'	31	0	0	0													
'GTA'	32	0	0	0													
'ACA'	33	0	0	0		'ACA'	33	0	0	0		'ACA'	33	0	0	0	
'TTG'	34	0	0	0		'TTG'	34	0	0	0		'TTG'	34	0	0	0	
'CAT'	35	0	0	0		'CAT'	35	0	0	0		'CAT'	35	0	0	0	
'TTC'	36	0	0	0		'TTC'	36	0	0	0		'TTC'	36	0	0	0	
'AAG'	37	0	0	0		'AAG'	37	0	0	0		'AAG'	37	0	0	0	

'CAA'	38	0	0	0		'CAA'	38	0	0	0		'CAA'	38	0	0	0	
'TAT'	39	0	0	0		'TAT'	39	0	0	0		'TAT'	39	0	0	0	
'GAA'	40	0	0	0		'GAA'	40	0	0	0		'GAA'	40	0	0	0	
'TTC'	41	0	0	0		'TTC'	41	0	0	0		'TTC'	41	0	0	0	
'AAC'	42	0	0	0		'AAC'	42	0	0	0		'AAC'	42	0	0	0	
'TCC'	43	0	0	0		'TCC'	43	0	0	0		'TCC'	43	0	0	0	
'CCC'	44	0	0	0		'CCC'	44	0	0	0		'CCC'	44	0	0	0	
'CCA'	45	0	0	0		'CCA'	45	0	0	0		'CCA'	45	0	0	0	
'AAC'	46	0	0	0		'AAC'	46	0	0	0		'AAC'	46	0	0	0	
'AAC'	47	0	0	0		'AAC'	47	0	0	0		'AAC'	47	0	0	0	
'CAA'	48	0	0	0		'CAA'	48	0	0	0		'CAA'	48	0	0	0	
'GTG'	49	0	0	0		'GTG'	49	0	0	0		'GTG'	49	0	0	0	
'ATG'	50	0	0	0		'ATG'	50	0	0	0		'ATG'	50	0	0	0	
'CTG'	51	0	0	0		'CTG'	51	0	0	0		'CTG'	51	0	0	0	
'TGT'	52	0	0	0		'TGT'	52	0	0	0		'TGT'	52	0	0	0	
'GAA'	53	0	0	0		'GAA'	53	0	0	0		'GAA'	53	0	0	0	
'CCA'	54	0	0	0		'CCA'	54	0	0	0		'CCA'	54	0	0	0	
'ACC'	55	0	0	0		'ACC'	55	0	0	0		'ACC'	55	ACA	T	Sin	99.71
'ATA'	56	0	0	0		'ATA'	56	0	0	0		'ATA'	56	0	0	0	
'ATA'	57	0	0	0		'ATA'	57	0	0	0		'ATA'	57	0	0	0	
'GAA'	58	0	0	0		'GAA'	58	0	0	0		'GAA'	58	0	0	0	
'AGA'	59	0	0	0		'AGA'	59	0	0	0		'AGA'	59	0	0	0	
'AAC'	60	0	0	0		'AAC'	60	0	0	0		'AAC'	60	0	0	0	

'ATA'	61	0	0	0		'ATA'	61	0	0	0		'ATA'	61	0	0	0	
'ACA'	62	0	0	0		'ACA'	62	0	0	0		'ACA'	62	0	0	0	
'GAG'	63	0	0	0		'GAG'	63	0	0	0		'GAG'	63	0	0	0	
'ATA'	64	0	0	0		'ATA'	64	0	0	0		'ATA'	64	0	0	0	
'GTT'	65	0	0	0		'GTT'	65	0	0	0		'GTT'	65	0	0	0	
'TAT'	66	0	0	0		'TAT'	66	0	0	0		'TAT'	66	0	0	0	
'CTG'	67	0	0	0		'CTG'	67	0	0	0		'CTG'	67	0	0	0	
'ACC'	68	0	0	0		'ACC'	68	0	0	0		'ACC'	68	0	0	0	
'AAC'	69	0	0	0		'AAC'	69	0	0	0		'AAC'	69	0	0	0	
'ACC'	70	0	0	0		'ACC'	70	0	0	0		'ACC'	70	0	0	0	
'ACC'	71	0	0	0		'TCC'	71	0	0	0		'TCC'	71	ACC	T	NoSin	99.77
'ATA'	72	0	0	0		'ATA'	72	0	0	0		'ATA'	72	0	0	0	
'GAG'	73	0	0	0		'GAG'	73	0	0	0		'GAG'	73	0	0	0	
'AAG'	74	0	0	0		'AAG'	74	0	0	0		'AAG'	74	0	0	0	
'GAA'	75	0	0	0		'GAA'	75	0	0	0		'GAA'	75	0	0	0	
'ATA'	76	0	0	0		'ATA'	76	0	0	0		'ATA'	76	0	0	0	
'TGC'	77	0	0	0		'TGC'	77	0	0	0		'TGC'	77	0	0	0	
'CCC'	78	0	0	0		'CCC'	78	0	0	0		'CCC'	78	0	0	0	
'AAC'	79	0	0	0		'AAA'	79	0	0	0		'AAA'	79	0	0	0	
'CAA'	80	0	0	0		'CCA'	80	0	0	0		'CCA'	80	0	0	0	
'GCA'	81	0	0	0		'GCT'	81	0	0	0		'GCA'	81	0	0	0	
'GAA'	82	0	0	0		'GAA'	82	0	0	0		'GAA'	82	0	0	0	
'TAC'	83	0	0	0		'TAC'	83	0	0	0		'TAC'	83	0	0	0	

'AGA'	84	0	0	0		'AGA'	84	0	0	0		'AGA'	84	0	0	0	
'AAT'	85	0	0	0		'AAT'	85	0	0	0		'AAT'	85	0	0	0	
'TGG'	86	0	0	0		'TGG'	86	0	0	0		'TGG'	86	0	0	0	
'TCA'	87	0	0	0		'TCA'	87	0	0	0		'TCA'	87	0	0	0	
'AAC'	88	0	0	0		'AAA'	88	0	0	0		'AAA'	88	0	0	0	
'CCG'	89	0	0	0		'CCG'	89	0	0	0		'CCG'	89	0	0	0	
'CAA'	90	0	0	0		'CAA'	90	0	0	0		'CAA'	90	0	0	0	
'TGT'	91	0	0	0		'TGT'	91	0	0	0		'TGT'	91	0	0	0	
'GGC'	92	0	0	0		'GGC'	92	0	0	0		'GGC'	92	0	0	0	
'ATT'	93	0	0	0		'ATT'	93	0	0	0		'ATT'	93	0	0	0	
'ACA'	94	0	0	0		'ACA'	94	0	0	0		'ACA'	94	0	0	0	
'GGG'	95	0	0	0		'CGA'	95	GGA	G	NoSin	99.92	'GGA'	95	0	0	0	
'ATT'	96	0	0	0		'TTT'	96	0	0	0		'TTT'	96	0	0	0	
'GCA'	97	0	0	0		'GCA'	97	0	0	0		'GCA'	97	0	0	0	
'CCT'	98	0	0	0		'CCT'	98	0	0	0		'CCT'	98	CCC	P	Sin	0.3
'TTC'	99	0	0	0		'TTT'	99	TTC	F	Sin	99.92	'TTT'	99	TTC	F	Sin	99.95
'TCT'	100	0	0	0		'TCT'	100	0	0	0		'TCT'	100	0	0	0	
'AAG'	101	0	0	0		'AAG'	101	0	0	0		'AAG'	101	0	0	0	
'GAC'	102	0	0	0		'GAC'	102	0	0	0		'GAC'	102	0	0	0	
'AAT'	103	0	0	0		'AAT'	103	0	0	0		'AAT'	103	0	0	0	
'TCG'	104	0	0	0		'TCG'	104	0	0	0		'TCG'	104	0	0	0	
'ATT'	105	0	0	0		'ATT'	105	0	0	0		'ATT'	105	0	0	0	
'AGG'	106	0	0	0		'AGG'	106	0	0	0		'AGG'	106	0	0	0	

'CTT'	107	0	0	0		'CTT'	107	0	0	0		'CTT'	107	0	0	0	
'TCC'	108	0	0	0		'TCC'	108	0	0	0		'TCC'	108	0	0	0	
'GCT'	109	0	0	0		'GCT'	109	0	0	0		'GCT'	109	0	0	0	
'GGT'	110	0	0	0		'GGT'	110	0	0	0		'GGT'	110	0	0	0	
'GGG'	111	0	0	0		'GGG'	111	0	0	0		'GGG'	111	0	0	0	
'GAC'	112	0	0	0		'GAC'	112	0	0	0		'GAC'	112	GGC	G	NoSin	0.22
'ATC'	113	0	0	0		'ATC'	113	0	0	0		'ATC'	113	0	0	0	
'TGG'	114	0	0	0		'TGG'	114	0	0	0		'TGG'	114	0	0	0	
'GTG'	115	0	0	0		'GTG'	115	0	0	0		'GTG'	115	0	0	0	
'ACA'	116	0	0	0		'ACA'	116	0	0	0		'ACA'	116	0	0	0	
'AGA'	117	0	0	0		'AGA'	117	0	0	0		'AGA'	117	AGG	R	Sin	0.23
'GAA'	118	0	0	0		'GAA'	118	0	0	0		'GAA'	118	GGA	G	NoSin	0.23
'CCT'	119	0	0	0		'CCT'	119	0	0	0		'CCT'	119	0	0	0	
'TAT'	120	0	0	0		'TAT'	120	0	0	0		'TAT'	120	0	0	0	
'GTG'	121	0	0	0		'GTG'	121	GCG	A	NoSin	0.4	'GTG'	121	GCG	A	NoSin	0.38
'TCA'	122	0	0	0		'TCA'	122	CCA	P	NoSin	0.3	'TCA'	122	CCA	P	NoSin	0.35
'TGC'	123	0	0	0		'TGC'	123	0	0	0		'TGC'	123	0	0	0	
'GAT'	124	0	0	0		'GAT'	124	0	0	0		'GAT'	124	GGT	G	NoSin	0.22
'CCT'	125	0	0	0		'CCT'	125	0	0	0		'CCT'	125	CCC	P	Sin	0.25
'GAC'	126	0	0	0		'GAC'	126	0	0	0		'GAC'	126	0	0	0	
'AAG'	127	0	0	0		'AAG'	127	AGG	R	NoSin	0.31	'AAG'	127	AGG	R	NoSin	0.24
'TGT'	128	0	0	0		'TGT'	128	0	0	0		'TGT'	128	0	0	0	
'TAT'	129	0	0	0		'TAT'	129	0	0	0		'TAT'	129	0	0	0	

'CAA'	130	0	0	0		'CAA'	130	0	0	0		'CAA'	130	0	0	0	
'TTT'	131	0	0	0		'TTT'	131	0	0	0		'TTT'	131	0	0	0	
'GCC'	132	0	0	0		'GCC'	132	0	0	0		'GCC'	132	0	0	0	
'CTT'	133	0	0	0		'CTT'	133	0	0	0		'CTT'	133	0	0	0	
'GGA'	134	0	0	0		'GGA'	134	0	0	0		'GGA'	134	GGG	G	Sin	0.31
'CAG'	135	0	0	0		'CAG'	135	CGG	R	NoSin	0.34	'CAG'	135	CGG	R	NoSin	0.23
'GGA'	136	0	0	0		'GGA'	136	0	0	0		'GGA'	136	0	0	0	
'ACA'	137	0	0	0		'ACA'	137	0	0	0		'ACA'	137	0	0	0	
'ACA'	138	0	0	0		'ACA'	138	0	0	0		'ACA'	138	GCA/ ACG	A/T	NoSin /Sin	0.22/ 0.22
'CTA'	139	0	0	0		'CTA'	139	0	0	0		'CTA'	139	0	0	0	
'AAC'	140	0	0	0		'AAC'	140	0	0	0		'AAC'	140	AGC	S	NoSin	0.25
'AAC'	141	0	0	0		'AAC'	141	0	0	0		'AAC'	141	0	0	0	
'GTG'	142	0	0	0		'GTG'	142	0	0	0		'GTG'	142	0	0	0	
'CAT'	143	0	0	0		'CAT'	143	0	0	0		'CAT'	143	0	0	0	
'TCA'	144	0	0	0		'TCA'	144	0	0	0		'TCA'	144	0	0	0	
'AAT'	145	0	0	0		'AAT'	145	0	0	0		'AAT'	145	0	0	0	
'AAC'	146	0	0	0		'AAC'	146	AGC	S	NoSin	0.29	'AAC'	146	AGC	S	NoSin	0.22
'ACA'	147	0	0	0		'ACA'	147	GCA/ ACG	A/T	NoSin /Sin	0.31/ 0.32	'ACA'	147	GCA	A	NoSin	0.31
'GTG'	148	0	0	0		'GTG'	148	0	0	0		'GTG'	148	0	0	0	
'CGT'	149	0	0	0		'CGT'	149	CGC	R	Sin	0.34	'CGT'	149	CAT	H	NoSin	1.23
'GAT'	150	0	0	0		'GAT'	150	0	0	0		'GAT'	150	0	0	0	

'AGG'	151	0	0	0		'AGG'	151	0	0	0		'AGG'	151	0	0	0	
'ACC'	152	0	0	0		'ACC'	152	GCC	A	NoSin	0.29	'ACC'	152	0	0	0	
'CCT'	153	0	0	0		'CCT'	153	CCC	P	Sin	0.26	'CCT'	153	0	0	0	
'TAT'	154	TAC	Y	Sin	0.38	'TAT'	154	0	0	0		'TAT'	154	0	0	0	
'CGG'	155	0	0	0		'CGG'	155	0	0	0		'CGG'	155	0	0	0	
'ACT'	156	ACC	T	Sin	0.34	'ACT'	156	ACC	T	Sin	0.38	'ACT'	156	ACC	T	Sin	0.25
'CTA'	157	0	0	0		'CTA'	157	0	0	0		'CTA'	157	0	0	0	
'TTG'	158	0	0	0		'TTG'	158	0	0	0		'TTG'	158	0	0	0	
'ATG'	159	0	0	0		'ATG'	159	0	0	0		'ATG'	159	0	0	0	
'AAT'	160	0	0	0		'AAT'	160	GAT/ AGT	D/S	NoSin	0.25/ 0.25	'AAT'	160	GAT/ AGT	D/S	NoSin	0.27/ 0.23
'GAG'	161	0	0	0		'GAG'	161	GGG	G	NoSin	0.25	'GAG'	161	0	0	0	
'TTG'	162	0	0	0		'TTG'	162	0	0	0		'TTG'	162	0	0	0	
'GGT'	163	GGC	G	Sin	0.38	'GGT'	163	GGC	G	Sin	0.35	'GGT'	163	GGC	G	NoSin	0.25
'GTC'	164	GCC	A	NoSin	0.37	'GTT'	164	GCT/ GTC	A/V	NoSin /Sin	0.4/ 0.27	'GTT'	164	GCT	A	NoSin	0.35
'CCT'	165	0	0	0		'CCT'	165	CCC	P	Sin	0.36	'CCT'	165	CCC	P	Sin	0.28
'TTT'	166	0	0	0		'TTT'	166	0	0	0		'TTT'	166	0	0	0	
'CAT'	167	0	0	0		'CAT'	167	0	0	0		'CAT'	167	0	0	0	
'CTG'	168	0	0	0		'CTG'	168	0	0	0		'CTG'	168	0	0	0	
'GGG'	169	0	0	0		'GGG'	169	0	0	0		'GGG'	169	0	0	0	
'ACC'	170	GCC	A	NoSin	0.31	'ACC'	170	GCC	A	NoSin	0.38	'ACC'	170	GCC	A	NoSin	0.39
'AAG'	171	AGG	R	NoSin	0.3	'AAG'	171	0	0	0		'AAG'	171	AGG	R	NoSin	0.33

'CAA'	172	0	0	0		'CAA'	172	0	0	0		'CAA'	172	CAG	Q	Sin	0.25
'GTG'	173	0	0	0		'GTG'	173	0	0	0		'GTG'	173	0	0	0	
'TGC'	174	0	0	0		'TGC'	174	CGC	R	NoSin	0.26	'TGC'	174	CGC/ TGT	R/C	NoSin /Sin	0.26/ 0.24
'ATA'	175	0	0	0		'ATA'	175	0	0	0		'ATA'	175	0	0	0	
'GCA'	176	0	0	0		'GCA'	176	0	0	0		'GCA'	176	0	0	0	
'TGG'	177	0	0	0		'TGG'	177	0	0	0		'TGG'	177	0	0	0	
'TCC'	178	CCC	P	NoSin	0.28	'TCC'	178	CCC	P	NoSin	0.32	'TCC'	178	CCC	P	NoSin	0.33
'AGC'	179	0	0	0		'AGC'	179	0	0	0		'AGC'	179	0	0	0	
'TCA'	180	CCA	P	NoSin	0.34	'TCA'	180	CCA	P	NoSin	0.4	'TCA'	180	CCA	P	NoSin	0.34
'AGT'	181	0	0	0		'AGT'	181	0	0	0		'AGT'	181	GGT	G	NoSin	0.23
'TGT'	182	0	0	0		'TGT'	182	CGT	R	NoSin	0.31	'TGT'	182	0	0	0	
'CAC'	183	0	0	0		'CAC'	183	CGC	R	NoSin	0.32	'CAC'	183	0	0	0	
'GAT'	184	0	0	0		'GAT'	184	0	0	0		'GAT'	184	0	0	0	
'GGA'	185	0	0	0		'GGA'	185	0	0	0		'GGA'	185	0	0	0	
'AAA'	186	GAA/ AAG	E/K	NoSin /Sin	0.3/ 0.47	'AAA'	186	AGA/ AAG	R/K	NoSin /Sin	0.35/ 0.43	'AAA'	186	AAG	K	Sin	0.55
'GCA'	187	0	0	0		'GCA'	187	0	0	0		'GCA'	187	0	0	0	
'TGG'	188	0	0	0		'TGG'	188	0	0	0		'TGG'	188	0	0	0	
'CTG'	189	CCG/ CTA	P/L	NoSin /Sin	0.49/ 0.32	'CTG'	189	CCG	P	NoSin	0.37	'CTG'	189	CCG/ CTA	P/L	NoSin /Sin	0.38/ 0.42
'CAT'	190	CGT/ CAC	R/H	NoSin /Sin	0.32/ 0.32	'CAT'	190	CGT/ CAC	R/H	NoSin /Sin	0.63/ 0.5	'CAT'	190	CGT/ CAC	R/H	NoSin /Sin	0.61/ 0.31
'GTT'	191	GCT	A	NoSin	0.51	'GTT'	191	GCT/	A/V	NoSin	0.6/	'GTT'	191	GCT/	A/V	NoSin	0.41/

								GTC		/Sin	0.33			GTC		/Sin	0.32
'TGT'	192	TGC	C	Sin	0.69	'TGT'	192	CGT/ TGC	R/C	NoSin /Sin	0.6/ 0.59	'TGT'	192	CGT/ TGC	R/C	NoSin /Sin	0.41/ 0.76
'ATA'	193	0	0	0		'ATA'	193	GTA	V	NoSin	0.28	'ATA'	193	ATG	M	NoSin	0.3
'ACG'	194	GCG	A	NoSin	0.55	'ACG'	194	GCG	A	NoSin	0.57	'ACG'	194	GCG	A	NoSin	0.87
'GGG'	195	0	0	0		'GGG'	195	0	0	0		'GGG'	195	0	0	0	
'GAT'	196	GGT	G	NoSin	0.29	'GAT'	196	GGT/ GAC	G/D	NoSin /Sin	0.29/ 0.31	'GAT'	196	GGT	G	NoSin	0.24
'GAT'	197	GGT	G	NoSin	0.29	'GAT'	197	GGT	G	NoSin	0.29	'GAT'	197	GGT	G	NoSin	0.31
'AAA'	198	0	0	0		'AAA'	198	0	0	0		'AAA'	198	0	0	0	
'AAT'	199	0	0	0		'AAT'	199	0	0	0		'AAT'	199	0	0	0	
'GCA'	200	0	0	0		'GCA'	200	GCG	A	Sin	0.34	'GCA'	200	0	0	0	
'ACT'	201	0	0	0		'ACT'	201	GCT	A	NoSin	0.26	'ACT'	201	GCT	A	NoSin	0.25
'GCT'	202	0	0	0		'GCT'	202	0	0	0		'GCT'	202	0	0	0	
'AGC'	203	GGC	G	NoSin	0.33	'AGC'	203	0	0	0		'AGC'	203	GGC	G	NoSin	0.25
'TTC'	204	CTC	L	NoSin	0.3	'TTC'	204	CTC	L	NoSin	0.45	'TTC'	204	CTC	L	NoSin	0.35
'ATT'	205	0	0	0		'ATT'	205	0	0	0		'ATT'	205	0	0	0	
'TAC'	206	0	0	0		'TAC'	206	0	0	0		'TAC'	206	0	0	0	
'AAT'	207	AGT	S	NoSin	0.35	'AAT'	207	AGT	S	NoSin	0.26	'AAT'	207	AGT	S	NoSin	0.33
'GGG'	208	0	0	0		'GGG'	208	0	0	0		'GGG'	208	0	0	0	
'AGG'	209	GGG	G	NoSin	0.33	'AGG'	209	GGG	G	NoSin	0.32	'AGG'	209	GGG	G	NoSin	0.29
'CTT'	210	CCT	P	NoSin	0.33	'CTT'	210	CCT	P	NoSin	0.37	'CTT'	210	CCT	P	NoSin	0.29
'GTA'	211	GTG	V	Sin	0.27	'GTA'	211	0	0	0		'GTA'	211	GTG	V	Sin	0.23

'GAT'	212	0	0	0		'GAT'	212	GGT	G	NoSin	0.27	'GAT'	212	GGT	G	NoSin	0.24
'AGT'	213	AGC	S	Sin	0.38	'AGT'	213	AGC	S	Sin	0.31	'AGT'	213	AGC	S	Sin	0.27
'GTT'	214	GCT	A	NoSin	0.3	'GTT'	214	GCT	A	NoSin	0.26	'GTT'	214	0	0	0	
'GTT'	215	GCT	A	NoSin	0.42	'GTT'	215	GCT	A	NoSin	0.41	'GTT'	215	GCT	A	NoSin	0.4
'TCA'	216	TCG	S	Sin	0.28	'TCA'	216	TCG	S	NoSin	0.3	'TCA'	216	CCA/ TCG	P/S	NoSin /Sin	0.22/ 0.22
'TGG'	217	CGG	R	NoSin	0.28	'TGG'	217	0	0	0		'TGG'	217	CGG	R	NoSin	0.23
'TCC'	218	0	0	0		'TCC'	218	CCC	P	NoSin	0.26	'TCC'	218	CCC	P	NoSin	0.23
'AAA'	219	0	0	0		'AAA'	219	AAG	K	Sin	0.26	'AAA'	219	AAG	K	NoSin	0.29
'GAA'	220	0	0	0		'GAA'	220	0	0	0		'GAA'	220	0	0	0	
'ATT'	221	0	0	0		'ATT'	221	0	0	0		'ATT'	221	0	0	0	
'CTC'	222	0	0	0		'CTC'	222	0	0	0		'CTC'	222	0	0	0	
'AGG'	223	GGG	G	NoSin	0.25	'AGG'	223	0	0	0		'AGG'	223	GGG	G	NoSin	0.21
'ACC'	224	0	0	0		'ACC'	224	0	0	0		'ACC'	224	GCC	A	NoSin	0.22
'CAG'	225	CGG	R	NoSin	0.25	'CAG'	225	0	0	0		'CAG'	225	CGG	R	NoSin	0.23
'GAG'	226	GGG	G	NoSin	0.32	'GAG'	226	GGG	G	NoSin	0.3	'GAG'	226	GGG	G	NoSin	0.24
'TCG'	227	0	0	0		'TCG'	227	0	0	0		'TCG'	227	CCG	P	NoSin	0.24
'GAA'	228	GGA	G	NoSin	0.25	'GAA'	228	0	0	0		'GAA'	228	0	0	0	
'TGC'	229	0	0	0		'TGC'	229	0	0	0		'TGC'	229	0	0	0	
'GTT'	230	0	0	0		'GTT'	230	0	0	0		'GTT'	230	0	0	0	
'TGT'	231	0	0	0		'TGT'	231	TGC	C	Sin	0.24	'TGT'	231	TGC	C	Sin	0.22
'ATC'	232	0	0	0		'ATC'	232	0	0	0		'ATC'	232	0	0	0	
'AAT'	233	0	0	0		'AAT'	233	0	0	0		'AAT'	233	0	0	0	

'GGA'	234	GGG	G	Sin	0.3	'GGA'	234	GGG	G	Sin	0.25	'GGA'	234	GGG	G	Sin	0.21
'ACT'	235	GCT	A	NoSin	0.29	'ACT'	235	GCT/ ACC	A/T	NoSin /Sin	0.25/ 0.23	'ACT'	235	GCT	A	NoSin	0.24
'TGT'	236	0	0	0		'TGT'	236	0	0	0		'TGT'	236	0	0	0	
'ACA'	237	0	0	0		'ACA'	237	GCA	A	NoSin	0.24	'ACA'	237	GCA	A	NoSin	0.21
'GTA'	238	0	0	0		'GTA'	238	0	0	0		'GTA'	238	0	0	0	
'GTA'	239	0	0	0		'GTA'	239	0	0	0		'GTA'	239	ATA	I	NoSin	
'ATG'	240	0	0	0		'ATG'	240	0	0	0		'ATG'	240	0	0	0	
'ACT'	241	0	0	0		'ACT'	241	GCT	A	NoSin	0.26	'ACT'	241	0	0	0	
'GAT'	242	GGT/ GAC	G/D	NoSin /Sin	0.25/ 0.26	'GAT'	242	GGT/ GAC	G/D	NoSin /Sin	0.28/ 0.26	'GAT'	242	GAC	D	NoSin	0.21
'GGA'	243	0	0	0		'GGA'	243	GGG	G	Sin	0.24	'GGA'	243	0	0	0	
'AGT'	244	GGT	G	NoSin	0.29	'AGT'	244	GGT/ AGC	G/S	NoSin /Sin	0.24/ 0.23	'AGT'	244	GGT	G	NoSin	0.2
'GCT'	245	0	0	0		'GCT'	245	GCC	A	Sin	0.25	'GCT'	245	GCC	A	Sin	0.2
'TCA'	246	TCG	S	Sin	0.23	'TCA'	246	TCG	S	Sin	0.24	'TCA'	246	TCG	S	Sin	0.22
'GGA'	247	0	0	0		'GGA'	247	0	0	0		'GGA'	247	0	0	0	
'AAA'	248	AAG	K	Sin	0.44	'AAA'	248	AAG	K	Sin	0.47	'AAA'	248	AAG	K	Sin	0.43
'GCT'	249	0	0	0		'GCT'	249	0	0	0		'GCT'	249	0	0	0	
'GAT'	250	GGT	G	NoSin	0.28	'GAT'	250	GGT	G	NoSin	0.23	'GAT'	250	GGT	G	NoSin	0.23
'ACT'	251	0	0	0		'ACT'	251	0	0	0		'ACT'	251	0	0	0	
'AAA'	252	0	0	0		'AAA'	252	0	0	0		'AAA'	252	0	0	0	
'ATA'	253	0	0	0		'ATA'	253	0	0	0		'ATA'	253	0	0	0	
'CTA'	254	0	0	0		'CTA'	254	0	0	0		'CTA'	254	0	0	0	

'TTC'	255	CTC	L	NoSin	0.28	'TTC'	255	CTC/ TCC	L/S	NoSin	0.35/ 0.23	'TTC'	255	CTC	L	NoSin	0.3
'ATT'	256	0	0	0		'ATT'	256	0	0	0		'ATT'	256	0	0	0	
'GAG'	257	GGG	G	NoSin	0.25	'GAG'	257	0	0	0		'GAG'	257	GGG	G	NoSin	0.24
'GAG'	258	0	0	0		'GAG'	258	0	0	0		'GAG'	258	0	0	0	
'GGG'	259	0	0	0		'GGG'	259	0	0	0		'GGG'	259	0	0	0	
'AAA'	260	0	0	0		'AAA'	260	0	0	0		'AAA'	260	0	0	0	
'ATC'	261	0	0	0		'ATC'	261	GTC	V	NoSin	0.23	'ATC'	261	0	0	0	
'GTT'	262	0	0	0		'GTT'	262	0	0	0		'GTT'	262	GTC	V	Sin	0.27
'CAT'	263	0	0	0		'CAT'	263	0	0	0		'CAT'	263	0	0	0	
'ACT'	264	0	0	0		'ACT'	264	0	0	0		'ACT'	264	0	0	0	
'AGC'	265	0	0	0		'AGC'	265	0	0	0		'AGC'	265	0	0	0	
'ACA'	266	0	0	0		'ACA'	266	GCA	A	NoSin	0.25	'ACA'	266	0	0	0	
'TTG'	267	0	0	0		'TTG'	267	0	0	0		'TTG'	267	0	0	0	
'TCA'	268	CCA/ TCG	P/S	NoSin /Sin	0.27/ 0.26	'TCA'	268	CCA/ TCG	P/S	NoSin /Sin	0.28/ 0.26	'TCA'	268	CCA	P	NoSin	0.24
'GGA'	269	0	0	0		'GGA'	269	0	0	0		'GGA'	269	0	0	0	
'AGT'	270	GGT/ AGC	G/S	NoSin /Sin	0.28/ 0.24	'AGT'	270	GGT	G	NoSin	0.31	'AGT'	270	GGT/ AGC	G/S	NoSin /Sin	0.2/ 0.21
'GCT'	271	0	0	0		'GCT'	271	0	0	0		'GCT'	271	GCC	A	Sin	0.23
'CAG'	272	CGG	R	NoSin	0.26	'CAG'	272	0	0	0		'CAG'	272	CGG	R	NoSin	0.2
'CAT'	273	0	0	0		'CAT'	273	0	0	0		'CAT'	273	CGT	R	NoSin	0.2
'GTC'	274	GCC	A	NoSin	0.29	'GTT'	274	GTC	V	Sin	0.36	'GTC'	274	GCC	A	NoSin	0.28
'GAA'	275	GAG	E	Sin	0.35	'GAA'	275	GAG	E	Sin	0.4	'GAA'	275	GAG	E	Sin	0.27

'GAG'	276	GGG	G	NoSin	0.23	'GAG'	276	GGG	G	NoSin	0.29	'GAG'	276	GGG	G	NoSin	0.27
'TGC'	277	0	0	0		'TGC'	277	0	0	0		'TGC'	277	CGC	R	NoSin	0.19
'TCT'	278	0	0	0		'TCT'	278	CCT	P	NoSin	0.32	'TCT'	278	CCT	P	NoSin	0.3
'TGC'	279	0	0	0		'TGC'	279	0	0	0		'TGC'	279	0	0	0	
'TAT'	280	0	0	0		'TAT'	280	CAT	H	NoSin	0.24	'TAT'	280	CAT	H	NoSin	0.23
'CCT'	281	0	0	0		'CCT'	281	CCC	P	Sin	0.22	'CCT'	281	0	0	0	
'CGA'	282	0	0	0		'CGA'	282	0	0	0		'CGA'	282	0	0	0	
'TAT'	283	0	0	0		'TAT'	283	0	0	0		'TAT'	283	0	0	0	
'CCT'	284	0	0	0		'CCT'	284	0	0	0		'CCT'	284	CCC	P	Sin	0.23
'GGT'	285	GGC	G	Sin	0.26	'GGT'	285	GGC	G	Sin	0.28	'GGT'	285	GGC	G	Sin	0.28
'GTC'	286	GCC	A	NoSin	0.27	'GTC'	286	GCC	A	NoSin	0.31	'GTC'	286	GCC	A	NoSin	0.28
'AGA'	287	0	0	0		'AGA'	287	AGG	R	Sin	0.23	'AGA'	287	GGA/ AGG	G/R	NoSin /Sin	0.23/ 0.19
'TGT'	288	CGT	R	NoSin	0.23	'TGT'	288	0	0	0		'TGT'	288	0	0	0	
'GTC'	289	GCC	A	NoSin	0.22	'GTC'	289	GCC	A	NoSin	0.32	'GTC'	289	GCC	A	NoSin	0.29
'TGC'	290	CGC	R	NoSin	0.29	'TGC'	290	CGC	R	NoSin	0.29	'TGC'	290	CGC	R	NoSin	0.28
'AGA'	291	0	0	0		'AGA'	291	GGA/ AGG	G/R	NoSin /Sin	0.23/ 0.26	'AGA'	291	GGA/ AGG	G/R	NoSin	0.24/ 0.26
'GAC'	292	GGC	G	NoSin	0.25	'GAC'	292	GGC	G	NoSin	0.3	'GAC'	292	GGC	G	NoSin	0.22
'AAC'	293	AGC	S	NoSin	0.23	'AAC'	293	0	0	0		'AAC'	293	GAC	D	NoSin	0.2
'TGG'	294	0	0	0		'TGG'	294	0	0	0		'TGG'	294	0	0	0	
'AAA'	295	AGA/ AAG	R/K	NoSin /Sin	0.3/ 0.46	'AAA'	295	AAG	K	Sin	0.46	'AAA'	295	AGA/ AAG	R/K	NoSin /Sin	0.22/ 0.45
'GGC'	296	0	0	0		'GGC'	296	0	0	0		'GGC'	296	0	0	0	

'TCC'	297	CCC	P	NoSin	0.38	'TCC'	297	CCC	P	NoSin	0.37	'TCC'	297	CCC	P	NoSin	0.32
'AAT'	298	0	0	0		'AAT'	298	0	0	0		'AAT'	298	0	0	0	
'CGG'	299	0	0	0		'CGG'	299	0	0	0		'CGG'	299	0	0	0	
'CCC'	300	0	0	0		'CCC'	300	0	0	0		'CCC'	300	0	0	0	
'ATC'	301	ACC	T	NoSin	0.28	'ATC'	301	ACC	T	NoSin	0.27	'ATC'	301	ACC	T	NoSin	0.31
'GTA'	302	0	0	0		'GTA'	302	0	0	0		'GTA'	302	0	0	0	
'GAT'	303	0	0	0		'GAT'	303	0	0	0		'GAT'	303	AAT	N	NoSin	
'ATA'	304	0	0	0		'ATA'	304	0	0	0		'ATA'	304	0	0	0	
'AAC'	305	0	0	0		'AAC'	305	0	0	0		'AAC'	305	0	0	0	
'ATA'	306	0	0	0		'ATA'	306	0	0	0		'ATA'	306	0	0	0	
'AAG'	307	0	0	0		'AAG'	307	0	0	0		'AAG'	307	AGG	R	NoSin	0.19
'GAT'	308	0	0	0		'GAT'	308	0	0	0		'GAT'	308	AAT	N	NoSin	
'CAT'	309	CGT	R	NoSin	0.26	'CAT'	309	CGT	R	NoSin	0.25	'CAT'	309	CGT/ CAC	R/H	NoSin /Sin	0.2/ 0.41
'AGC'	310	0	0	0		'AGC'	310	0	0	0		'AGC'	310	0	0	0	
'ATT'	311	0	0	0		'ATT'	311	0	0	0		'ATT'	311	0	0	0	
'GTT'	312	GCT	A	NoSin	0.31	'GTT'	312	0	0	0		'GTT'	312	GCT	A	NoSin	0.2
'TCC'	313	0	0	0		'TCC'	313	0	0	0		'TCC'	313	0	0	0	
'AGT'	314	0	0	0		'AGT'	314	AGC	S	Sin	0.25	'AGT'	314	AGC	S	Sin	0.25
'TAT'	315	0	0	0		'TAT'	315	0	0	0		'TAT'	315	TGT	C	NoSin	
'GTG'	316	GCG	A	NoSin	0.23	'GTG'	316	GCG	A	NoSin	0.27	'GTG'	316	GCG	A	NoSin	0.24
'TGT'	317	0	0	0		'TGT'	317	CGT/ TGC	R/C	NoSin /Sin	0.29/ 0.23	'TGT'	317	CGT/ TGC	R/C	NoSin /Sin	0.28/ 0.19

'TCA'	318	TCG	S	Sin	0.29	'TCA'	318	TCG	S	Sin	0.28	'TCA'	318	TCG	S	Sin	0.27
'GGA'	319	0	0	0		'GGA'	319	0	0	0		'GGA'	319	GGG	G	NoSin	0.23
'CTT'	320	0	0	0		'CTT'	320	0	0	0		'CTT'	320	0	0	0	
'GTT'	321	GCT	A	NoSin	0.27	'GTT'	321	GCT	A	NoSin	0.24	'GTT'	321	GCT	A	NoSin	0.23
'GGA'	322	GGG	G	Sin	0.41	'GGA'	322	GGG	G	Sin	0.31	'GGA'	322	GGG	G	Sin	0.29
'GAC'	323	GGC	G	NoSin	0.37	'GAC'	323	GGC	G	NoSin	0.34	'GAC'	323	GGC	G	NoSin	0.39
'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.32/ 0.33	'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.38/ 0.3	'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.25/ 0.3
'CCC'	325	0	0	0		'CCC'	325	TCC	S	NoSin	0.24	'CCC'	325	TCC/ CAC/ CCT	S/H/P	NoSin /NoSi n/Sin	0.41/ 0.28/ 0.31
'AGA'	326	0	0	0		'AGA'	326	0	0	0		'AGA'	326	0	0	0	
'AAA'	327	0	0	0		'AAA'	327	0	0	0		'AAA'	327	0	0	0	
'AAC'	328	0	0	0		'AAC'	328	0	0	0		'AAC'	328	0	0	0	
'GAC'	329	GGC	G	NoSin	0.27	'GAC'	329	GGC	G	NoSin	0.25	'GAC'	329	GGC	G	NoSin	0.21
'AGC'	330	GGC	G	NoSin	0.28	'AGC'	330	GGC	G	NoSin	0.37	'AGC'	330	GGC	G	NoSin	0.33
'TCC'	331	CCC	P	NoSin	0.27	'TCC'	331	CCC	P	NoSin	0.26	'TCC'	331	CCC	P	NoSin	0.2
'AGC'	332	GGC	G	NoSin	0.23	'AGC'	332	GGC	G	NoSin	0.25	'AGC'	332	GGC	G	NoSin	0.22
'AGT'	333	0	0	0		'AGT'	333	0	0	0		'AGT'	333	GGT	G	NoSin	0.2
'AGC'	334	0	0	0		'AGC'	334	0	0	0		'AGC'	334	0	0	0	
'CAT'	335	0	0	0		'CAT'	335	0	0	0		'CAT'	335	0	0	0	
'TGT'	336	TGC	C	Sin	0.25	'TGT'	336	TGC	C	Sin	0.24	'TGT'	336	TGC	C	Sin	0.19
'TTG'	337	0	0	0		'TTG'	337	0	0	0		'TTG'	337	0	0	0	
'GAT'	338	0	0	0		'GAT'	338	0	0	0		'GAT'	338	0	0	0	

'CCT'	339	0	0	0		'CCT'	339	0	0	0		'CCT'	339	TCT	S	NoSin	
'AAC'	340	0	0	0		'AAC'	340	0	0	0		'AAC'	340	0	0	0	
'AAT'	341	0	0	0		'AAT'	341	0	0	0		'AAT'	341	0	0	0	
'GAA'	342	GAG	E	Sin	0.4	'GAA'	342	GAG	E	Sin	0.42	'GAA'	342	GAG	E	Sin	0.36
'GAA'	343	GAG	E	Sin	0.27	'GAA'	343	GGA/ GAG	G/E	NoSin /Sin	0.23/ 0.25	'GAA'	343	GGA/ GAG	G/E	NoSin /Sin	0.24/ 0.26
'GGT'	344	0	0	0		'GGT'	344	0	0	0		'GGT'	344	GGC	G	Sin	0.2
'GGT'	345	GGC	G	NoSin	0.36	'GGT'	345	GGC	G	Sin	0.3	'GGT'	345	GGC	G	Sin	0.36
'CAT'	346	0	0	0		'CAT'	346	0	0	0		'CAT'	346	0	0	0	
'GGA'	347	GGG	G	Sin	0.33	'GGA'	347	GGG	G	Sin	0.35	'GGA'	347	GGG	G	Sin	0.31
'GTG'	348	0	0	0		'GTG'	348	GCG	A	NoSin	0.24	'GTG'	348	GCG	A	NoSin	0.22
'AAA'	349	AAG	K	Sin	0.45	'AAA'	349	AAG	K	Sin	0.39	'AAA'	349	AAG	K	Sin	0.38
'GGC'	350	0	0	0		'GGC'	350	0	0	0		'GGC'	350	0	0	0	
'TGG'	351	CGG	R	NoSin	0.27	'TGG'	351	CGG	R	NoSin	0.22	'TGG'	351	CGG	R	NoSin	0.22
'GCC'	352	0	0	0		'GCC'	352	0	0	0		'GCC'	352	0	0	0	
'TTT'	353	CTT	L	NoSin	0.29	'TTT'	353	CTT	L	NoSin	0.36	'TTT'	353	CTT	L	NoSin	0.28
'GAT'	354	0	0	0		'GAT'	354	GGT	G	NoSin	0.25	'GAT'	354	0	0	0	
'GAT'	355	GGT	G	NoSin	0.28	'GAT'	355	GGT	G	NoSin	0.3	'GAT'	355	GGT	G	NoSin	0.3
'GGA'	356	0	0	0		'GGA'	356	0	0	0		'GGA'	356	0	0	0	
'AAT'	357	0	0	0		'AAT'	357	0	0	0		'AAT'	357	0	0	0	
'GAC'	358	GGC	G	NoSin	0.34	'GAC'	358	GGC	G	NoSin	0.25	'GAC'	358	GGC	G	NoSin	0.33
'GTG'	359	0	0	0		'GTG'	359	0	0	0		'GTG'	359	GCG	A	NoSin	0.19
'TGG'	360	0	0	0		'TGG'	360	0	0	0		'TGG'	360	0	0	0	

'ATG'	361	0	0	0		'ATG'	361	0	0	0		'ATG'	361	0	0	0	
'GGA'	362	0	0	0		'GGA'	362	0	0	0		'GGA'	362	0	0	0	
'AGA'	363	0	0	0		'AGA'	363	0	0	0		'AGA'	363	0	0	0	
'ACA'	364	GCA/ ACG	A/T	NoSin /Sin	0.35/ 0.24	'ACA'	364	GCA/ ACG	A/T	NoSin /Sin	0.24/ 0.31	'ACA'	364	GCA/ ACG	A/T	NoSin /Sin	0.23/ 0.26
'ATC'	365	0	0	0		'ATC'	365	0	0	0		'ATC'	365	GTC	V	NoSin	0.2
'AAC'	366	0	0	0		'AAC'	366	0	0	0		'AAC'	366	AGC	S	NoSin	0.22
'GAG'	367	GGG	G	NoSin	0.27	'GAG'	367	GGG	G	NoSin	0.29	'GAG'	367	GGG	G	NoSin	0.21
'ACG'	368	GCG	A	NoSin	0.38	'ACG'	368	GCG	A	NoSin	0.33	'ACG'	368	GCG	A	NoSin	0.27
'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.33/ 0.25	'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.44/ 0.24	'TCA'	369	CCA/ CCG	P/P	NoSin	0.43/ 0.24
'CGC'	370	0	0	0		'CGC'	370	0	0	0		'CGC'	370	0	0	0	
'TTA'	371	0	0	0		'TTA'	371	0	0	0		'TTA'	371	0	0	0	
'GGG'	372	0	0	0		'GGG'	372	0	0	0		'GGG'	372	0	0	0	
'TAT'	373	0	0	0		'TAT'	373	0	0	0		'TAT'	373	0	0	0	
'GAA'	374	0	0	0		'GAA'	374	0	0	0		'GAA'	374	0	0	0	
'ACC'	375	GCC	A	NoSin	0.34	'ACC'	375	GCC	A	NoSin	0.4	'ACC'	375	GCC	A	NoSin	0.36
'TTC'	376	0	0	0		'TTC'	376	CTC	L	NoSin	0.27	'TTC'	376	CTC/ TCC	L/S	NoSin	0.23/ 0.21
'AAA'	377	0	0	0		'AAA'	377	0	0	0		'AAA'	377	0	0	0	
'GTC'	378	GCC	A	NoSin	0.25	'GTC'	378	GCC	A	NoSin	0.24	'GTC'	378	GCC	A	NoSin	0.32
'ATT'	379	0	0	0		'ATT'	379	0	0	0		'ATT'	379	0	0	0	
'GAA'	380	GAG	E	Sin	0.29	'GAA'	380	GAG	E	Sin	0.23	'GAA'	380	0	0	0	
'GGC'	381	0	0	0		'GGC'	381	0	0	0		'GGC'	381	0	0	0	

'TGG'	382	CGG	R	NoSin	0.25	'TGG'	382	0	0	0		'TGG'	382	CGG	R	NoSin	0.22
'TCC'	383	0	0	0		'TCC'	383	CCC	P	NoSin	0.3	'TCC'	383	CCC	P	NoSin	0.25
'AAC'	384	0	0	0		'AAC'	384	0	0	0		'AAC'	384	0	0	0	
'CCT'	385	0	0	0		'CCT'	385	0	0	0		'CCT'	385	0	0	0	
'AAG'	386	0	0	0		'AAG'	386	0	0	0		'AAG'	386	0	0	0	
'TCC'	387	0	0	0		'TCC'	387	0	0	0		'TCC'	387	CCC	P	NoSin	0.22
'AAA'	388	0	0	0		'AAA'	388	0	0	0		'AAA'	388	0	0	0	
'TTG'	389	0	0	0		'TTG'	389	0	0	0		'TTG'	389	0	0	0	
'CAG'	390	CGG	R	NoSin	0.36	'CAG'	390	CGG	R	NoSin	0.37	'CAG'	390	CGG	R	NoSin	0.34
'ATA'	391	0	0	0		'ATA'	391	0	0	0		'ATA'	391	0	0	0	
'AAT'	392	0	0	0		'AAT'	392	0	0	0		'AAT'	392	0	0	0	
'AGG'	393	0	0	0		'AGG'	393	0	0	0		'AGG'	393	GGG	G	NoSin	0.21
'CAA'	394	0	0	0		'CAA'	394	CAG	Q	Sin	0.26	'CAA'	394	CAG	Q	Sin	0.32
'GTC'	395	GCC	A	NoSin	0.3	'GTC'	395	GCC	A	NoSin	0.28	'GTC'	395	GCC	A	NoSin	0.29
'ATA'	396	0	0	0		'ATA'	396	0	0	0		'ATA'	396	0	0	0	
'GTT'	397	0	0	0		'GTT'	397	0	0	0		'GTT'	397	0	0	0	
'GAC'	398	GGC	G	NoSin	0.41	'GAC'	398	GGC	G	NoSin	0.45	'GAC'	398	GGC	G	NoSin	0.41
'AGA'	399	AGG	R	Sin	0.36	'AGA'	399	AGG	R	Sin	0.41	'AGA'	399	AGG	R	Sin	0.35
'GGT'	400	GGG	G	Sin	0.33	'GGT'	400	GGC	G	Sin	0.3	'GGT'	400	0	0	0	
'GAT'	401	GGT	G	NoSin	0.28	'GAT'	401	GGT	G	NoSin	0.34	'GAT'	401	GGT	G	NoSin	0.24
'AGG'	402	0	0	0		'AGG'	402	GGG	G	NoSin	0.36	'AGG'	402	GGG	G	NoSin	0.3
'TCC'	403	0	0	0		'TCC'	403	0	0	0		'TCC'	403	0	0	0	
'GGT'	404	0	0	0		'GGT'	404	0	0	0		'GGT'	404	0	0	0	

'TAT'	405	0	0	0		'TAT'	405	0	0	0		'TAT'	405	0	0	0	
'TCT'	406	0	0	0		'TCT'	406	0	0	0		'TCT'	406	0	0	0	
'GGT'	407	0	0	0		'GGT'	407	0	0	0		'GGT'	407	0	0	0	
'ATT'	408	0	0	0		'ATT'	408	0	0	0		'ATT'	408	0	0	0	
'TTC'	409	0	0	0		'TTC'	409	0	0	0		'TTC'	409	0	0	0	
'TCT'	410	CCT	P	NoSin	0.32	'TCT'	410	0	0	0		'TCT'	410	0	0	0	
'GTT'	411	0	0	0		'GTT'	411	0	0	0		'GTT'	411	0	0	0	
'GAA'	412	GAG	E	NoSin	0.48	'GAA'	412	GGA/ GAG	G/E	NoSin /Sin	0.32/ 0.36	'GAA'	412	GAG	E	Sin	0.45
'GGC'	413	0	0	0		'GGC'	413	0	0	0		'GGC'	413	0	0	0	
'AAA'	414	AGA	R	NoSin	0.41	'AAA'	414	AGA/ AAG	R/K	NoSin /Sin	0.57/ 0.32	'AAA'	414	AGA	R	NoSin	0.52
'AGC'	415	GGC	G	NoSin	0.43	'AGC'	415	GGC	G	NoSin	0.49	'AGC'	415	GGC	G	NoSin	0.57
'TGC'	416	0	0	0		'TGC'	416	0	0	0		'TGC'	416	CGC	R	NoSin	0.3
'ATC'	417	GTC	V	NoSin	0.37	'ATC'	417	GTC	V	NoSin	0.45	'ATC'	417	GTC	V	NoSin	0.41
'AAT'	418	0	0	0		'AAT'	418	0	0	0		'AAT'	418	0	0	0	
'CGG'	419	0	0	0		'CGG'	419	0	0	0		'CGG'	419	0	0	0	
'TGC'	420	0	0	0		'TGC'	420	0	0	0		'TGC'	420	0	0	0	
'TTC'	421	0	0	0		'TTT'	421	0	0	0		'TTT'	421	0	0	0	
'TAT'	422	0	0	0		'TAT'	422	0	0	0		'TAT'	422	0	0	0	
'GTG'	423	0	0	0		'GTG'	423	0	0	0		'GTG'	423	0	0	0	
'GAA'	424	0	0	0		'GAG'	424	0	0	0		'GAG'	424	0	0	0	
'TTG'	425	0	0	0		'TTG'	425	0	0	0		'TTG'	425	0	0	0	

'ATT'	426	0	0	0		'ATT'	426	GTT	V	NoSin	0.5	'ATT'	426	GTT	V	NoSin	0.41
'AGG'	427	0	0	0		'AGG'	427	0	0	0		'AGG'	427	0	0	0	
'GGA'	428	0	0	0		'GGA'	428	0	0	0		'GGA'	428	0	0	0	
'AGA'	429	0	0	0		'AGA'	429	0	0	0		'AGA'	429	0	0	0	
'AAA'	430	AAG	K	Sin	0.68	'AAA'	430	AAG	K	Sin	0.81	'AAA'	430	AAG	K	Sin	0.58
'GAG'	431	0	0	0		'GAG'	431	0	0	0		'GAG'	431	0	0	0	
'GAA'	432	0	0	0		'GAA'	432	GAG	E	Sin	0.62	'GAA'	432	GGA/ GAG	G/E	NoSin /Sin	0.44/ 0.68
'ACT'	433	0	0	0		'ACT'	433	0	0	0		'ACT'	433	GCT	A	NoSin	0.43
'GAA'	434	GAG	E	Sin	0.73	'GAA'	434	GAG	E	Sin	1.22	'GAA'	434	GAG	E	Sin	0.54
'GTC'	435	0	0	0		'GTC'	435	0	0	0		'GTC'	435	0	0	0	
'TTG'	436	0	0	0		'TTG'	436	0	0	0		'TTG'	436	CTG	L	Sin	0.49
'TGG'	437	0	0	0		'TGG'	437	0	0	0		'TGG'	437	0	0	0	
'ACC'	438	0	0	0		'ACC'	438	0	0	0		'ACC'	438	0	0	0	
'TCA'	439	0	0	0		'TCA'	439	0	0	0		'TCA'	439	TTA	L	NoSin	0.72
'AAC'	440	0	0	0		'AAC'	440	AGC	S	NoSin	1.42	'AAC'	440	AGC	S	NoSin	1.7
'AGT'	441	0	0	0		'AGT'	441	0	0	0		'AGT'	441	0	0	0	
'ATT'	442	0	0	0		'ATT'	442	0	0	0		'ATT'	442	0	0	0	

Muestra 224						Muestra 250						Muestra 264					
Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%	Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%	Codon	nºcodon	Var nt	VarAA	Sin ó NoSin	Frec%
'ATG'	23	0	0	0								'ATG'	23	0	0	0	
'CAA'	24	0	0	0								'CAA'	24	0	0	0	
'ATT'	25	0	0	0								'ATT'	25	0	0	0	
'GCC'	26	0	0	0								'GCC'	26	0	0	0	
'ATC'	27	0	0	0								'ATC'	27	0	0	0	
'TTG'	28	0	0	0								'TTG'	28	0	0	0	
'ATA'	29	0	0	0								'ATA'	29	0	0	0	
'ACT'	30	0	0	0		'ACT'	30	0	0	0		'ACT'	30	0	0	0	
'ACT'	31	0	0	0		'ACT'	31	0	0	0		'ACT'	31	0	0	0	
'GTA'	32	0	0	0		'GTA'	32	GCA	A	NoSin	0.83	'GTA'	32	0	0	0	
'ACA'	33	0	0	0		'ACA'	33	0	0	0		'ACA'	33	0	0	0	
'TTG'	34	0	0	0		'TTG'	34	CTG/ TCG	L/S	Sin/ NoSin	0.93/ 0.84	'TTG'	34	CTG	L	Sin	0.94
'CAT'	35	0	0	0		'CAT'	35	CAC	H	Sin	0.54	'CAT'	35	0	0	0	
'TTC'	36	0	0	0		'TTC'	36	0	0	0		'TTC'	36	0	0	0	
'AAG'	37	0	0	0		'AAG'	37	0	0	0		'AAG'	37	0	0	0	

'CAA'	38	0	0	0		'CAA'	38	0	0	0		'CAA'	38	0	0	0	
'TAT'	39	0	0	0		'TAT'	39	0	0	0		'TAT'	39	0	0	0	
'GAA'	40	0	0	0		'GAA'	40	0	0	0		'GAA'	40	0	0	0	
'TTC'	41	0	0	0		'TTC'	41	0	0	0		'TTC'	41	CTC/ TGC	L/C	NoSin	1.13/ 5.21
'AAC'	42	0	0	0		'AAC'	42	GAC	D	NoSin	0.39	'AAC'	42	0	0	0	
'TCC'	43	0	0	0		'TCC'	43	CCC	P	NoSin	0.57	'TCC'	43	CCC	P	NoSin	0.72
'CCC'	44	0	0	0		'CCC'	44	0	0	0		'CCC'	44	0	0	0	
'CCA'	45	0	0	0		'CCA'	45	0	0	0		'CCA'	45	0	0	0	
'AAC'	46	0	0	0		'AAC'	46	0	0	0		'AAC'	46	0	0	0	
'AAC'	47	0	0	0		'AAC'	47	0	0	0		'AAC'	47	0	0	0	
'CAA'	48	0	0	0		'CAA'	48	0	0	0		'CAA'	48	0	0	0	
'GTG'	49	0	0	0		'GTG'	49	GCG	A	NoSin	0.4	'GTG'	49	GCG	A	NoSin	0.5
'ATG'	50	0	0	0		'ATG'	50	0	0	0		'ATG'	50	0	0	0	
'CTG'	51	0	0	0		'CTG'	51	CCG	P	NoSin	0.4	'CTG'	51	CCG	P	NoSin	0.49
'TGT'	52	0	0	0		'TGT'	52	TGC	C	Sin	0.33	'TGT'	52	TGC	C	Sin	0.41
'GAA'	53	0	0	0		'GAA'	53	GGA	G	NoSin	0.31	'GAA'	53	GAG	E	Sin	0.43
'CCA'	54	0	0	0		'CCA'	54	0	0	0		'CCA'	54	0	0	0	
'ACC'	55	ACA	T	Sin	99.87	'ACA'	55	GCA	A	NoSin	0.48	'ACA'	55	0	0	0	
'ATA'	56	0	0	0		'ATA'	56	0	0	0		'ATA'	56	0	0	0	
'ATA'	57	0	0	0		'ATA'	57	ACA	T	NoSin	0.28	'ATA'	57	0	0	0	
'GAA'	58	0	0	0		'GAA'	58	GGA	G	NoSin	0.38	'GAA'	58	0	0	0	
'AGA'	59	0	0	0		'AGA'	59	GGA	G	NoSin	0.41	'AGA'	59	AGG	R	Sin	0.35

'AAC'	60	0	0	0		'AAC'	60	AGC	S	NoSin	0.46	'AAC'	60	AGC	S	NoSin	0.43
'ATA'	61	0	0	0		'ATA'	61	0	0	0		'ATA'	61	0	0	0	
'ACA'	62	0	0	0		'ACA'	62	ACG	T	Sin	0.26	'ACA'	62	0	0	0	
'GAG'	63	0	0	0		'GAG'	63	GGG	G	NoSin	0.4	'GAG'	63	GGG	G	NoSin	0.41
'ATA'	64	0	0	0		'ATA'	64	ATG	M	NoSin	0.26	'ATA'	64	0	0	0	
'GTT'	65	0	0	0		'GTG'	65	0	0	0		'GTG'	65	0	0	0	
'TAT'	66	0	0	0		'TAT'	66	CAT	H	NoSin	0.25	'TAT'	66	0	0	0	
'CTG'	67	0	0	0		'CTG'	67	CCG/ CTA	P/L	NoSin /Sin	0.22/ 1.86	'TTG'	67	0	0	0	
'ACC'	68	0	0	0		'ACC'	68	GCC	A	NoSin	0.24	'ACC'	68	GCC	A	NoSin	0.31
'AAC'	69	0	0	0		'AAC'	69	AGC	S	NoSin	0.27	'AAC'	69	AGC	S	NoSin	0.37
'ACC'	70	0	0	0		'ACC'	70	GCC	A	NoSin	0.25	'ACC'	70	GCC	A	NoSin	0.29
'TCC'	71	ACC	T	NoSin	99.75	'ACC'	71	GCC	A	NoSin	0.5	'ACC'	71	0	0	0	
'ATA'	72	0	0	0		'ATA'	72	GTA/ ATG	V/M	NoSin	0.23/ 0.25	'ATA'	72	0	0	0	
'GAG'	73	0	0	0		'GAG'	73	0	0	0		'GAG'	73	0	0	0	
'AAG'	74	0	0	0		'AAG'	74	0	0	0		'AAG'	74	AGG	R	NoSin	0.28
'GAA'	75	0	0	0		'GAA'	75	GGA	G	NoSin	0.22	'GAA'	75	0	0	0	
'ATA'	76	0	0	0		'ATA'	76	GTA	V	NoSin	0.28	'ATA'	76	0	0	0	
'TGC'	77	0	0	0		'TGC'	77	0	0	0		'TGC'	77	0	0	0	
'CCC'	78	0	0	0		'CCC'	78	0	0	0		'CCC'	78	0	0	0	
'AAA'	79	0	0	0		'AAA'	79	0	0	0		'AAA'	79	0	0	0	
'CCA'	80	0	0	0		'CCA'	80	0	0	0		'CCA'	80	0	0	0	

'GCA'	81	0	0	0		'GCA'	81	ACA/ GCG	T/A	NoSin /Sin	0.87/ 0.21	'GCA'	81	0	0	0	
'GAA'	82	0	0	0		'GAA'	82	0	0	0		'GAA'	82	0	0	0	
'TAC'	83	0	0	0		'TAC'	83	0	0	0		'TAC'	83	TGC	C	NoSin	0.33
'AGA'	84	0	0	0		'AGA'	84	GGA	G	NoSin	0.27	'AGA'	84	0	0	0	
'AAT'	85	0	0	0		'AAT'	85	0	0	0		'AAT'	85	0	0	0	
'TGG'	86	0	0	0		'TGG'	86	0	0	0		'TGG'	86	0	0	0	
'TCA'	87	0	0	0		'TCA'	87	CCA	P	NoSin	0.27	'TCA'	87	CCA	P	NoSin	0.29
'AAA'	88	0	0	0		'AAA'	88	AAG	K	Sin	0.28	'AAA'	88	0	0	0	
'CCG'	89	0	0	0		'CCG'	89	0	0	0		'CCG'	89	0	0	0	
'CAA'	90	0	0	0		'CAA'	90	CGA	R	NoSin	1.3	'CAA'	90	0	0	0	
'TGT'	91	0	0	0		'TGT'	91	TGC	C	Sin	0.41	'TGT'	91	TGC	C	Sin	0.54
'GGC'	92	0	0	0		'GGC'	92	0	0	0		'GGC'	92	0	0	0	
'ATT'	93	0	0	0		'ATT'	93	0	0	0		'ATT'	93	0	0	0	
'ACA'	94	0	0	0		'ACA'	94	ACG	T	Sin	0.35	'ACA'	94	ACG	T	Sin	0.33
'GGA'	95	0	0	0		'GGA'	95	0	0	0		'GGA'	95	0	0	0	
'TTT'	96	0	0	0		'TTT'	96	0	0	0		'TTT'	96	0	0	0	
'GCA'	97	0	0	0		'GCA'	97	GCG	A	Sin	0.2	'GCA'	97	GCG	A	Sin	0.29
'CCT'	98	0	0	0		'CCT'	98	CCC	P	Sin	0.63	'CCT'	98	CCC	P	Sin	0.36
'TTT'	99	TTC	F	Sin	99.92	'TTC'	99	0	0	0		'TTC'	99	0	0	0	
'TCT'	100	0	0	0		'TCT'	100	CCT	P	NoSin	0.26	'TCT'	100	0	0	0	
'AAG'	101	0	0	0		'AAG'	101	AGG	R	NoSin	0.2	'AAG'	101	AGG	R	NoSin	0.26
'GAC'	102	0	0	0		'GAC'	102	GGC	G	NoSin	0.43	'GAC'	102	GGC	G	NoSin	0.3

'AAT'	103	0	0	0		'AAT'	103	0	0	0		'AAT'	103	0	0	0	
'TCG'	104	0	0	0		'TCG'	104	0	0	0		'TCG'	104	0	0	0	
'ATT'	105	0	0	0		'ATT'	105	0	0	0		'ATT'	105	0	0	0	
'AGG'	106	0	0	0		'AGG'	106	0	0	0		'AGG'	106	0	0	0	
'CTT'	107	0	0	0		'CTT'	107	CCT	P	NoSin	0.2	'CTT'	107	0	0	0	
'TCC'	108	0	0	0		'TCC'	108	0	0	0		'TCC'	108	0	0	0	
'GCT'	109	0	0	0		'GCT'	109	ACT/ GCC	T/A	NoSin /Sin	1.97/ 0.19	'GCT'	109	0	0	0	
'GGT'	110	0	0	0		'GGT'	110	0	0	0		'GGT'	110	0	0	0	
'GGG'	111	0	0	0		'GGG'	111	0	0	0		'GGG'	111	0	0	0	
'GAC'	112	0	0	0		'GAC'	112	GGC	G	NoSin	0.29	'GAC'	112	GGC	G	NoSin	0.3
'ATC'	113	0	0	0		'ATC'	113	0	0	0		'ATC'	113	0	0	0	
'TGG'	114	0	0	0		'TGG'	114	0	0	0		'TGG'	114	0	0	0	
'GTG'	115	GCG	A	NoSin	0.25	'GTG'	115	GCG	A	NoSin	0.23	'GTG'	115	0	0	0	
'ACA'	116	GCA	A	NoSin	0.24	'ACA'	116	GCA/ ACG	A/T	NoSin /Sin	0.34/ 0.33	'ACA'	116	GCA	A	NoSin	0.32
'AGA'	117	0	0	0		'AGA'	117	GGA/ AGG	G/R	NoSin /Sin	0.3/ 0.24	'AGA'	117	GGA/ AGG	G/R	NoSin /Sin	0.31/ 0.28
'GAA'	118	GGA	G	NoSin	0.25	'GAA'	118	GGA/ GAG	G/E	NoSin /Sin	0.33/ 0.27	'GAA'	118	GGA/ GAG	G/E	NoSin	0.35/ 0.28
'CCT'	119	0	0	0		'CCT'	119	TCT/ CCC	S/P	NoSin /Sin	0.24/ 0.22	'CCT'	119	0	0	0	
'TAT'	120	0	0	0		'TAT'	120	0	0	0		'TAT'	120	0	0	0	
'GTG'	121	GCG	A	NoSin	0.33	'GTG'	121	GCG	A	NoSin	0.41	'GTG'	121	GCG	A	NoSin	0.52

'TCA'	122	CCA	P	NoSin	0.29	'TCA'	122	CCA	P	NoSin	0.39	'TCA'	122	CCA/ TCG	P/S	NoSin /Sin	0.37/ 0.29
'TGC'	123	0	0	0		'TGC'	123	0	0	0		'TGC'	123	CGC	R	NoSin	1.67
'GAT'	124	0	0	0		'GAT'	124	GGT	G	NoSin	0.21	'GAT'	124	GGT	G	NoSin	0.26
'CCT'	125	CCC	P	Sin	0.31	'CCC'	125	0	0	0		'CCT'	125	0	0	0	
'GAC'	126	GGC	G	NoSin	0.24	'GAC'	126	GGC	G	NoSin	0.33	'GAC'	126	GGC	G	NoSin	
'AAG'	127	0	0	0		'AAG'	127	AGG	R	NoSin	0.24	'AAG'	127	AGG	R	NoSin	0.29
'TGT'	128	0	0	0		'TGT'	128	TGC	C	Sin	0.21	'TGT'	128	0	0	0	
'TAT'	129	0	0	0		'TAT'	129	TAC	Y	Sin	0.21	'TAT'	129	0	0	0	
'CAA'	130	0	0	0		'CAA'	130	0	0	0		'CAA'	130	0	0	0	
'TTT'	131	0	0	0		'TTT'	131	0	0	0		'TTT'	131	0	0	0	
'GCC'	132	0	0	0		'GCC'	132	0	0	0		'GCC'	132	GCT	A	Sin	0.34
'CTT'	133	0	0	0		'CTT'	133	CCT/ CTC	P/L	NoSin /Sin	0.21/ 0.57	'CTT'	133	CCT	P	NoSin	0.33
'GGA'	134	GGG	G	NoSin	0.29	'GGA'	134	GGG	G	Sin	0.36	'GGA'	134	GGG	G	Sin	0.58
'CAG'	135	CGG	R	NoSin	0.22	'CAG'	135	CGG	R	NoSin	0.28	'CAG'	135	CGG	R	NoSin	0.36
'GGA'	136	GGG	G	Sin	0.27	'GGA'	136	GGG	G	Sin	0.19	'GGA'	136	0	0	0	
'ACA'	137	GCA	A	NoSin	0.31	'ACA'	137	GCA/ ACG	A/T	NoSin /Sin	0.33/ 0.22	'ACA'	137	0	0	0	
'ACA'	138	ACG	T	Sin	0.25	'ACA'	138	GCA/ ACG	A/T	NoSin /Sin	0.29/ 0.21	'ACA'	138	0	0	0	
'CTA'	139	0	0	0		'CTA'	139	0	0	0		'CTA'	139	0	0	0	
'AAC'	140	AGC	S	NoSin	0.23	'AAC'	140	AGC	S	NoSin	0.29	'AAC'	140	AGC	S	NoSin	0.36
'AAC'	141	0	0	0		'AAC'	141	0	0	0		'AAC'	141	GAC/	D/S	NoSin	0.6/

													AGC			0.29
'GTG'	142	0	0	0		'GTG'	142	0	0	0		'GTG'	142	0	0	0	
'CAT'	143	0	0	0		'CAT'	143	0	0	0		'CAT'	143	0	0	0	
'TCA'	144	0	0	0		'TCA'	144	0	0	0		'TCA'	144	0	0	0	
'AAT'	145	0	0	0		'AAT'	145	0	0	0		'AAT'	145	0	0	0	
'AAC'	146	0	0	0		'AAC'	146	AGC	S	NoSin	0.25	'AAC'	146	0	0	0	
'ACA'	147	GCA	A	NoSin	0.32	'ACA'	147	GCA	A	NoSin	0.21	'ACC'	147	GCC	A	NoSin	0.32
'GTG'	148	0	0	0		'GTA'	148	GTG	V	Sin	0.2	'GTA'	148	0	0	0	
'CGT'	149	0	0	0		'CGT'	149	CGC	R	Sin	0.21	'CGT'	149	0	0	0	
'GAT'	150	GGT	G	NoSin	0.24	'GAT'	150	GGT	G	NoSin	0.24	'GAT'	150	GGT	G	NoSin	0.35
'AGG'	151	GGG	G	NoSin	0.24	'AGG'	151	GGG	G	NoSin	0.21	'AGG'	151	GGG	G	NoSin	0.3
'ACC'	152	GCC	A	NoSin	0.25	'ACC'	152	GCC/ ATC	A/I	NoSin	0.38/ 1.66	'ACC'	152	0	0	0	
'CCT'	153	CCC	P	Sin	0.24	'CCT'	153	CCC	P	Sin	0.19	'CCT'	153	0	0	0	
'TAT'	154	0	0	0		'TAT'	154	CAT	H	NoSin	0.32	'TAT'	154	0	0	0	
'CGG'	155	0	0	0		'CGG'	155	0	0	0		'CGG'	155	0	0	0	
'ACT'	156	GCT/ ACC	A/T	NoSin /Sin	0.22/ 0.32	'ACT'	156	ACC	T	Sin	0.26	'ACT'	156	ACC	T	Sin	0.34
'CTA'	157	0	0	0		'CTA'	157	0	0	0		'CTA'	157	CCA	P	NoSin	0.71
'TTG'	158	0	0	0		'TTG'	158	0	0	0		'TTG'	158	0	0	0	
'ATG'	159	0	0	0		'ATG'	159	0	0	0		'ATG'	159	GTG	V	NoSin	0.29
'AAT'	160	GAT	D	NoSin	0.25	'AAT'	160	GAT/ AGT	D/S	NoSin	0.27/ 0.23	'AAT'	160	AGT	S	NoSin	0.56
'GAG'	161	GGG	G	NoSin	0.23	'GAG'	161	GGG	G	NoSin	0.27	'GAG'	161	GGG	G	NoSin	0.28

'TTG'	162	0	0	0		'TTG'	162	0	0	0		'TTG'	162	0	0	0	
'GGT'	163	GGC	G	Sin	0.25	'GGT'	163	GGC	G	Sin	0.31	'GGT'	163	0	0	0	
'GTT'	164	GCT	A	NoSin	0.37	'GTT'	164	GCT/ GTC	A/V	NoSin /Sin	0.26/ 0.95	'GTT'	164	GCT	A	NoSin	0.28
'CCT'	165	CCC	P	Sin	0.34	'CCT'	165	CTT/ CCC	L/P	NoSin /Sin	0.47/ 0.28	'CCT'	165	CCC	P	Sin	0.28
'TTT'	166	0	0	0		'TTT'	166	CTT	L	NoSin	0.19	'TTT'	166	0	0	0	
'CAT'	167	0	0	0		'CAT'	167	0	0	0		'CAT'	167	0	0	0	
'CTG'	168	0	0	0		'CTG'	168	0	0	0		'CTG'	168	0	0	0	
'GGG'	169	0	0	0		'GGG'	169	AGG	R	NoSin	0.23	'GGG'	169	0	0	0	
'ACC'	170	GCC	A	NoSin	0.32	'ACC'	170	GCC/ AAC	A/N	NoSin	0.26/ 144	'ACC'	170	0	0	0	
'AAG'	171	AGG	R	NoSin	0.27	'AAG'	171	GAG/ AGG	E/R	NoSin	0.2/ 0.45	'AAG'	171	GAG/ AGG	E/R	NoSin	0.3/ 0.3
'CAA'	172	CAG	Q	Sin	0.26	'CAA'	172	CAG	Q	Sin	0.23	'CAA'	172	0	0	0	
'GTG'	173	0	0	0		'GTG'	173	ATG	M	NoSin	0.7	'GTG'	173	0	0	0	
'TGC'	174	CGC	R	NoSin	0.28	'TGC'	174	CGC	R	NoSin	0.26	'TGC'	174	0	0	0	
'ATA'	175	0	0	0		'ATA'	175	0	0	0		'ATA'	175	0	0	0	
'GCA'	176	0	0	0		'GCA'	176	GCG	A	Sin	0.35	'GCA'	176	0	0	0	
'TGG'	177	0	0	0		'TGG'	177	0	0	0		'TGG'	177	0	0	0	
'TCC'	178	CCC	P	NoSin	0.38	'TCC'	178	CCC	P	NoSin	0.26	'TCC'	178	CCC	P	NoSin	0.58
'AGC'	179	0	0	0		'AGC'	179	GGC	G	NoSin	0.22	'AGC'	179	0	0	0	
'TCA'	180	CCA	P	NoSin	0.38	'TCA'	180	CCA	P	NoSin	0.26	'TCA'	180	CCA	P	NoSin	0.51
'AGT'	181	0	0	0		'AGT'	181	GGT	G	NoSin	0.28	'AGT'	181	0	0	0	

'TGT'	182	0	0	0		'TGT'	182	CGT	R	NoSin	0.27	'TGT'	182	0	0	0	
'CAC'	183	CGC	R	NoSin	0.24	'CAC'	183	CGC	R	NoSin	0.25	'CAC'	183	0	0	0	
'GAT'	184	0	0	0		'GAC'	184	GGC	G	NoSin	0.2	'GAT'	184	0	0	0	
'GGA'	185	0	0	0		'GGA'	185	AGA	R	NoSin	1.43	'GGA'	185	0	0	0	
'AAA'	186	AAG	K	Sin	0.45	'AAA'	186	GAA/ AAG	E/K	NoSin /Sin	0.24/ 0.57	'AAA'	186	GAA/ AAG	E/K	NoSin /Sin	0.3/ 0.64
'GCA'	187	0	0	0		'GCA'	187	GCG	A	Sin	0.21	'GCA'	187	0	0	0	
'TGG'	188	0	0	0		'TGG'	188	0	0	0		'TGG'	188	0	0	0	
'CTG'	189	CCG	P	NoSin	0.46	'CTG'	189	CCG/ CTA	P/L	NoSin /Sin	0.36/ 0.26	'CTG'	189	CCG	P	NoSin	1.69
'CAT'	190	CGT	R	NoSin	0.39	'CAT'	190	CGT/ CAC	R/H	NoSin /Sin	0.55/ 0.34	'CAT'	190	CGT	R	NoSin	0.32
'GTT'	191	GCT	A	NoSin	0.36	'GTT'	191	GCT/ GTC	A/V	NoSin /Sin	0.57/ 0.3	'GTT'	191	0	0	0	
'TGT'	192	CGT/ TGC	R/C	NoSin /Sin	0.35/ 0.45	'TGT'	192	CGT/ TGC	R/C	NoSin /Sin	0.43/ 0.68	'TGT'	192	TGC	C	Sin	0.4
'ATA'	193	ATG	M	NoSin	0.28	'ATA'	193	GTA/ ATG	V/M	NoSin	0.25/ 0.32	'ATA'	193	0	0	0	
'ACG'	194	GCG	A	NoSin	0.51	'ACG'	194	GCG	A	NoSin	1.17	'ACG'	194	GCG	A	NoSin	0.64
'GGG'	195	0	0	0		'GGG'	195	0	0	0		'GGG'	195	0	0	0	
'GAT'	196	GGT	G	NoSin	0.25	'GAT'	196	GGT	G	NoSin	0.27	'GAT'	196	0	0	0	
'GAT'	197	GGT	G	NoSin	0.31	'GAT'	197	GGT	G	NoSin	0.38	'GAT'	197	GGT	G	NoSin	0.3
'AAA'	198	0	0	0		'AAA'	198	0	0	0		'AAA'	198	0	0	0	
'AAT'	199	0	0	0		'AAT'	199	0	0	0		'AAT'	199	0	0	0	

'GCA'	200	GCG	A	Sin	0.23	'GCA'	200	GCG	A	Sin	0.27	'GCA'	200	0	0	0	
'ACT'	201	GCT	A	NoSin	0.23	'ACT'	201	GCT	A	NoSin	0.25	'ACT'	201	GCT/ ACC	A/T	NoSin /Sin	0.29/ 0.36
'GCT'	202	0	0	0		'GCT'	202	0	0	0		'GCT'	202	0	0	0	
'AGC'	203	GGC	G	NoSin	0.25	'AGC'	203	GGC	G	NoSin	0.29	'AGC'	203	0	0	0	
'TTC'	204	CTC	S	NoSin	0.29	'TTC'	204	CTC	L	NoSin	0.27	'TTC'	204	CTC	L	NoSin	0.29
'ATT'	205	0	0	0		'ATT'	205	0	0	0		'ATT'	205	0	0	0	
'TAC'	206	0	0	0		'TAC'	206	0	0	0		'TAC'	206	0	0	0	
'AAT'	207	AGT	S	NoSin	0.26	'AAT'	207	AGT	S	NoSin	0.33	'AAT'	207	AGT	S	NoSin	0.31
'GGG'	208	0	0	0		'GGG'	208	0	0	0		'GGG'	208	0	0	0	
'AGG'	209	GGG	G	NoSin	0.35	'AGG'	209	GGG	G	NoSin	0.26	'AGG'	209	GGG	G	NoSin	0.3
'CTT'	210	CCT	P	NoSin	0.31	'CTT'	210	CCT	P	NoSin	0.29	'CTT'	210	CCT	P	NoSin	0.32
'GTA'	211	GTG	V	Sin	0.25	'GTA'	211	ATA	I	NoSin	1.09	'GTA'	211	0	0	0	
'GAT'	212	0	0	0		'GAT'	212	GGT	G	NoSin	0.27	'GAT'	212	0	0	0	
'AGT'	213	AGC	S	Sin	0.33	'AGT'	213	AGC	S	Sin	0.41	'AGT'	213	AGC	S	Sin	0.34
'GTT'	214	GCT/ GTC	A/V	NoSin /Sin	0.26/ 0.23	'GTT'	214	GCT	A	NoSin	0.21	'GTT'	214	GTC	V	Sin	0.28
'GTT'	215	GCT	A	NoSin	0.4	'GTT'	215	GCT	A	NoSin	0.37	'GTT'	215	GCT	A	NoSin	0.35
'TCA'	216	CCA/ TCG	P/S	NoSin /Sin	0.22/ 0.27	'TCA'	216	CCA/ TCG	P/S	NoSin /Sin	0.22/ 0.35	'TCA'	216	0	0	0	
'TGG'	217	0	0	0		'TGG'	217	0	0	0		'TGG'	217	0	0	0	
'TCC'	218	CCC	P	NoSin	0.29	'TCC'	218	CCC	P	NoSin	0.23	'TCC'	218	CCC	P	NoSin	0.37
'AAA'	219	AAG	K	Sin	0.28	'AAA'	219	AAG	K	Sin	0.25	'AAA'	219	AAG	K	Sin	0.27
'GAA'	220	0	0	0		'GAA'	220	0	0	0		'GAA'	220	0	0	0	

'ATT'	221	0	0	0		'ATT'	221	0	0	0		'ATT'	221	0	0	0	
'CTC'	222	0	0	0		'CTC'	222	0	0	0		'CTC'	222	0	0	0	
'AGG'	223	GGG	G	NoSin	0.23	'AGG'	223	GGG	G	NoSin	0.23	'AGG'	223	GGG	G	NoSin	0.26
'ACC'	224	GCC	A	NoSin	0.25	'ACC'	224	GCC/ TCC	A/S	NoSin	0.25/ 0.29	'ACC'	224	0	0	0	
'CAG'	225	0	0	0		'CAG'	225	0	0	0		'CAG'	225	0	0	0	
'GAG'	226	GGG	G	NoSin	0.21	'GAG'	226	GGG	G	NoSin	0.41	'GAG'	226	GGG	G	NoSin	0.28
'TCG'	227	0	0	0		'TCA'	227	TCG	S	Sin	0.33	'TCA'	227	CCA/ TCG	P/S	NoSin /Sin	0.31/ 0.28
'GAA'	228	0	0	0		'GAA'	228	GGA	G	NoSin	0.42	'GAA'	228	GGA/ GAG	G/E	NoSin /Sin	0.25/ 0.24
'TGC'	229	0	0	0		'TGC'	229	0	0	0		'TGC'	229	0	0	0	
'GTT'	230	GCT	A	NoSin	0.22	'GTT'	230	0	0	0		'GTT'	230	GCT/ GTC	A/V	NoSin /Sin	0.27/ 0.3
'TGT'	231	TGC	C	Sin	0.24	'TGT'	231	TGC	C	Sin	0.2	'TGT'	231	0	0	0	
'ATC'	232	0	0	0		'ATC'	232	0	0	0		'ATC'	232	0	0	0	
'AAT'	233	0	0	0		'AAT'	233	0	0	0		'AAT'	233	0	0	0	
'GGA'	234	GGG	G	Sin	0.23	'GGA'	234	GGG	G	Sin	0.24	'GGA'	234	GGG	G	Sin	0.26
'ACT'	235	GCT/ ACC	A/T	NoSin /Sin	0.27/ 0.25	'ACT'	235	GCT	A	NoSin	1.42	'ACT'	235	GCT	A	NoSin	0.41
'TGT'	236	0	0	0		'TGT'	236	0	0	0		'TGT'	236	0	0	0	
'ACA'	237	GCA	A	NoSin	0.2	'ACA'	237	GCA	A	NoSin	0.28	'ACA'	237	GCA	A	NoSin	0.27
'GTA'	238	0	0	0		'GTA'	238	0	0	0		'GTA'	238	0	0	0	
'GTA'	239	0	0	0		'GTA'	239	0	0	0		'GTA'	239	0	0	0	

'ATG'	240	0	0	0		'ATG'	240	0	0	0		'ATG'	240	0	0	0	
'ACT'	241	0	0	0		'ACT'	241	GCT	A	NoSin	0.24	'ACT'	241	0	0	0	
'GAT'	242	GGT	G	NoSin	0.23	'GAT'	242	GGT/ GAC	G/D	NoSin /Sin	0.31/ 0.23	'GAT'	242	0	0	0	
'GGA'	243	0	0	0		'GGA'	243	0	0	0		'GGA'	243	0	0	0	
'AGT'	244	GGT	G	NoSin	0.24	'AGT'	244	GGT/ AGC	G/S	NoSin /Sin	0.26/ 0.49	'AGT'	244	GGT	G	NoSin	0.26
'GCT'	245	GCC	A	Sin	0.44	'GCT'	245	ACT	T	NoSin	0.32	'GCT'	245	0	0	0	
'TCA'	246	TCG	S	Sin	0.23	'TCA'	246	TCG	S	NoSin	0.26	'TCA'	246	TCG	S	Sin	0.27
'GGA'	247	0	0	0		'GGA'	247	0	0	0		'GGA'	247	0	0	0	
'AAA'	248	AAG	K	Sin	0.53	'AAA'	248	AAG	K	Sin	0.54	'AAA'	248	AAG	K	Sin	0.55
'GCT'	249	0	0	0		'GCT'	249	0	0	0		'GCT'	249	0	0	0	
'GAT'	250	GGT	G	NoSin	0.21	'GAT'	250	GGT	G	NoSin	0.25	'GAT'	250	TAT/ GGT	Y/G	NoSin	2.28/ 0.44
'ACT'	251	0	0	0		'ACT'	251	ACC	T	Sin	1.51	'ACT'	251	ACC	T	Sin	0.28
'AAA'	252	0	0	0		'AAA'	252	0	0	0		'AAA'	252	0	0	0	
'ATA'	253	0	0	0		'ATA'	253	0	0	0		'ATA'	253	0	0	0	
'CTA'	254	0	0	0		'CTA'	254	0	0	0		'CTA'	254	0	0	0	
'TTC'	255	CTC	L	NoSin	33	'TTC'	255	CTC	L	NoSin	0.33	'TTC'	255	0	0	0	
'ATT'	256	0	0	0		'ATT'	256	GTT	V	NoSin	0.97	'ATT'	256	0	0	0	
'GAG'	257	GGG	G	NoSin	0.29	'GAG'	257	GGG/ GAA	G/E	NoSin /Sin	0.21/ 0.51	'GAG'	257	0	0	0	
'GAG'	258	0	0	0		'GAG'	258	0	0	0		'GAG'	258	0	0	0	
'GGG'	259	0	0	0		'GGG'	259	0	0	0		'GGG'	259	0	0	0	

'AAA'	260	0	0	0		'AAA'	260	0	0	0		'AAA'	260	0	0	0	
'ATC'	261	GTC	V	NoSin	0.2	'ATC'	261	0	0	0		'ATC'	261	GTC	V	NoSin	0.23
'GTT'	262	GTC	V	Sin	0.32	'GTT'	262	GTC	V	NoSin	0.23	'GTT'	262	0	0	0	
'CAT'	263	0	0	0		'CAT'	263	0	0	0		'CAT'	263	0	0	0	
'ACT'	264	0	0	0		'ACT'	264	ACC	T	Sin	0.77	'ACT'	264	0	0	0	
'AGC'	265	GGC	G	NoSin	0.21	'AGC'	265	0	0	0		'AGC'	265	0	0	0	
'ACA'	266	0	0	0		'ACA'	266	GCA	A	NoSin	0.25	'ACA'	266	0	0	0	
'TTG'	267	0	0	0		'TTG'	267	0	0	0		'TTG'	267	0	0	0	
'TCA'	268	CCA/ TCG	P/S	NoSin /Sin	0.25/ 0.29	'TCA'	268	CCA/ TCG	P/S	NoSin /Sin	0.23/ 0.29	'TCA'	268	CCA/ TCG	P/S	NoSin /Sin	0.29/ 0.25
'GGA'	269	0	0	0		'GGA'	269	0	0	0		'GGA'	269	0	0	0	
'AGT'	270	GGT/ AGC	G/S	NoSin /Sin	0.26/ 0.2	'AGT'	270	GGT	G	NoSin	0.24	'AGT'	270	GGT	G	NoSin	0.43
'GCT'	271	GCC	A	Sin	0.23	'GCT'	271	0	0	0		'GCT'	271	0	0	0	
'CAG'	272	CGG	R	NoSin	0.23	'CAG'	272	CGG/ CAA	R/Q	NoSin /Sin	0.56/ 0.41	'CAG'	272	0	0	0	
'CAT'	273	CGT	R	NoSin	0.2	'CAT'	273	CGT/ CAC	R/H	NoSin /Sin	0.71/ 3.41	'CAT'	273	CGT/ CAA	R/Q	NoSin	0.25/ 2.4
'GTC'	274	GCC	A	NoSin	0.24	'GTC'	274	GCC	A	NoSin	0.55	'GTC'	274	GCC	A	NoSin	0.27
'GAA'	275	GAG	E	Sin	0.36	'GAG'	275	GGG	G	NoSin	0.26	'GAA'	275	0	0	0	
'GAG'	276	GGG	G	NoSin	0.27	'GAG'	276	GGG	G	NoSin	1.43	'GAG'	276	GGG	G	NoSin	0.38
'TGC'	277	0	0	0		'TGC'	277	0	0	0		'TGC'	277	0	0	0	
'TCT'	278	CCT	P	NoSin	0.26	'TCT'	278	CCT	P	NoSin	0.29	'TCT'	278	CCT	P	NoSin	0.29
'TGC'	279	0	0	0		'TGC'	279	0	0	0		'TGC'	279	0	0	0	

'TAT'	280	CAT	H	NoSin	0.27	'TAT'	280	CAT/ TAC	H/Y	NoSin /Sin	0.27/ 0.26	'TAT'	280	CAT	H	NoSin	0.25
'CCT'	281	0	0	0		'CCT'	281	0	0	0		'CCT'	281	0	0	0	
'CGA'	282	0	0	0		'CGA'	282	0	0	0		'CGA'	282	0	0	0	
'TAT'	283	0	0	0		'TAT'	283	0	0	0		'TAT'	283	0	0	0	
'CCT'	284	0	0	0		'CCT'	284	CCC	P	Sin	0.25	'CCT'	284	CCC	P	Sin	0.23
'GGT'	285	GGC	G	NoSin	0.28	'GGT'	285	GGC	G	Sin	0.34	'GGT'	285	GGC	G	Sin	0.27
'GTC'	286	GCC	A	NoSin	0.21	'GTC'	286	GCC	A	NoSin	0.23	'GTC'	286	GCC	A	NoSin	0.23
'AGA'	287	GGA	G	NoSin	0.21	'AGA'	287	0	0	0		'AGA'	287	0	0	0	
'TGT'	288	0	0	0		'TGT'	288	0	0	0		'TGT'	288	0	0	0	
'GTC'	289	GCC	A	NoSin	0.25	'GTC'	289	GCC	A	NoSin	0.27	'GTC'	289	GCC	A	NoSin	0.25
'TGC'	290	CGC	R	NoSin	0.29	'TGC'	290	CGC	R	NoSin	0.3	'TGC'	290	CGC	R	NoSin	0.35
'AGA'	291	GGA	G	NoSin	0.21	'AGA'	291	GGA/ AGG	G/R	NoSin /Sin	0.27/ 0.24	'AGA'	291	AGG	R	Sin	0.27
'GAC'	292	GGC	G	NoSin	0.26	'GAC'	292	GGC	G	NoSin	0.24	'GAC'	292	GGC	G	NoSin	0.27
'AAC'	293	0	0	0		'AAC'	293	0	0	0		'AAC'	293	GAC/ AAA	D/K	NoSin	0.26/ 3.9
'TGG'	294	TAG	STOP	NoSin	0.03	'TGG'	294	0	0	0		'TGG'	294	0	0	0	
'AAA'	295	AAG	K	Sin	0.43	'AAA'	295	GAA/ AAG	E/K	NoSin /Sin	1.6/ 0.35	'AAA'	295	GAA/ AGA/ AAG	E/R/K	NoSin /NoSi n/Sin	0.24/ 0.34/ 0.61
'GGC'	296	0	0	0		'GGC'	296	0	0	0		'GGC'	296	0	0	0	
'TCC'	297	CCC	P	NoSin	0.3	'TCC'	297	CCC	P	NoSin	0.32	'TCC'	297	CCC	P	NoSin	0.26
'AAT'	298	0	0	0		'AAT'	298	0	0	0		'AAT'	298	0	0	0	

'CGG'	299	0	0	0		'CGG'	299	CAG	Q	NoSin	2.65	'CGG'	299	0	0	0	
'CCC'	300	0	0	0		'CCC'	300	0	0	0		'CCC'	300	0	0	0	
'ATC'	301	ACC	T	NoSin	0.27	'ATC'	301	ACC/ ATT	T/I	NoSin /Sin	1.2/ 1.22	'ATC'	301	ACC	T	NoSin	0.27
'GTA'	302	0	0	0		'GTA'	302	0	0	0		'GTA'	302	0	0	0	
'GAT'	303	0	0	0		'GAT'	303	0	0	0		'GAT'	303	0	0	0	
'ATA'	304	0	0	0		'ATA'	304	0	0	0		'ATA'	304	0	0	0	
'AAC'	305	0	0	0		'AAC'	305	0	0	0		'AAC'	305	0	0	0	
'ATA'	306	0	0	0		'ATA'	306	0	0	0		'ATA'	306	0	0	0	
'AAG'	307	0	0	0		'AAG'	307	0	0	0		'AAG'	307	0	0	0	
'GAT'	308	0	0	0		'GAT'	308	0	0	0		'GAT'	308	0	0	0	
'CAT'	309	CGT	R	NoSin	0.21	'CAT'	309	0	0	0		'CAT'	309	CGT	R	NoSin	0.31
'AGC'	310	0	0	0		'AGC'	310	GGC	G	NoSin	0.2	'AGC'	310	AAC	N	NoSin	3.45
'ATT'	311	0	0	0		'ATT'	311	0	0	0		'ATT'	311	0	0	0	
'GTG'	312	GTC/ GTT	V/V	Sin	0.19/ 99.79	'GTT'	312	GCT	A	NoSin	0.42	'GTT'	312	0	0	0	
'TCC'	313	0	0	0		'TCC'	313	CCC	P	NoSin	0.22	'TCC'	313	TCG	S	Sin	0.88
'AGT'	314	0	0	0		'AGT'	314	0	0	0		'AGT'	314	GGT/ AGC	G/S	NoSin /Sin	1.25/ 0.24
'TAT'	315	0	0	0		'TAT'	315	0	0	0		'TAT'	315	0	0	0	
'GTG'	316	0	0	0		'GTG'	316	GCG	A	NoSin	0.27	'GTG'	316	GCG	A	NoSin	0.26
'TGT'	317	0	0	0		'TGT'	317	CGT/ TGC	R/C	NoSin /Sin	0.25/ 0.26	'TGT'	317	CGT	R	NoSin	0.49
'TCA'	318	0	0	0		'TCA'	318	TCG	S	Sin	0.57	'TCA'	318	TCG	S	Sin	0.36

'GGA'	319	0	0	0		'GGA'	319	GGG	G	Sin	0.23	'GGA'	319	GGG	G	Sin	0.27
'CTT'	320	0	0	0		'CTT'	320	0	0	0		'CTT'	320	0	0	0	
'GTT'	321	GCT	A	NoSin	0.26	'GTT'	321	GCT/ GTC	A/V	NoSin /Sin	0.29/ 0.19	'GTT'	321	GCT	A	NoSin	0.3
'GGA'	322	GGG	G	Sin	0.35	'GGA'	322	GGG	G	Sin	0.56	'GGA'	322	GGG	G	Sin	0.42
'GAC'	323	GGC	G	NoSin	0.32	'GAC'	323	GGC	G	NoSin	0.36	'GAC'	323	GGC	G	NoSin	0.4
'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.26/ 0.34	'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.29/ 0.34	'ACA'	324	GCA/ ACG	A/T	NoSin /Sin	0.34/ 0.34
'CCC'	325	TCC	S	NoSin	0.24	'CCC'	325	0	0	0		'CCC'	325	0	0	0	
'AGA'	326	0	0	0		'AGA'	326	0	0	0		'AGA'	326	0	0	0	
'AAA'	327	0	0	0		'AAA'	327	0	0	0		'AAA'	327	0	0	0	
'AAC'	328	0	0	0		'AAC'	328	0	0	0		'AAC'	328	0	0	0	
'GAC'	329	GGC	G	NoSin	0.27	'GAC'	329	GGC	G	NoSin	0.27	'GAC'	329	GGC	G	NoSin	0.25
'AGC'	330	GGC	G	NoSin	0.25	'AGC'	330	GGC	G	NoSin	0.25	'AGC'	330	GGC	G	NoSin	0.3
'TCC'	331	CCC	P	NoSin	0.22	'TCC'	331	CCC	P	NoSin	0.48	'TCC'	331	CCC	P	NoSin	0.26
'AGC'	332	0	0	0		'AGC'	332	GGC	G	NoSin	0.2	'AGC'	332	GGC	G	NoSin	0.23
'AGT'	333	0	0	0		'AGT'	333	0	0	0		'AGT'	333	0	0	0	
'AGC'	334	0	0	0		'AGC'	334	GGC	G	NoSin	0.23	'AGC'	334	0	0	0	
'CAT'	335	0	0	0		'CAT'	335	CAA	Q	NoSin	1.19	'CAT'	335	0	0	0	
'TGT'	336	TGC	C	Sin	0.21	'TGT'	336	TGC	C	Sin	0.21	'TGT'	336	0	0	0	
'TTG'	337	0	0	0		'TTG'	337	ATG	M	NoSin	1.19	'TTG'	337	0	0	0	
'GAT'	338	0	0	0		'GAT'	338	0	0	0		'GAT'	338	0	0	0	
'CCT'	339	0	0	0		'CCT'	339	0	0	0		'CCT'	339	CCA	P	Sin	0.54

'AAC'	340	AGC	S	NoSin	0.2	'AAC'	340	0	0	0		'AAC'	340	AGC	S	NoSin	0.25
'AAT'	341	0	0	0		'AAT'	341	0	0	0		'AAT'	341	0	0	0	
'GAA'	342	GAG	E	Sin	0.32	'GAA'	342	GAG	E	Sin	0.31	'GAA'	342	GAG	E	Sin	0.48
'GAA'	343	GGA/ GAG	G/E	NoSin /Sin	0.24/ 0.25	'GAA'	343	GAG	E	Sin	0.31	'GAA'	343	0	0	0	
'GGT'	344	0	0	0		'GGT'	344	0	0	0		'GGT'	344	0	0	0	
'GGT'	345	GGC	G	Sin	0.27	'GGT'	345	GGC	G	Sin	0.24	'GGT'	345	GGC	G	Sin	0.34
'CAT'	346	0	0	0		'CAT'	346	0	0	0		'CAT'	346	0	0	0	
'GGA'	347	GGG	G	Sin	0.26	'GGA'	347	GGG	G	Sin	0.36	'GGA'	347	GGG	G	Sin	0.42
'GTG'	348	0	0	0		'GTG'	348	0	0	0		'GTG'	348	0	0	0	
'AAA'	349	AAG	K	Sin	0.39	'AAA'	349	AAG	K	Sin	0.45	'AAA'	349	GAA/ AAG	E/K	NoSin /Sin	0.24/ 0.44
'GGC'	350	0	0	0		'GGC'	350	TGC	C	NoSin	0.22	'GGC'	350	0	0	0	
'TGG'	351	CGG	R	NoSin	0.27	'TGG'	351	CGG	R	NoSin	0.24	'TGG'	351	CGG	R	NoSin	0.28
'GCC'	352	0	0	0		'GCC'	352	0	0	0		'GCC'	352	0	0	0	
'TTT'	353	CTT	L	NoSin	0.32	'TTT'	353	CTT	L	NoSin	0.29	'TTT'	353	CTT	L	NoSin	0.42
'GAT'	354	0	0	0		'GAT'	354	0	0	0		'GAT'	354	GGT	G	NoSin	0.3
'GAT'	355	GGT	G	NoSin	0.23	'GAT'	355	GGT	G	NoSin	0.44	'GAT'	355	GGT	G	NoSin	0.29
'GGA'	356	0	0	0		'GGA'	356	0	0	0		'GGA'	356	0	0	0	
'AAT'	357	0	0	0		'AAT'	357	0	0	0		'AAT'	357	0	0	0	
'GAC'	358	GGC	G	NoSin	0.34	'GAC'	358	GGC	G	NoSin	0.28	'GAC'	358	GGC	G	NoSin	0.38
'GTG'	359	GCG	A	NoSin	0.21	'GTG'	359	GCG	A	NoSin	0.21	'GTA'	359	0	0	0	
'TGG'	360	0	0	0		'TGG'	360	CGG	R	NoSin	0.35	'TGG'	360	TGA	STOP	NoSin	3.05

'ATG'	361	GTG	V	NoSin	0.28	'ATG'	361	GTG/ ACG	V/T	NoSin	0.22/ 0.24	'ATG'	361	0	0	0	
'GGA'	362	0	0	0		'GGA'	362	0	0	0		'GGA'	362	GAA	E	NoSin	2.57
'AGA'	363	0	0	0		'AGA'	363	0	0	0		'AGA'	363	0	0	0	
'ACA'	364	GCA/ ACG	A/T	NoSin /Sin	0.23/ 0.27	'ACA'	364	GCA	A	NoSin	0.25	'ACA'	364	GCA/ ACG	A/T	NoSin /Sin	0.26/ 0.34
'ATC'	365	0	0	0		'ATC'	365	0	0	0		'ATC'	365	GTC	V	NoSin	0.35
'AAC'	366	0	0	0		'AAC'	366	0	0	0		'AAC'	366	0	0	0	
'GAG'	367	GGG	G	NoSin	0.23	'GAG'	367	GGG	G	NoSin	1.07	'GAG'	367	0	0	0	
'ACG'	368	GCG	A	NoSin	0.33	'ACG'	368	GCG	A	NoSin	0.36	'ACG'	368	GCG	A	NoSin	0.37
'TCA'	369	CCA	P	NoSin	0.32	'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.34/ 0.33	'TCA'	369	CCA/ TCG	P/S	NoSin /Sin	0.27/ 0.29
'CGC'	370	0	0	0		'CGC'	370	0	0	0		'CGC'	370	0	0	0	
'TTA'	371	0	0	0		'TTA'	371	0	0	0		'TTA'	371	0	0	0	
'GGG'	372	0	0	0		'GGG'	372	0	0	0		'GGG'	372	0	0	0	
'TAT'	373	0	0	0		'TAT'	373	0	0	0		'TAT'	373	0	0	0	
'GAA'	374	GGA	G	NoSin	0.22	'GAA'	374	0	0	0		'GAA'	374	0	0	0	
'ACC'	375	GCC	A	NoSin	0.34	'ACC'	375	GCC	A	NoSin	0.36	'ACC'	375	GCC	A	NoSin	0.34
'TTC'	376	0	0	0		'TTC'	376	0	0	0		'TTC'	376	CTC	L	NoSin	0.25
'AAA'	377	0	0	0		'AAA'	377	0	0	0		'AAA'	377	0	0	0	
'GTC'	378	GCC	A	NoSin	0.21	'GTC'	378	0	0	0		'GTC'	378	GCC	A	NoSin	0.65
'ATT'	379	0	0	0		'ATT'	379	0	0	0		'ATT'	379	0	0	0	
'GAA'	380	0	0	0		'GAA'	380	GAG	E	Sin	0.21	'GAA'	380	0	0	0	
'GGC'	381	0	0	0		'GGC'	381	0	0	0		'GGC'	381	0	0	0	

'TGG'	382	CGG	R	NoSin	0.26	'TGG'	382	0	0	0		'TGG'	382	CGG	R	NoSin	0.37
'TCC'	383	CCC	P	NoSin	0.28	'TCC'	383	CCC	P	NoSin	0.28	'TCC'	383	0	0	0	
'AAC'	384	AGC	S	NoSin	0.22	'AAC'	384	0	0	0		'AAC'	384	0	0	0	
'CCT'	385	0	0	0		'CCT'	385	0	0	0		'CCT'	385	0	0	0	
'AAG'	386	0	0	0		'AAG'	386	0	0	0		'AAG'	386	0	0	0	
'TCC'	387	CCC	P	NoSin	0.23	'TCC'	387	CCC	P	NoSin	0.24	'TCC'	387	0	0	0	
'AAA'	388	0	0	0		'AAA'	388	0	0	0		'AAA'	388	0	0	0	
'TTG'	389	0	0	0		'TTG'	389	TTA	L	Sin	1.47	'TTG'	389	0	0	0	
'CAG'	390	CGG	R	NoSin	0.36	'CAG'	390	CGG	R	NoSin	0.65	'CAG'	390	CGG	R	NoSin	0.42
'ATA'	391	0	0	0		'ATA'	391	0	0	0		'ATA'	391	0	0	0	
'AAT'	392	0	0	0		'AAT'	392	0	0	0		'AAT'	392	0	0	0	
'AGG'	393	0	0	0		'AGG'	393	GGG	G	NoSin	0.26	'AGG'	393	GGG	G	NoSin	0.75
'CAA'	394	CAG	Q	Sin	0.33	'CAA'	394	CAG	Q	Sin	0.27	'CAA'	394	CAG	Q	Sin	0.33
'GTC'	395	GCC	A	NoSin	0.26	'GTC'	395	GCC	A	NoSin	0.74	'GTC'	395	GCC	A	NoSin	0.31
'ATA'	396	GTA	V	NoSin	0.26	'ATA'	396	ATC	I	Sin	0.27	'ATA'	396	0	0	0	
'GTT'	397	0	0	0		'GTT'	397	0	0	0		'GTT'	397	0	0	0	
'GAC'	398	GGC	G	NoSin	0.36	'GAC'	398	GGC	G	NoSin	0.45	'GAC'	398	GGC	G	NoSin	0.46
'AGA'	399	AGG	R	Sin	0.35	'AGA'	399	AGG	R	Sin	0.42	'AGA'	399	AGG	R	Sin	0.5
'GGT'	400	0	0	0		'GGT'	400	0	0	0		'GGT'	400	0	0	0	
'GAT'	401	GGT	G	NoSin	0.31	'GAT'	401	GGT	G	NoSin	0.3	'GAT'	401	GGT	G	NoSin	0.44
'AGG'	402	GGG	G	NoSin	0.3	'AGG'	402	GGG	G	NoSin	0.23	'AGG'	402	GGG	G	NoSin	0.44
'TCC'	403	0	0	0		'TCC'	403	0	0	0		'TCC'	403	0	0	0	
'GGT'	404	0	0	0		'GGT'	404	0	0	0		'GGT'	404	TGT/	C/G	NoSin	4.03/

														GGC		/Sin	0.52
'TAT'	405	0	0	0		'TAT'	405	0	0	0		'TAT'	405	0	0	0	
'TCT'	406	0	0	0		'TCT'	406	TCC	S	Sin	0.33	'TCT'	406	0	0	0	
'GGT'	407	0	0	0		'GGT'	407	0	0	0		'GGT'	407	0	0	0	
'ATT'	408	0	0	0		'ATT'	408	0	0	0		'ATT'	408	0	0	0	
'TTC'	409	0	0	0		'TTC'	409	0	0	0		'TTC'	409	0	0	0	
'TCT'	410	0	0	0		'TCT'	410	CCT	P	NoSin	0.25	'TCT'	410	0	0	0	
'GTT'	411	0	0	0		'GTT'	411	0	0	0		'GTT'	411	0	0	0	
'GAA'	412	GAG	E	Sin	0.44	'GAA'	412	GAG	E	Sin	0.37	'GAA'	412	GAG	E	Sin	0.62
'GGC'	413	0	0	0		'GGC'	413	0	0	0		'GGC'	413	0	0	0	
'AAA'	414	AGA	R	NoSin	0.36	'AAA'	414	AGA	R	NoSin	0.36	'AAA'	414	AGA	R	NoSin	0.38
'AGC'	415	GGC	G	NoSin	0.49	'AGC'	415	GGC	G	NoSin	0.38	'AGC'	415	GGC	G	NoSin	0.45
'TGC'	416	CGC	R	NoSin	0.28	'TGC'	416	CGC	R	NoSin	0.26	'TGC'	416	0	0	0	
'ATC'	417	GTC/ ACC	V/T	NoSin	0.31/ 0.3	'ATC'	417	GTC	V	NoSin	0.36	'ATC'	417	GTC	V	NoSin	0.41
'AAT'	418	0	0	0		'AAT'	418	0	0	0		'AAT'	418	0	0	0	
'CGG'	419	0	0	0		'CGG'	419	0	0	0		'CGG'	419	0	0	0	
'TGC'	420	0	0	0		'TGC'	420	0	0	0		'TGC'	420	0	0	0	
'TTT'	421	0	0	0		'TTT'	421	0	0	0		'TTT'	421	0	0	0	
'TAT'	422	0	0	0		'TAT'	422	0	0	0		'TAT'	422	0	0	0	
'GTG'	423	0	0	0		'GTG'	423	0	0	0		'GTG'	423	0	0	0	
'GAG'	424	0	0	0		'GAG'	424	0	0	0		'GAG'	424	0	0	0	
'TTG'	425	0	0	0		'TTG'	425	0	0	0		'TTG'	425	0	0	0	

'ATT'	426	GTT	V	NoSin	0.56	'ATT'	426	GTT	V	NoSin	0.48	'ATT'	426	GTT	V	NoSin	0.45
'AGG'	427	0	0	0		'AGG'	427	0	0	0		'AGG'	427	0	0	0	
'GGA'	428	0	0	0		'GGA'	428	GGG	G	Sin	0.39	'GGA'	428	0	0	0	
'AGA'	429	0	0	0		'AGG'	429	0	0	0		'AGA'	429	0	0	0	
'AAA'	430	AAG	K	Sin	0.73	'AAA'	430	AAG	K	Sin	0.86	'AAA'	430	AAG	K	Sin	0.85
'GAG'	431	0	0	0		'GAG'	431	0	0	0		'GAG'	431	0	0	0	
'GAA'	432	GAG	E	Sin	0.81	'GAA'	432	GAG	E	Sin	0.81	'GAA'	432	GAG	E	Sin	1.23
'ACT'	433	GCT	A	NoSin	0.72	'ACT'	433	GCT	A	NoSin	0.56	'ACT'	433	0	0	0	
'GAA'	434	GGA/ GAG	G/E	NoSin /Sin	0.53/ 0.83	'GAA'	434	GGA/ GAG	G/E	NoSin /Sin	0.57/ 0.5	'GAA'	434	GAG	E	Sin	0.77
'GTC'	435	0	0	0		'GTC'	435	0	0	0		'GTC'	435	0	0	0	
'TTG'	436	0	0	0		'TTG'	436	0	0	0		'TTG'	436	0	0	0	
'TGG'	437	0	0	0		'TGG'	437	0	0	0		'TGG'	437	0	0	0	
'ACC'	438	0	0	0		'ACC'	438	0	0	0		'ACC'	438	0	0	0	
'TCA'	439	0	0	0		'TCA'	439	0	0	0		'TCA'	439	0	0	0	
'AAC'	440	AGC	S	NoSin	0.92	'AAC'	440	AGC	S	NoSin	0.91	'AAC'	440	0	0	0	
'AGT'	441	0	0	0		'AGT'	441	0	0	0		'AGT'	441	0	0	0	
'ATT'	442	0	0	0		'ATT'	442	0	0	0		'ATT'	442	0	0	0	

Tabal 7. Resistencia a inhibidores de la Neuraminidasa del VIA H3N2. Adquisición de mutaciones hacia los Inhibidores de la Neuraminidasa INAs. Mutaciones de resistencia a INAs reportadas en la bibliografía: Q136K, N142S, R292K, G320E, I222T+S331R. Otros cambios relevantes: I222V/T/R/K, D198E, S246G/N, N294S, Del 244-247, Q136K/R, E119K, T156I+D213G, I222T+S331R, I222R/V+H274Y, S246N+H274Y.

Cambio	Muestras	Frecuencia (%)
E119G	116, 132, 159, 220, 224, 250, 264	0.26, 0.47, 0.45, 0.23, 0.25, 0.27, 0.28
H274R	116, 132, 159, 220, 224, 250, 264	0.68, 0.21, 0.61, 0.20, 0.23, 0.71, 0.25
H274Q	264	2.4
R292K	220	0.26
R292G	132, 159, 206, 220, 224, 250	0.28, 0.24, 0.23, 0.24, 0.21, 0.27
N294S	202	0.25
N294D	132,220, 264	0.25, 0.22, 0.27
N294K	264	0.26
G320R	116	1.17
N142S	132, 264	0.23, 0.6
N142D	264	0.37
S331G	116, 132, 159, 202, 206, 220, 224, 250, 264	0.38, 0.24, 0.28, 0.27, 0.25, 0.21, 0.27, 0.25, 0.25
Q136R	116, 132, 159, 206, 220, 224, 250, 264	0.34, 0.37, 0.26, 0.34, 0.31, 0.29, 0.28, 0.58
S246A	116, 132, 159, 202, 206, 220, 224, 264	100
S246T	250	100
D213G	116, 132, 206, 220, 250	0.32, 0.25, 0.27, 0.23, 0.27
T156A	132, 159, 224	0.25, 0.24, 0.22
D198G	116, 132, 159, 202, 206, 220, 224, 250, 264	0.29, 0.3, 0.32, 0.3, 0.29, 0.24, 0.25, 0.38, 0.3
R151G	116, 132, 159, 224, 250, 264	0.28, 0.3, 0.24, 0.24, 0.21, 0.35

Tabla 8. Resultados ensamblado de haplotipos por tres abordajes diferentes. QuRe, PredictHaplo y QuasiRecomb. #var representa el número de variantes ensambladas, se indica para cada muestra y gen analizado.

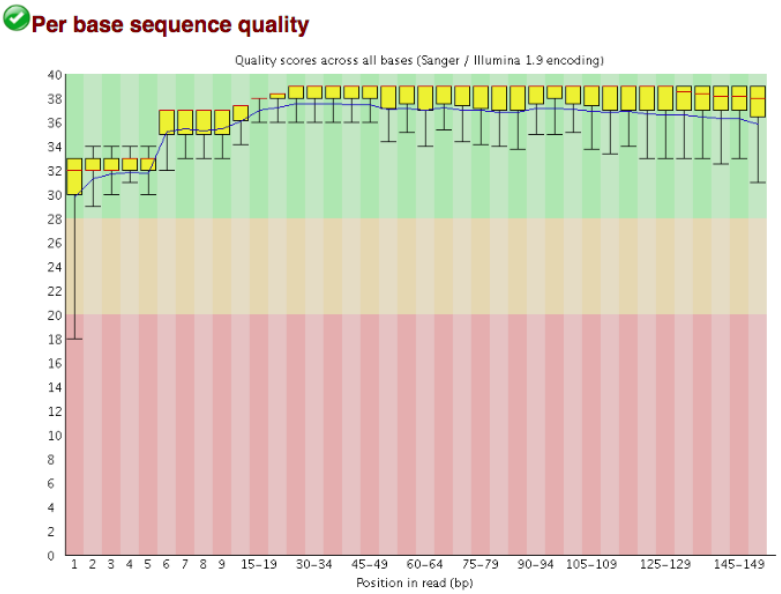
Gen	Método	#var_m116	#var_m132	#var_m159	#var_m202	#var_m206	#var_m220	#var_m224	#var_m250	#var_m264
HA	QuRe	126	66	66	359	107	127	136	266	268
	PredictHaplo	2	3	7	16	4	1	1	5	11
	QuasiRecomb	69	40	49	53	31	65	58	43	83
NA	QuRe	155	126	87	138	273	223	115	251	78
	PredictHaplo	4	7	14	10	5	13	4	29	13
	QuasiRecomb	39	80	43	102	91	129	91	131	102

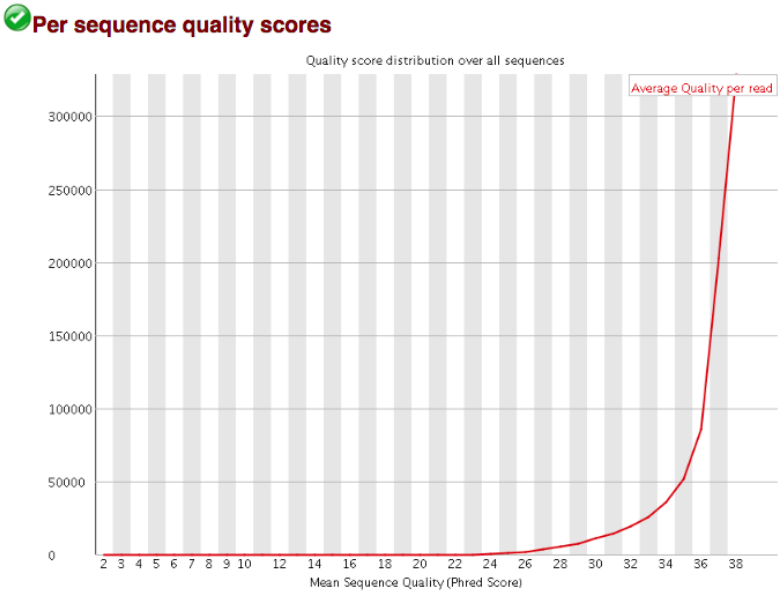
Anexo II

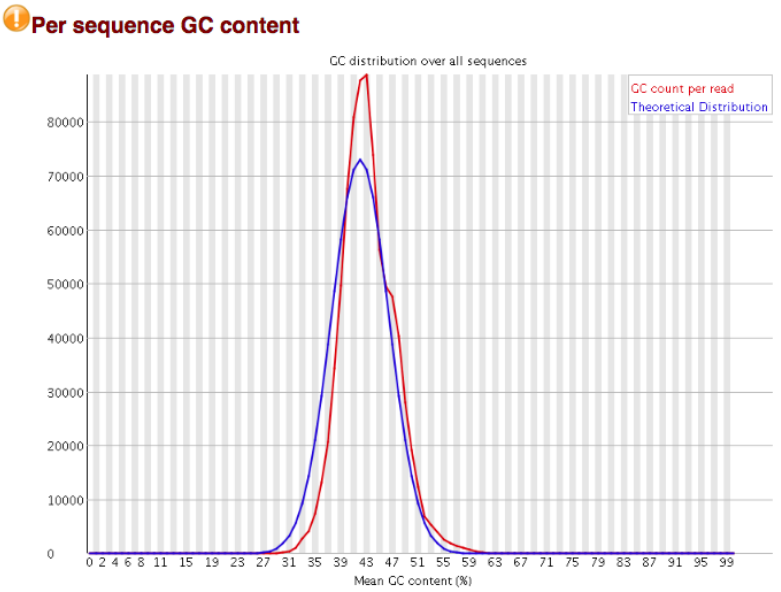
Figura 1. Evaluación de los datos masivos con el programa FastQC para las nueve muestras analizadas.
Se aprecian los outputs del programa FastQC para los datos crudos emitidos por el secuenciador MiSeq de Illumina, así como para la salida de datos de todas las etapas de pre-procesamiento (trimming de adaptadores, primers y calidad)

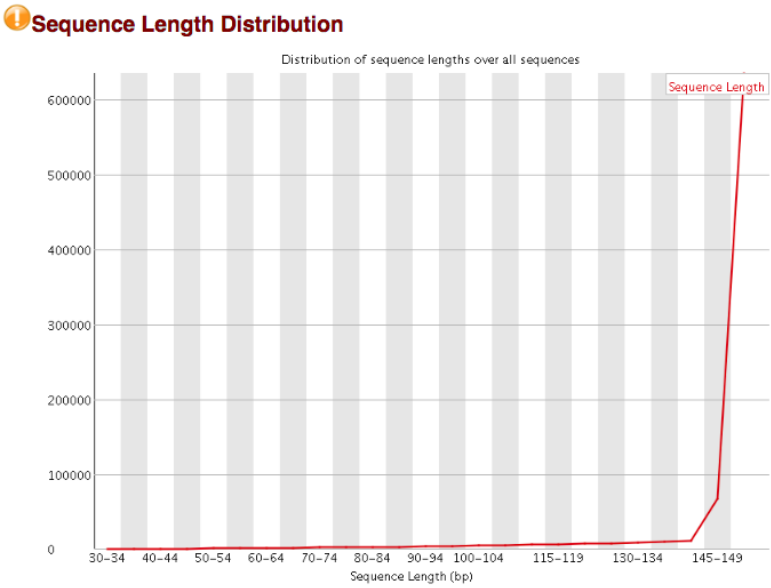
m116:

Datos Crudos R1



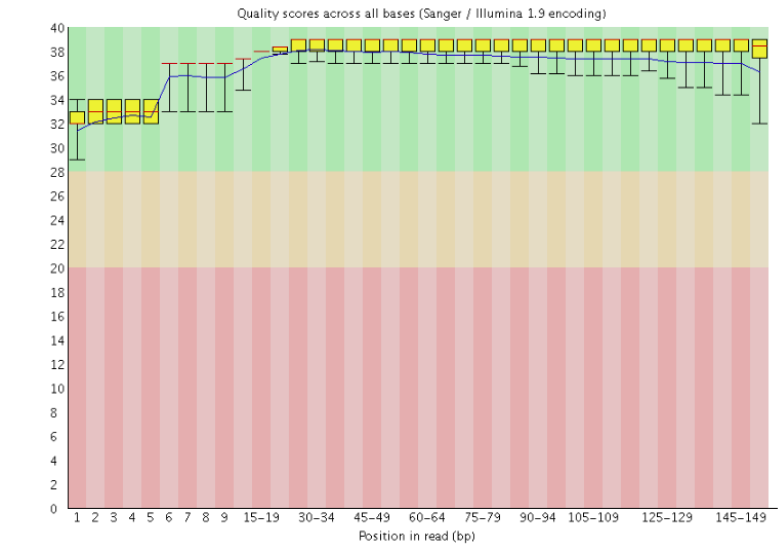




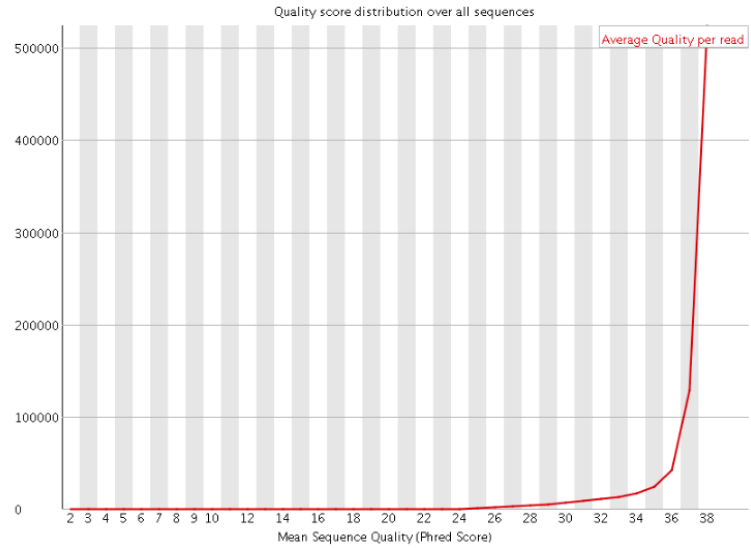


Datos crudos R2

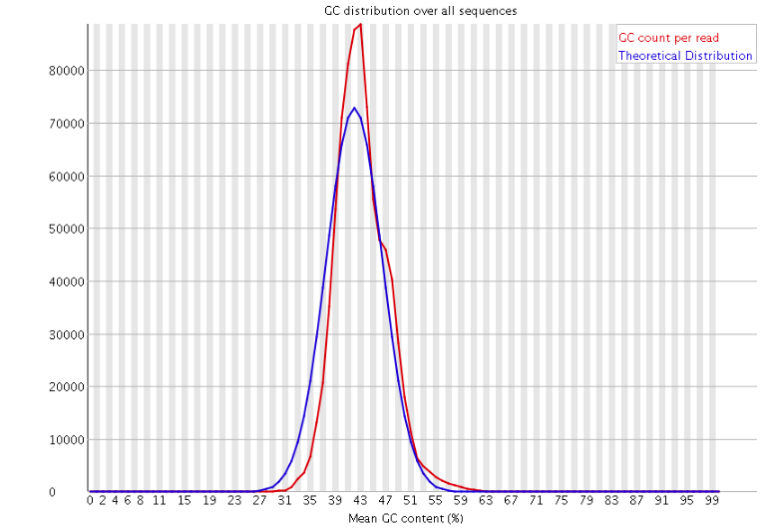
Per base sequence quality



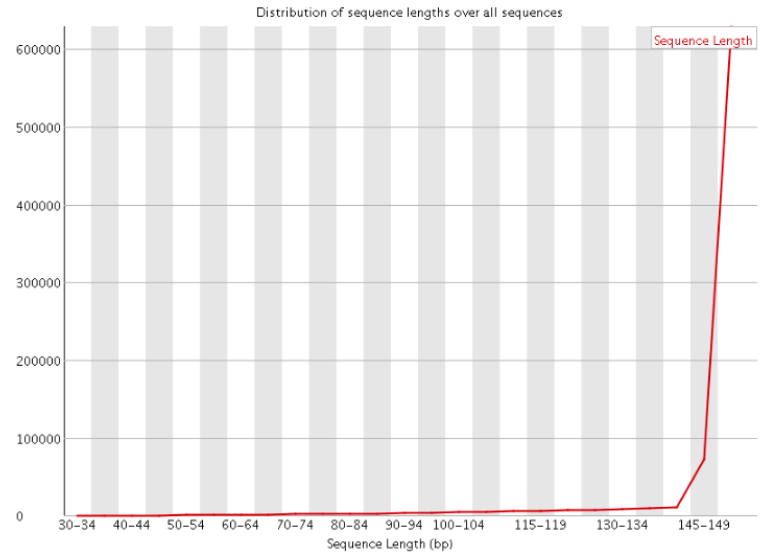
Per sequence quality scores



Per sequence GC content

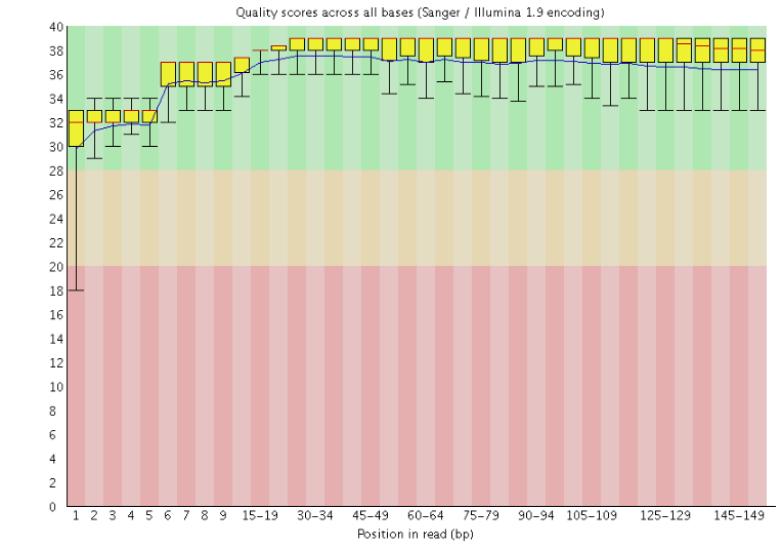


Sequence Length Distribution

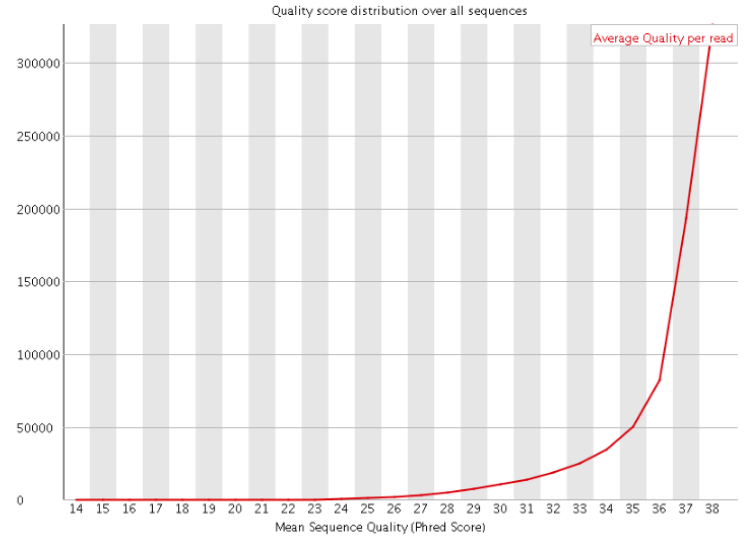


Datos trimados para primers y adaptadores R1

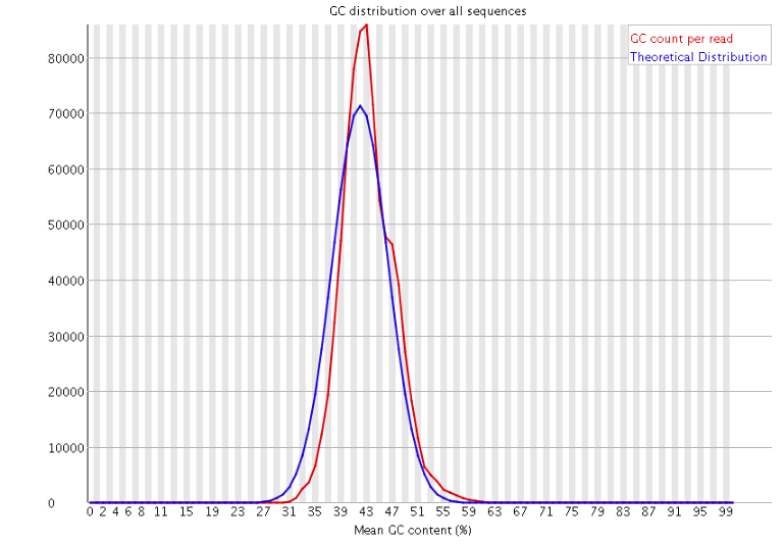
Per base sequence quality



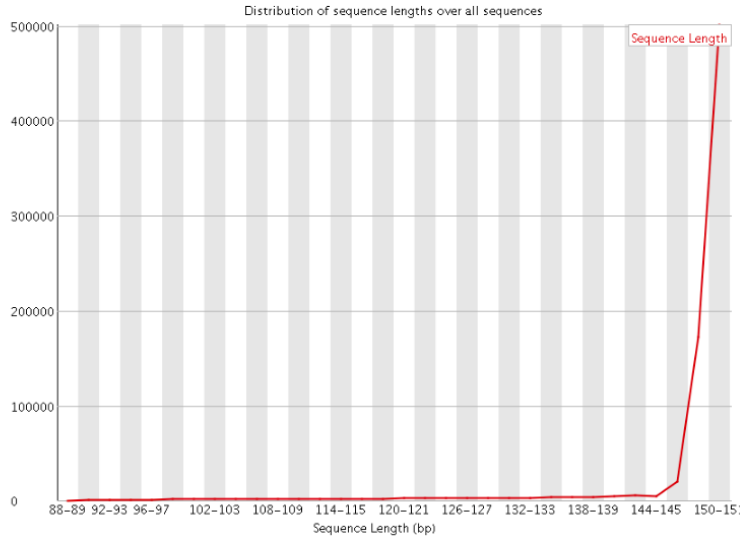
Per sequence quality scores



Per sequence GC content

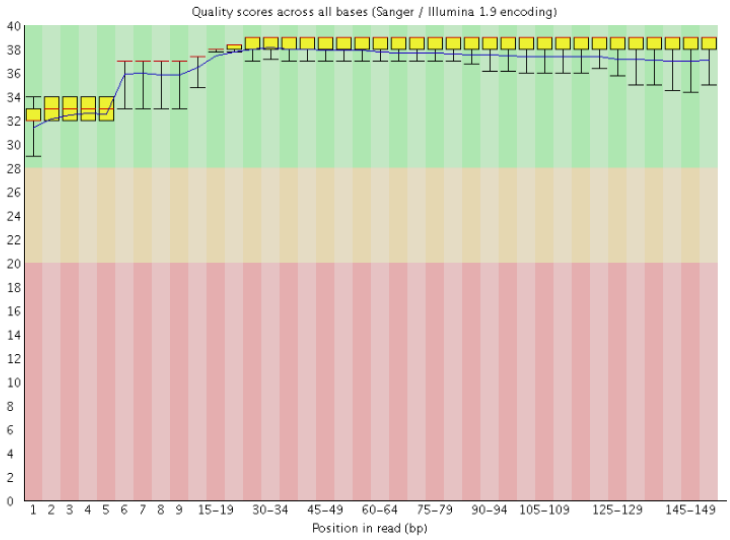


Sequence Length Distribution

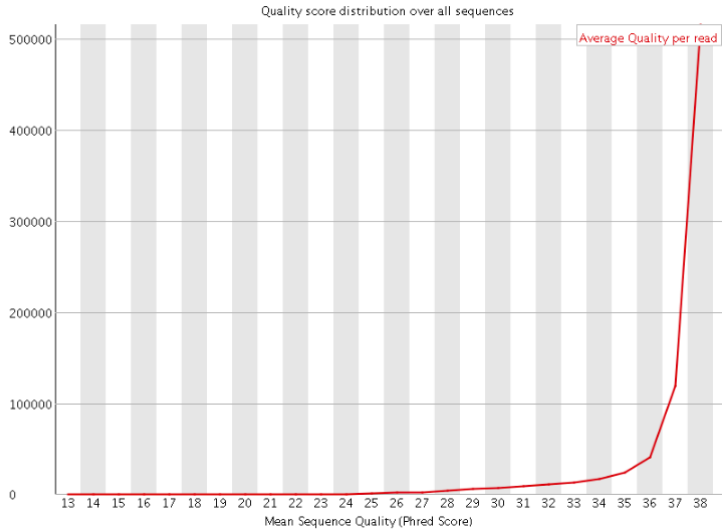


Datos trimados para primers y adaptadores R2

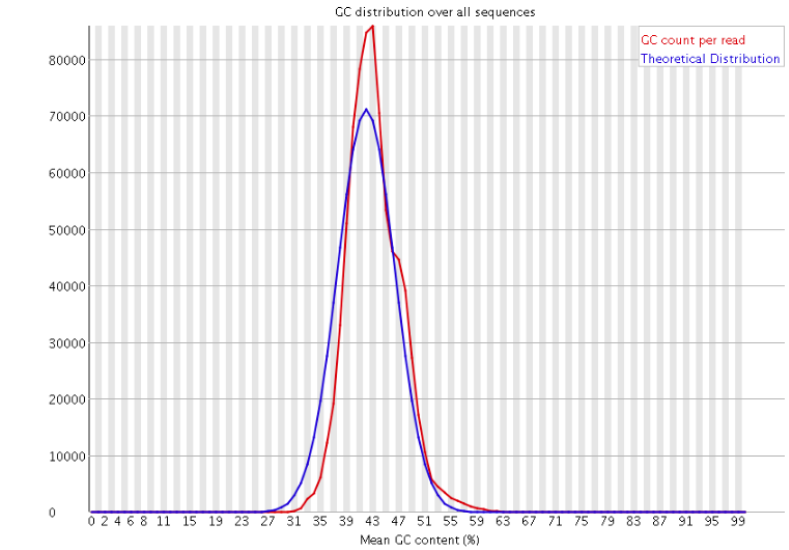
Per base sequence quality



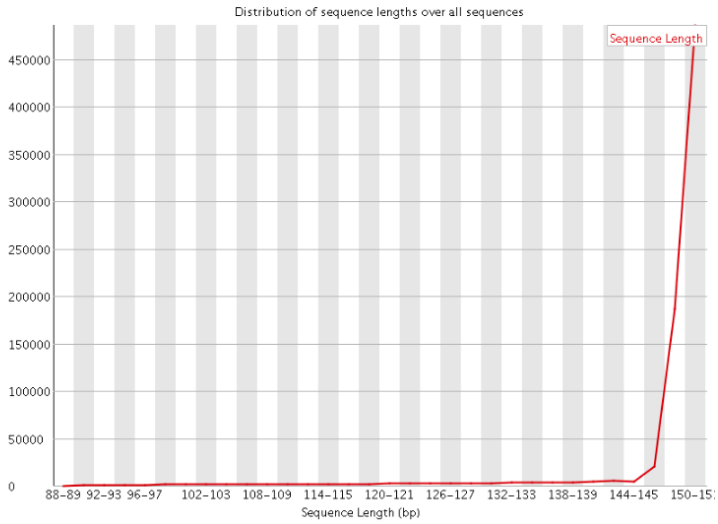
Per sequence quality scores



Per sequence GC content

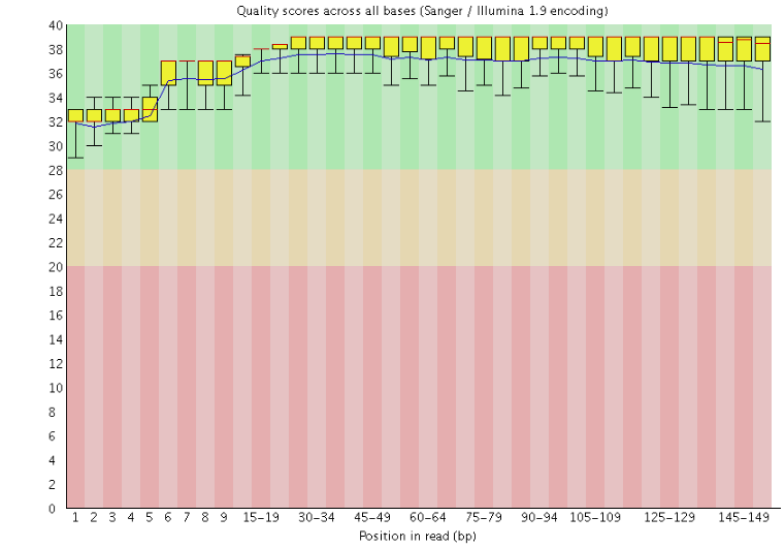


Sequence Length Distribution

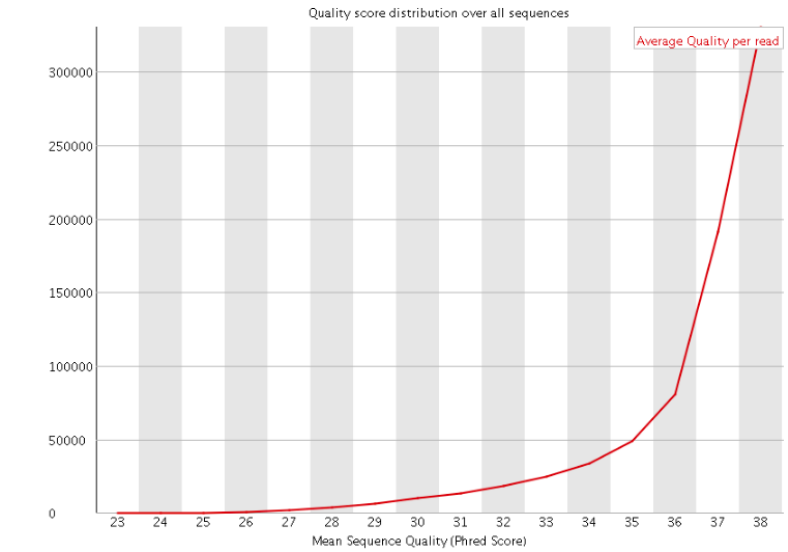


Datos trimados por calidad R1

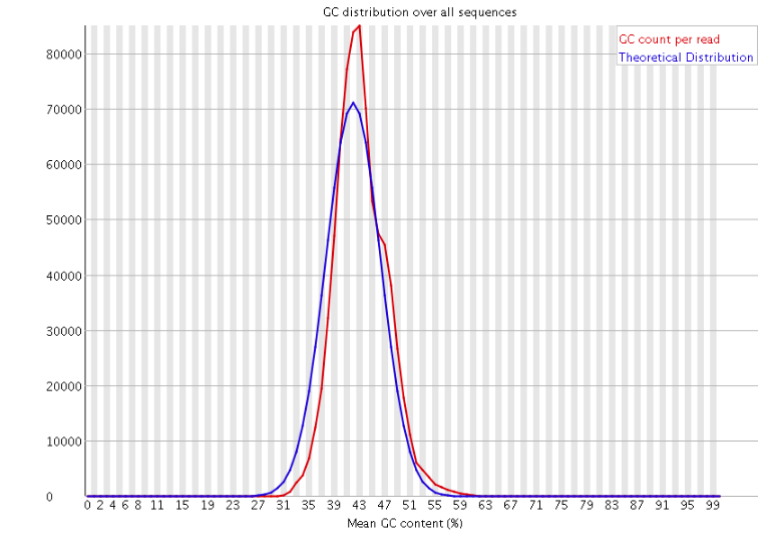
Per base sequence quality



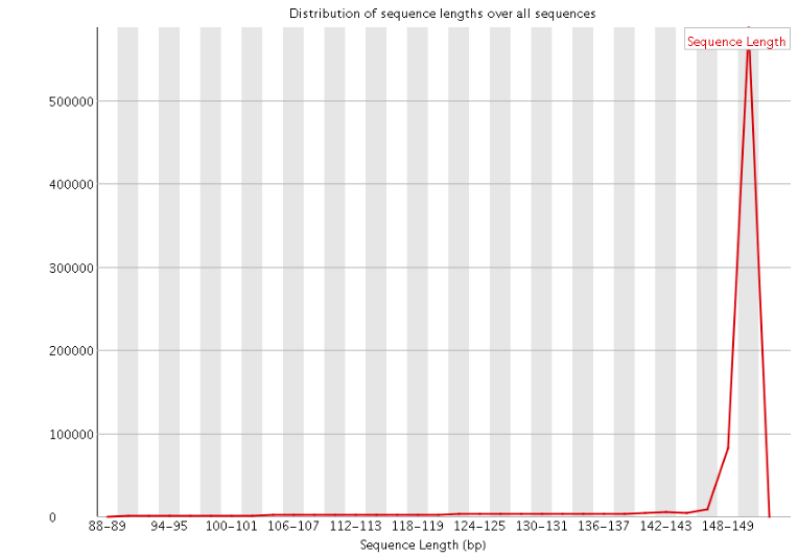
Per sequence quality scores



Per sequence GC content

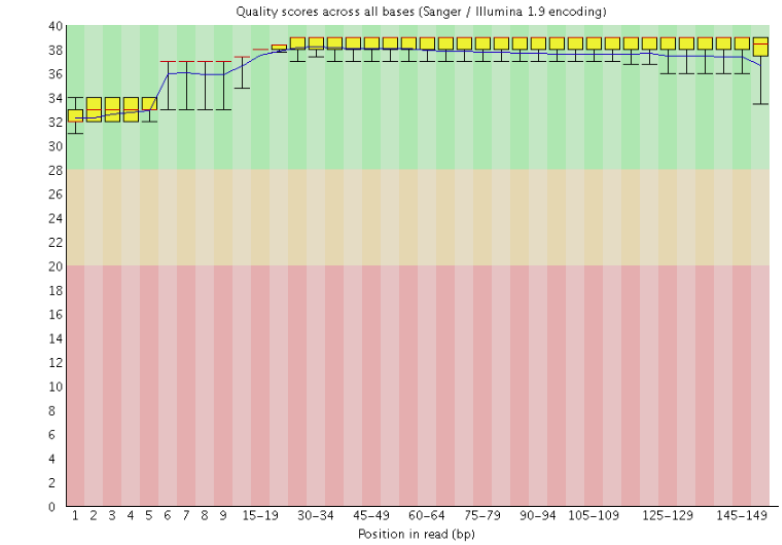


Sequence Length Distribution

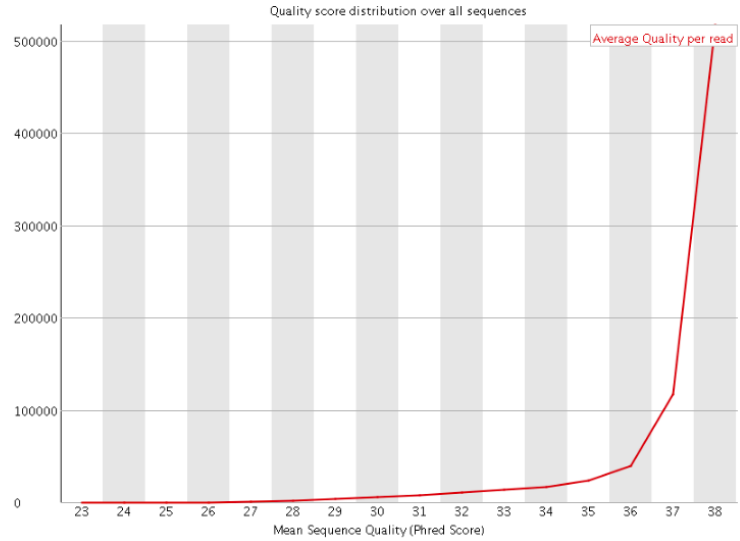


Datos trimados por calidad R2

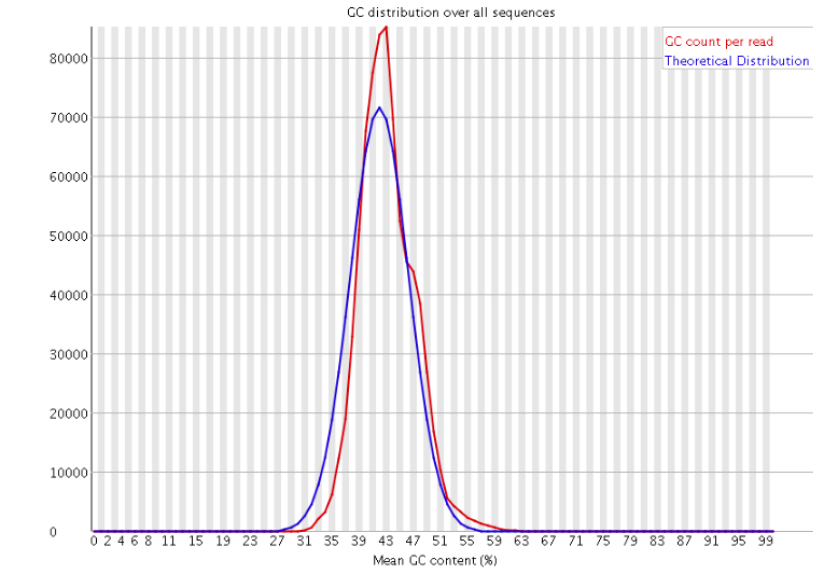
Per base sequence quality



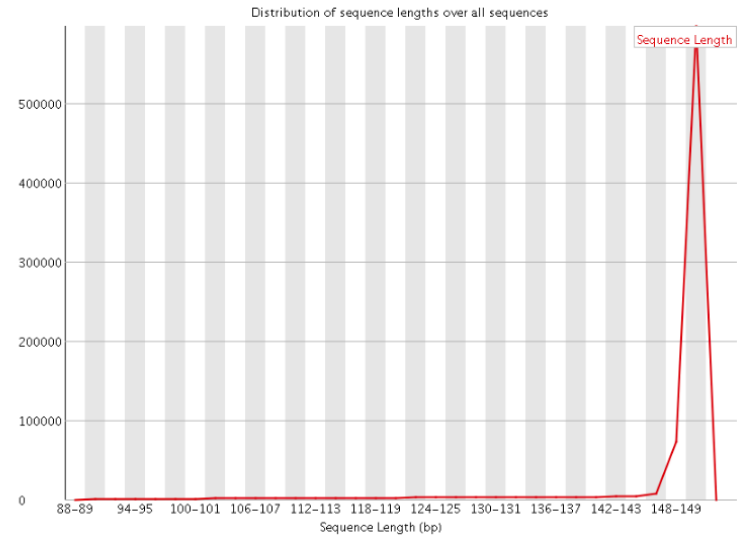
Per sequence quality scores



Per sequence GC content



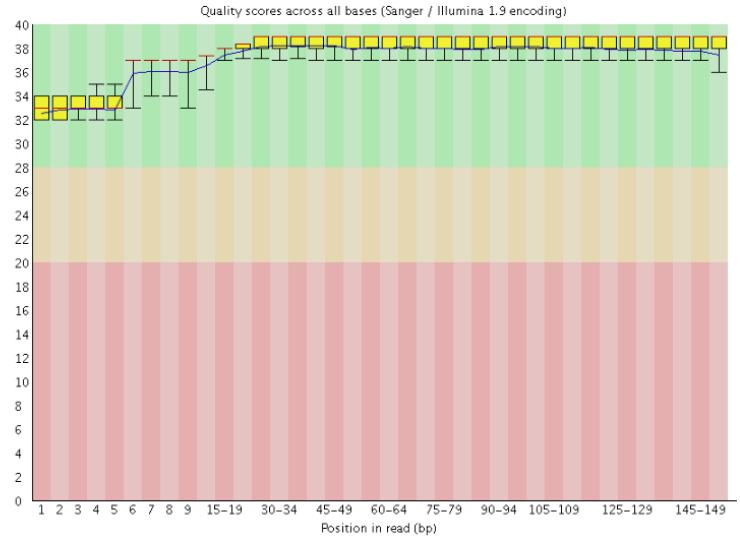
Sequence Length Distribution



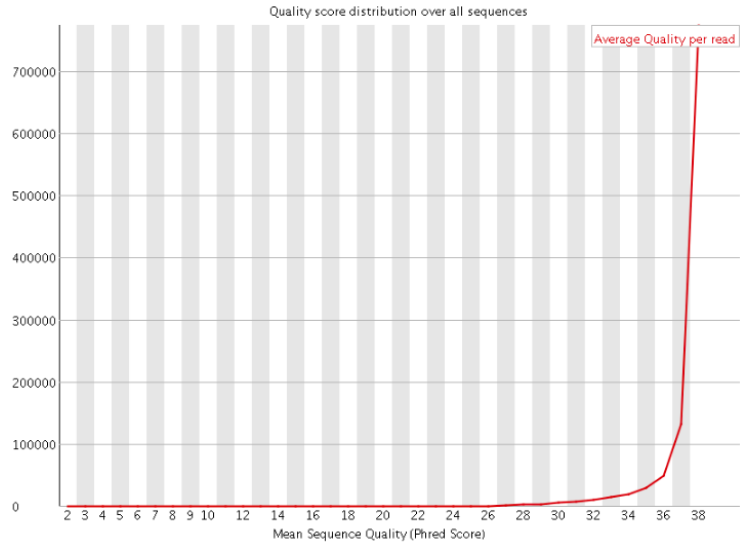
m132:

Datos Crudos R1

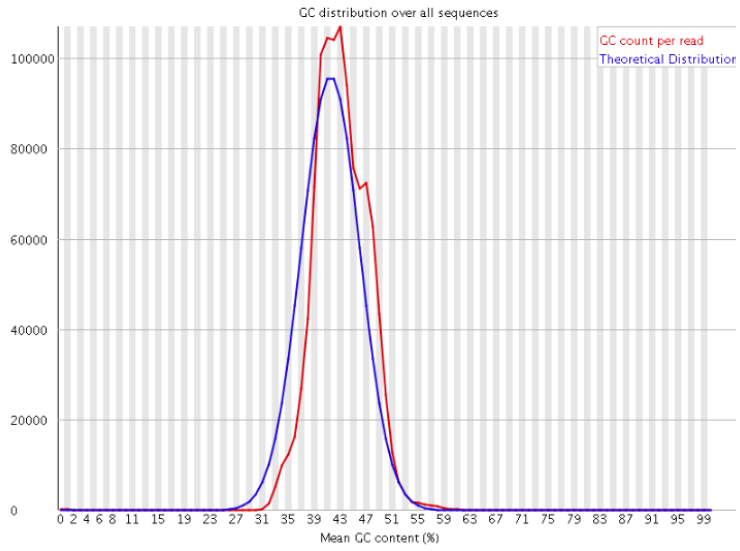
Per base sequence quality



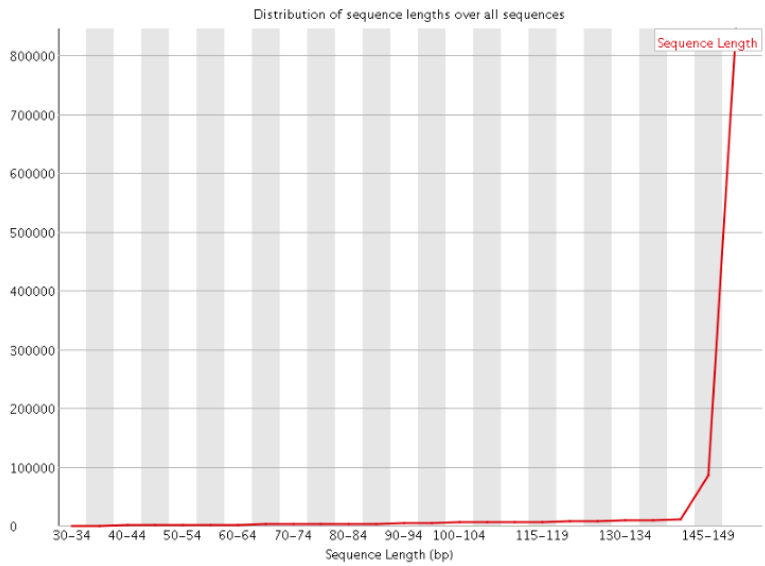
Per sequence quality scores



Per sequence GC content

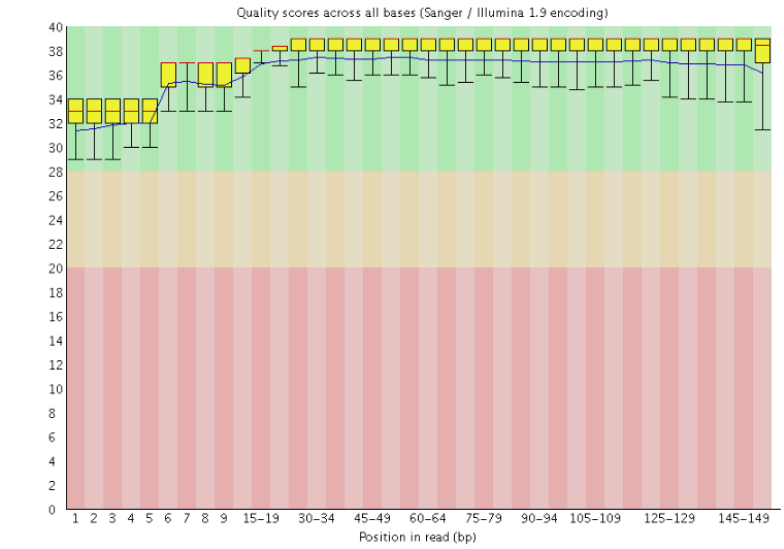


Sequence Length Distribution

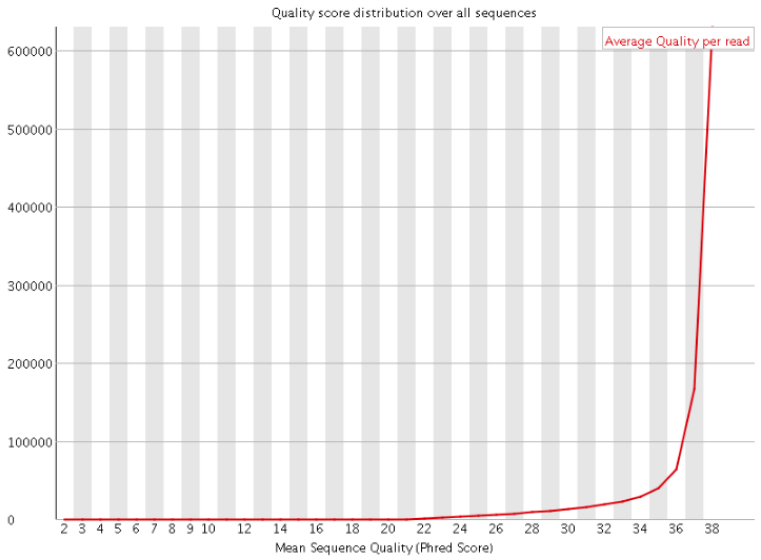


Datos Crudos R2

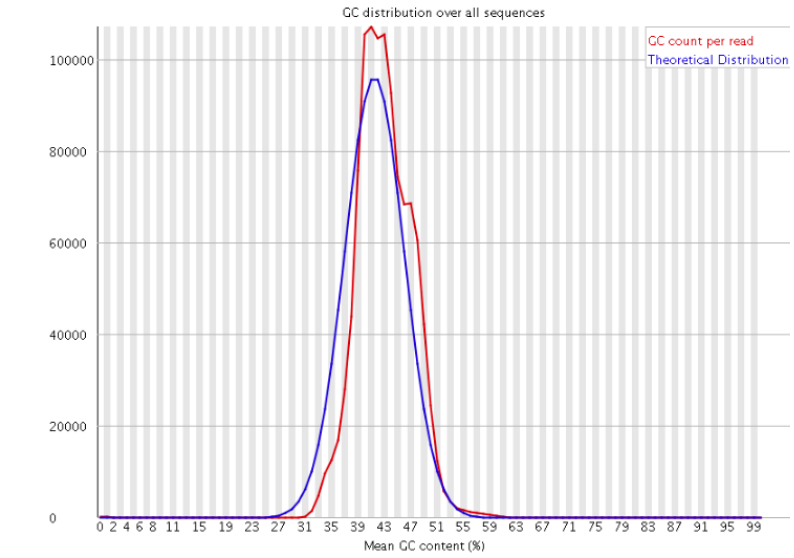
Per base sequence quality



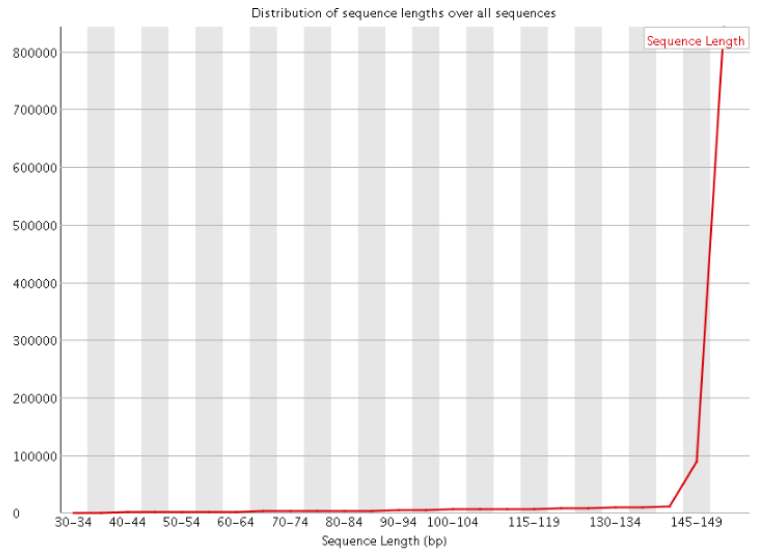
Per sequence quality scores



Per sequence GC content

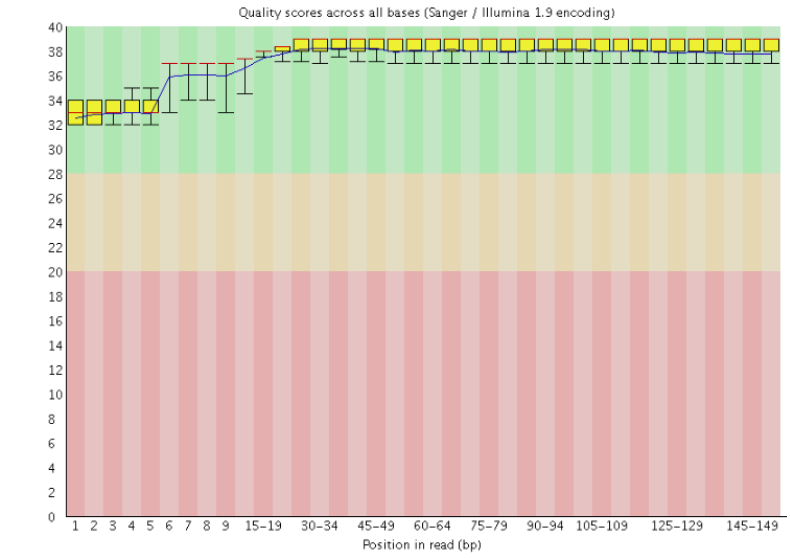


Sequence Length Distribution

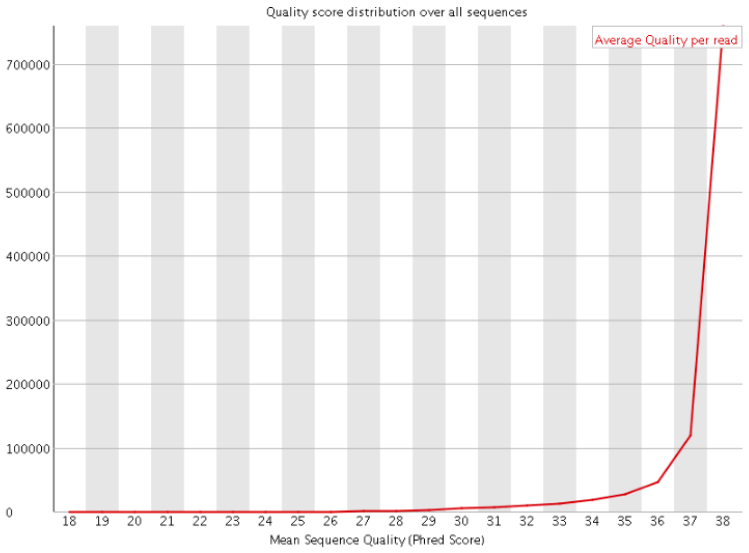


Datos trimados para primers y adaptadores R1

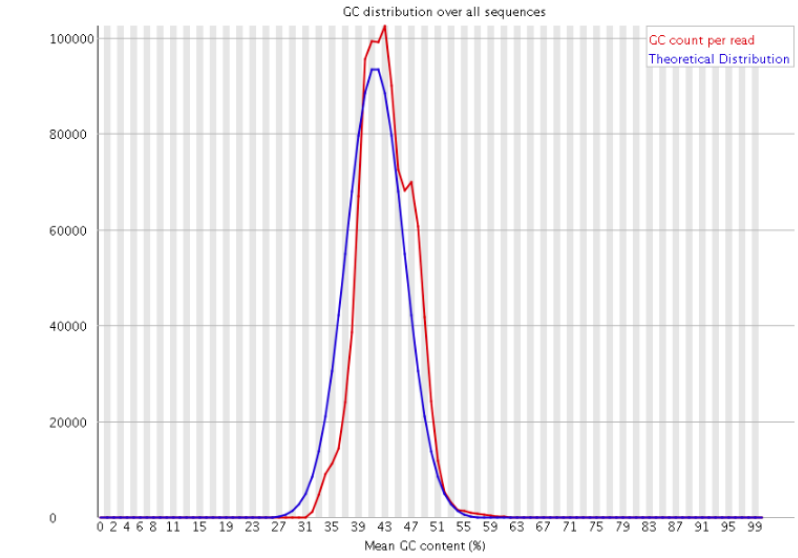
Per base sequence quality



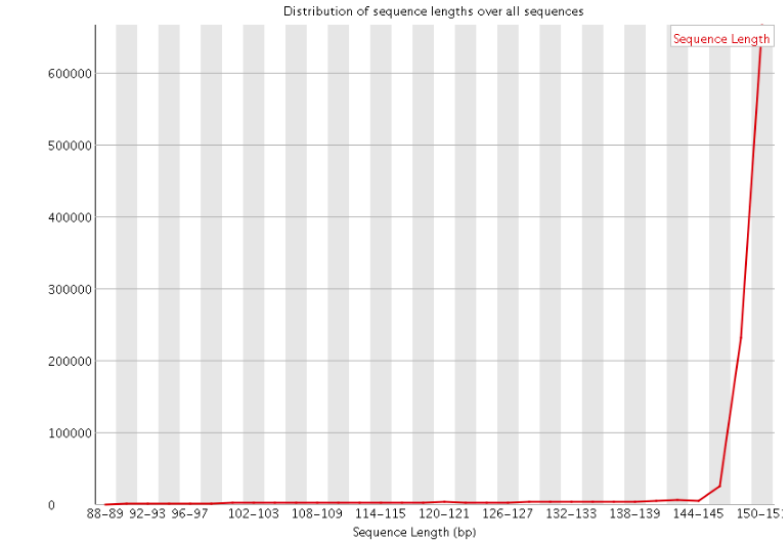
Per sequence quality scores



Per sequence GC content

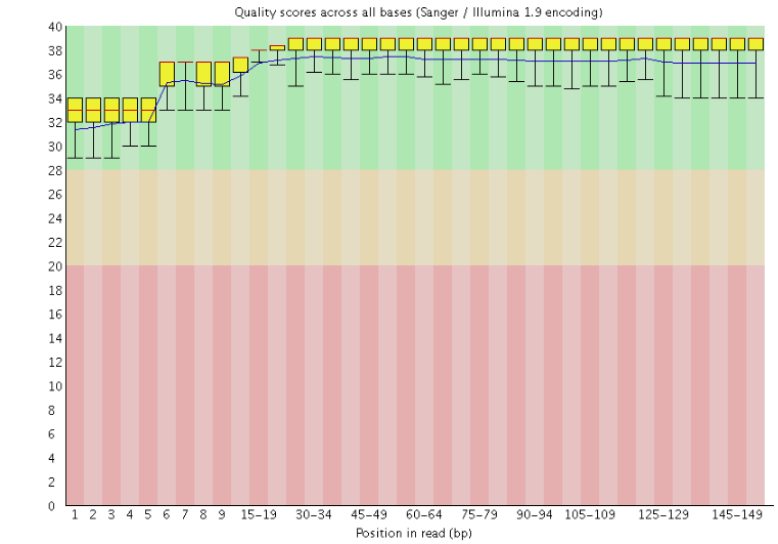


Sequence Length Distribution

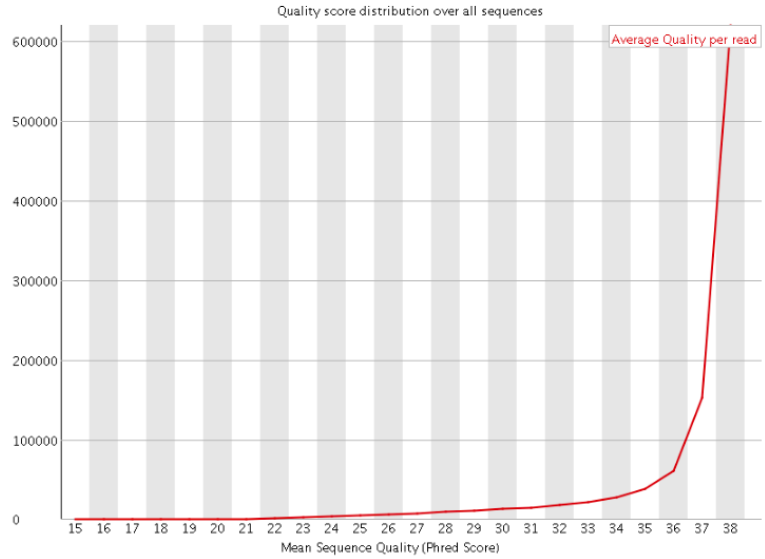


Datos trimados para primers y adaptadores R2

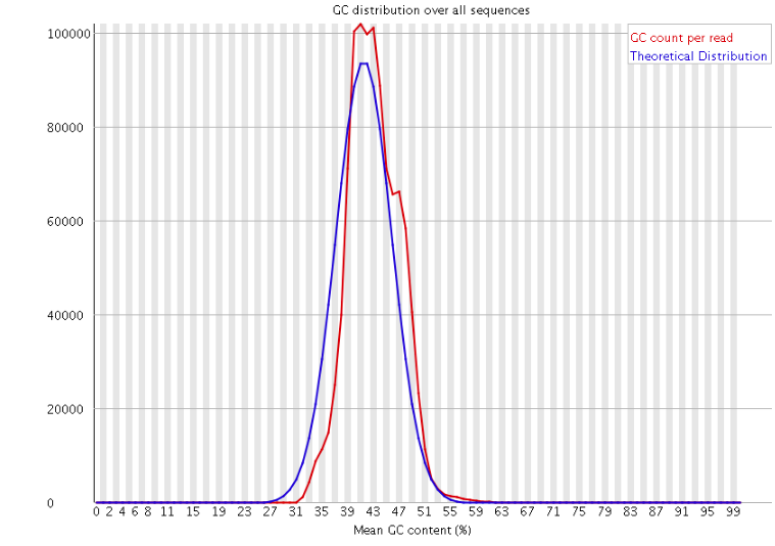
Per base sequence quality



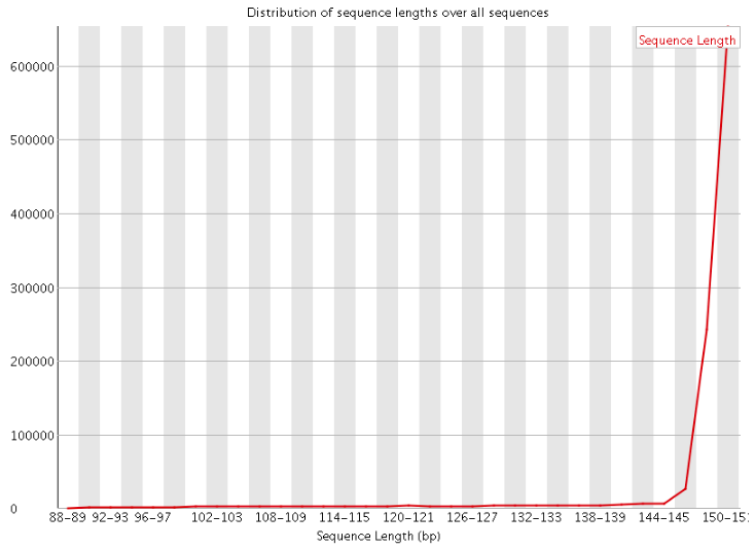
Per sequence quality scores



Per sequence GC content

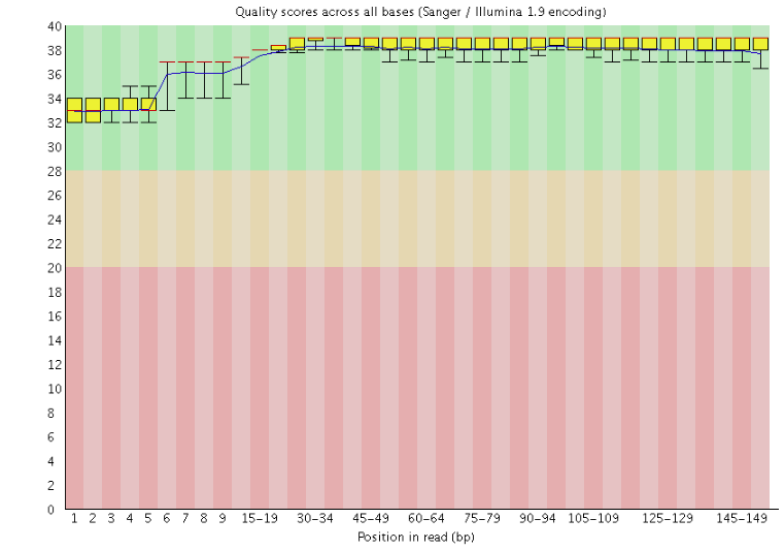


Sequence Length Distribution

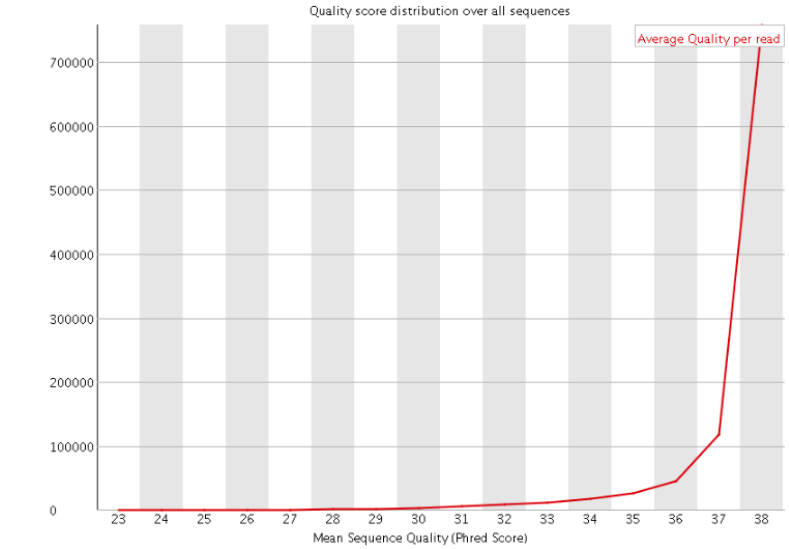


Datos trimados por calidad R1

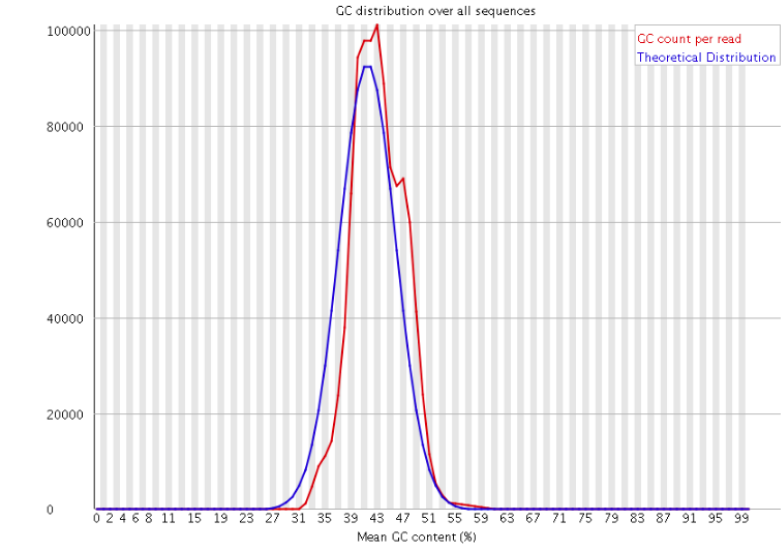
Per base sequence quality



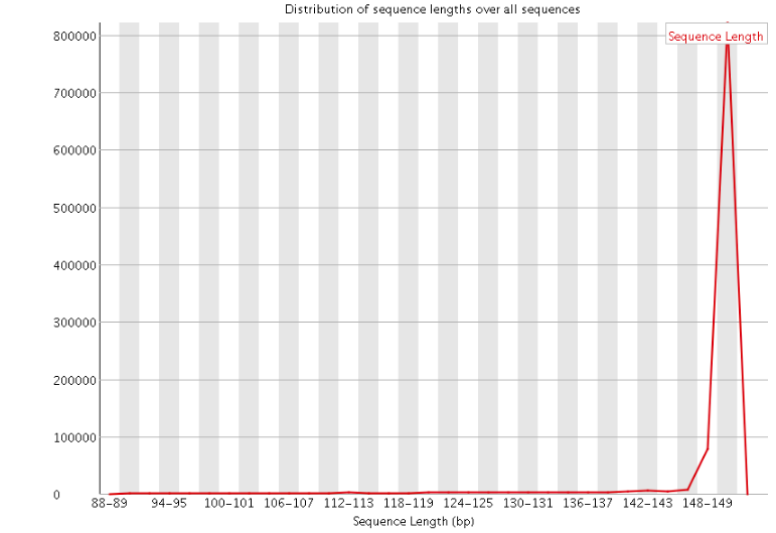
Per sequence quality scores



Per sequence GC content

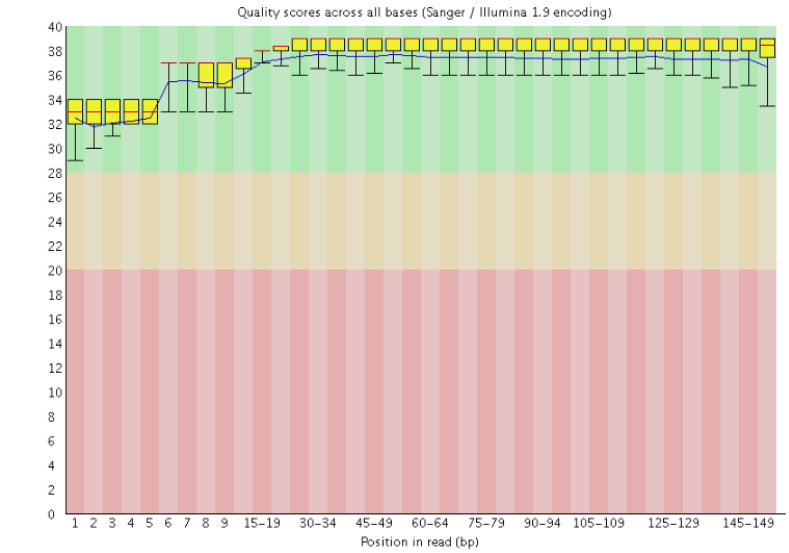


Sequence Length Distribution

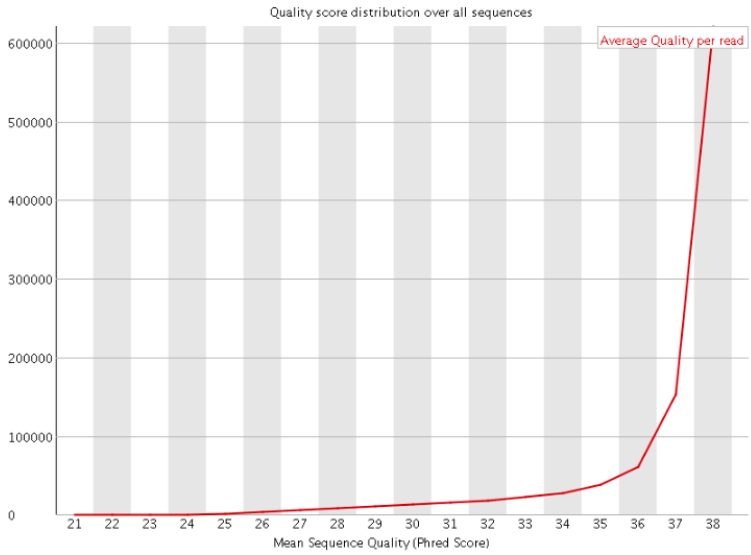


Datos trimados por calidad R2

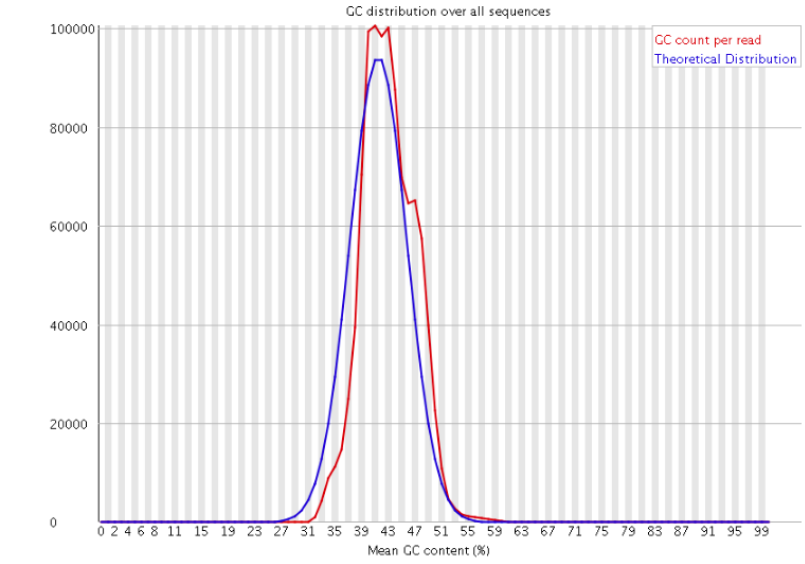
Per base sequence quality



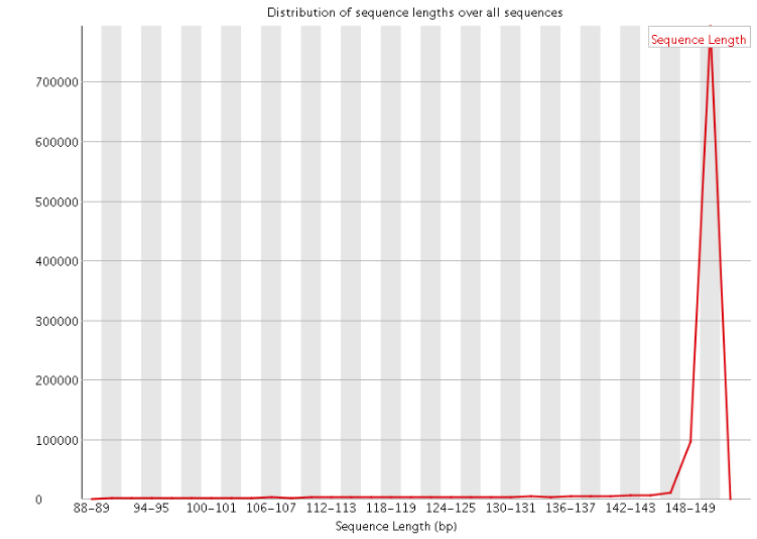
Per sequence quality scores



Per sequence GC content



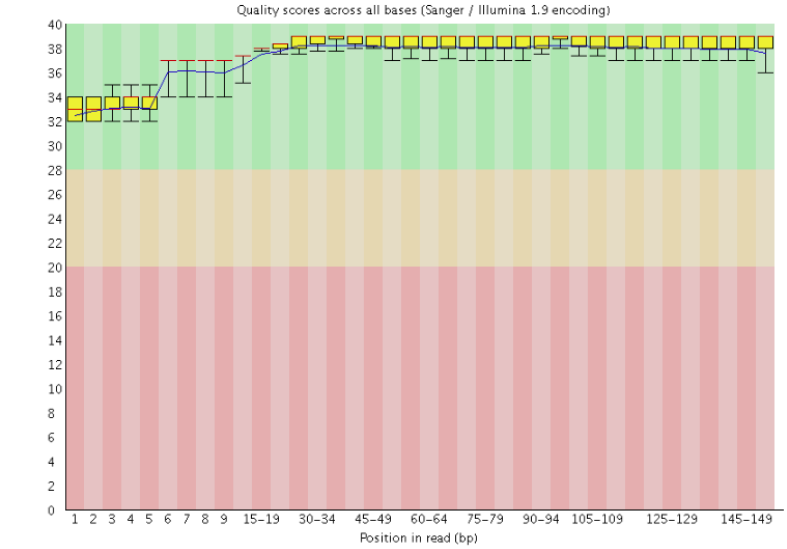
Sequence Length Distribution



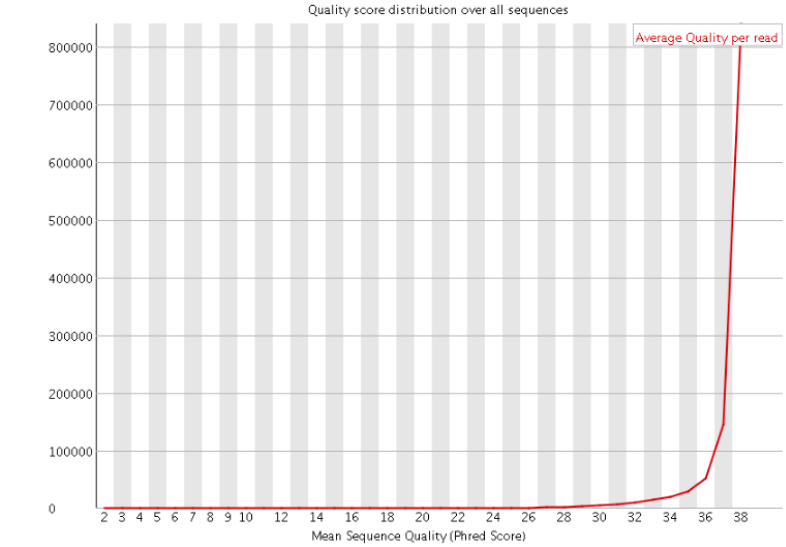
m159:

Datos Crudos R1

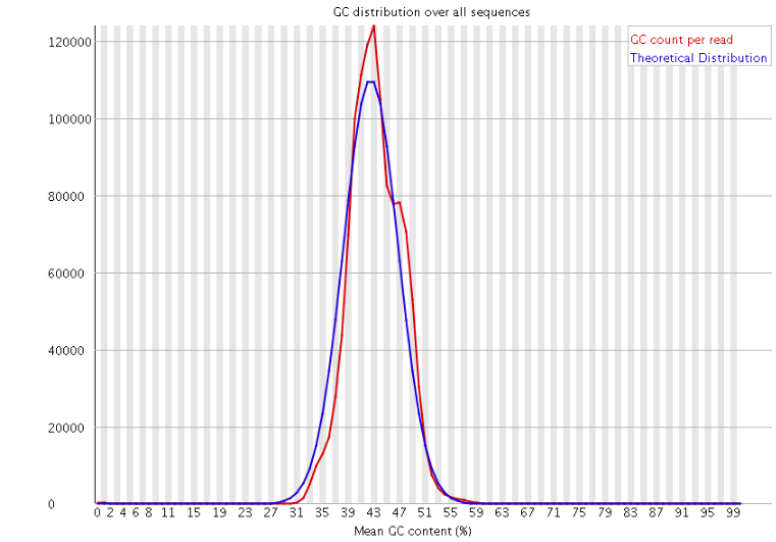
Per base sequence quality



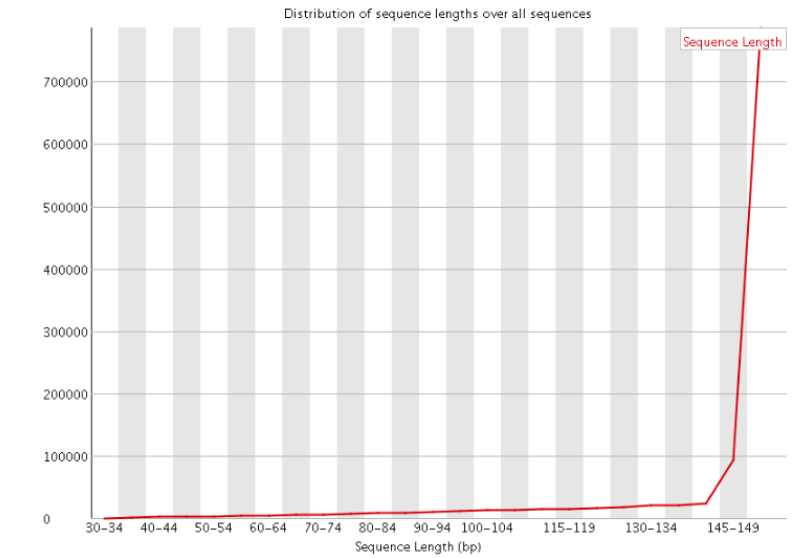
Per sequence quality scores



Per sequence GC content

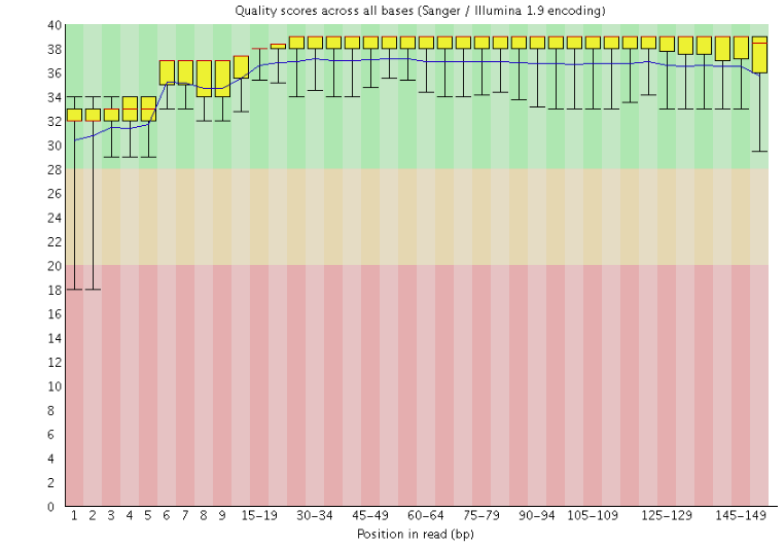


Sequence Length Distribution

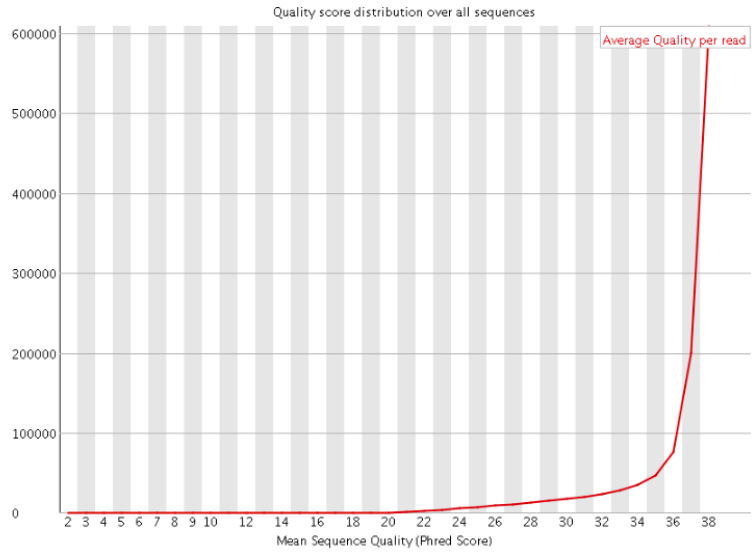


Datos Crudos R2

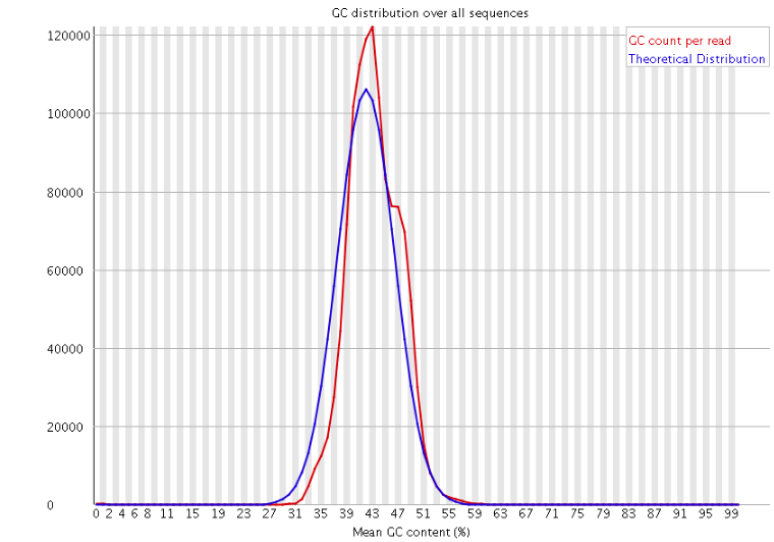
Per base sequence quality



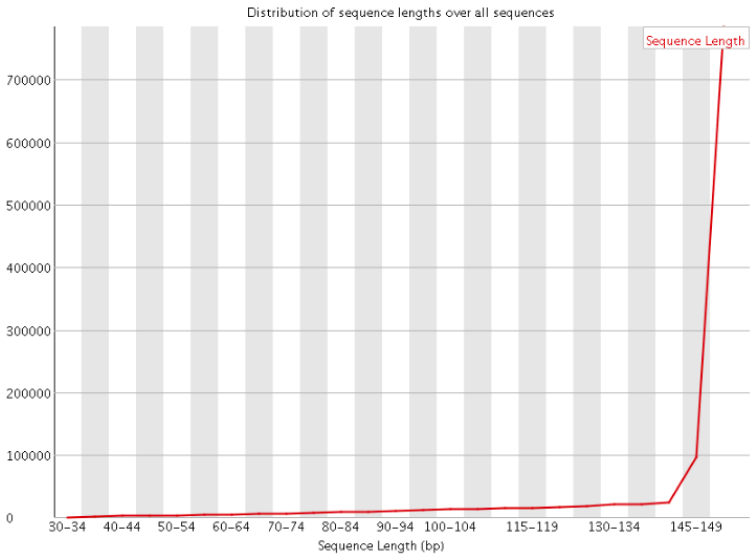
Per sequence quality scores



Per sequence GC content

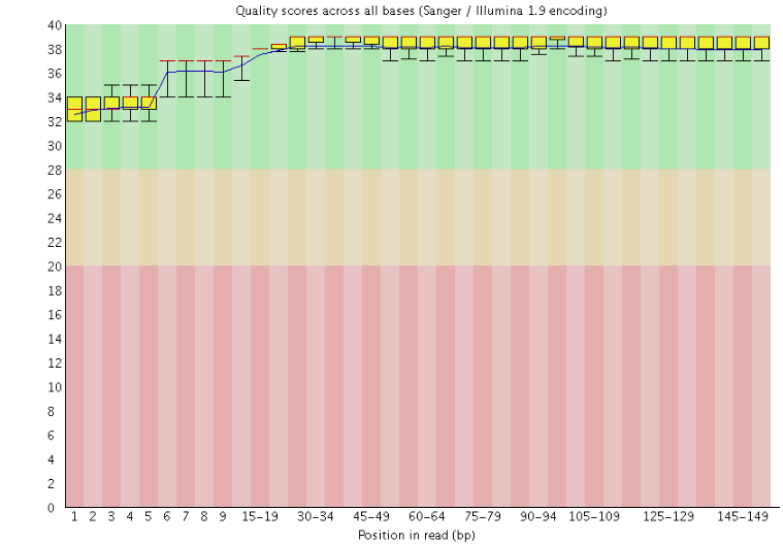


Sequence Length Distribution

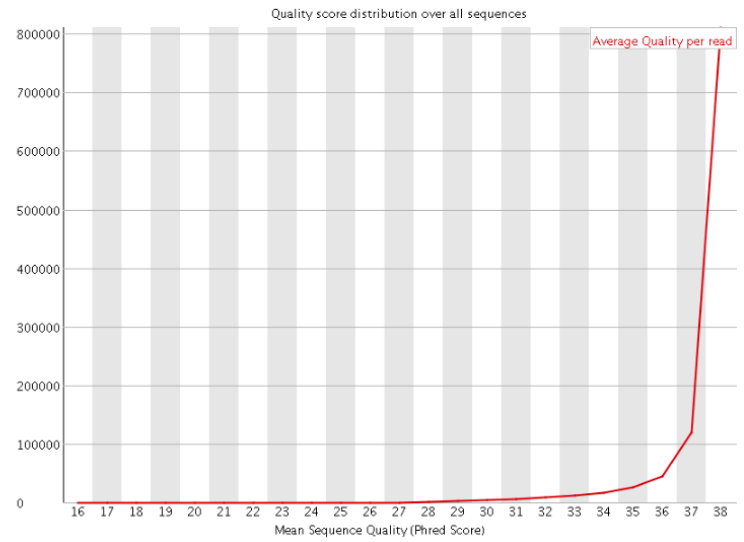


Datos trimados para primers y adaptadores R1

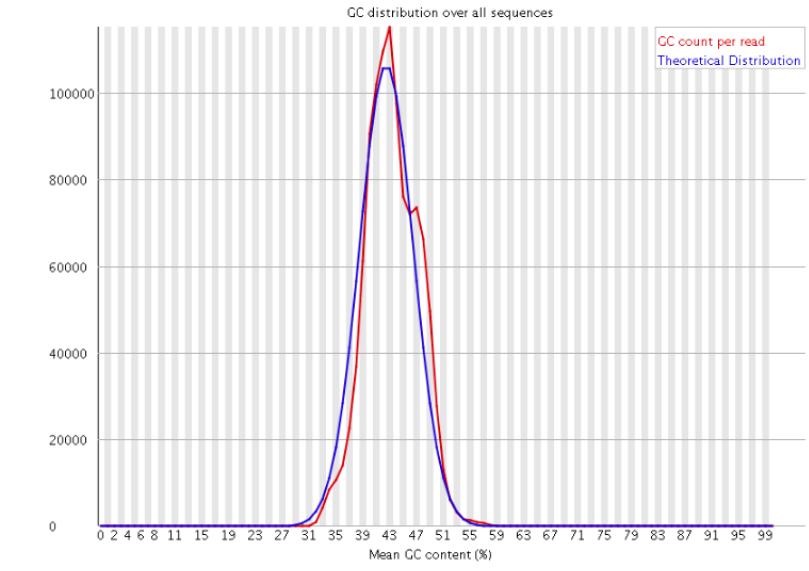
Per base sequence quality



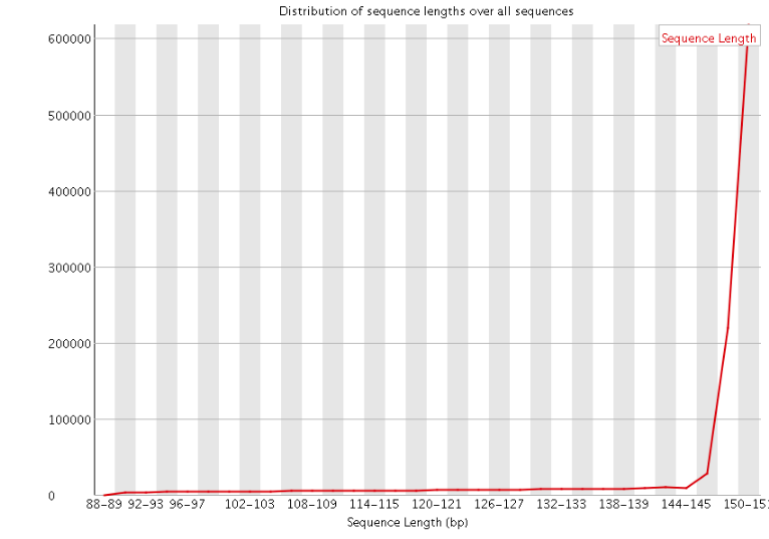
Per sequence quality scores



Per sequence GC content

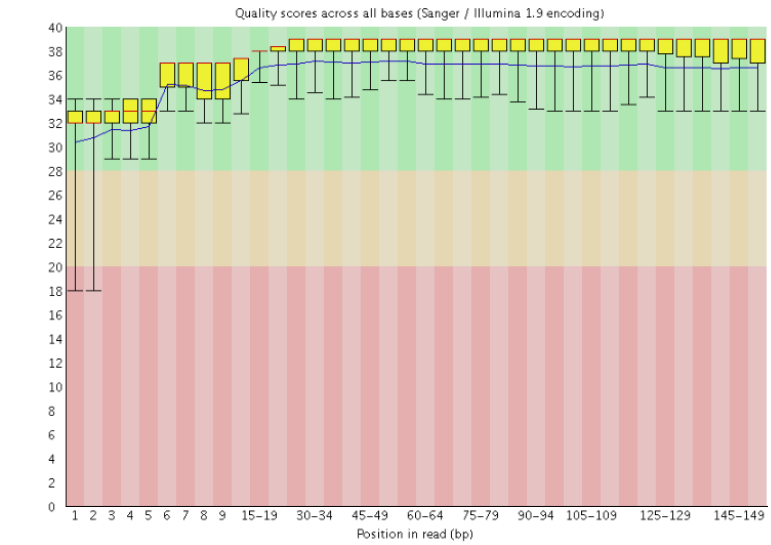


Sequence Length Distribution

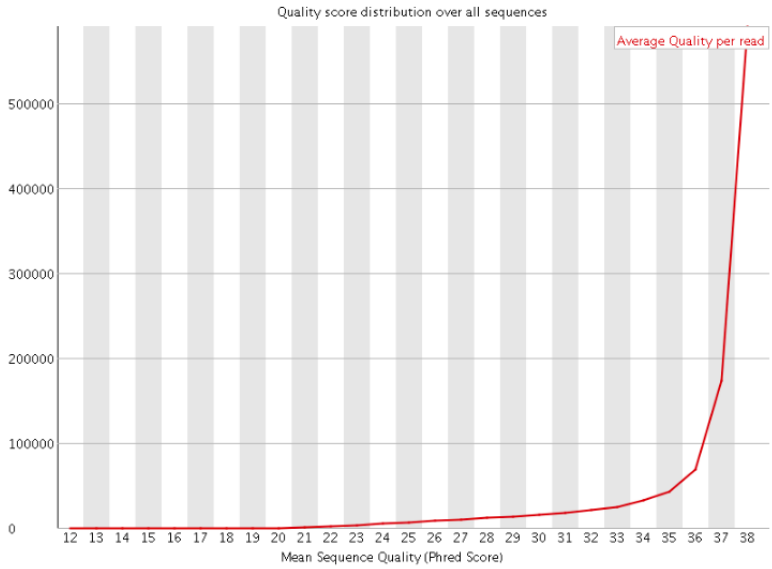


Datos trimados para primers y adaptadores R2

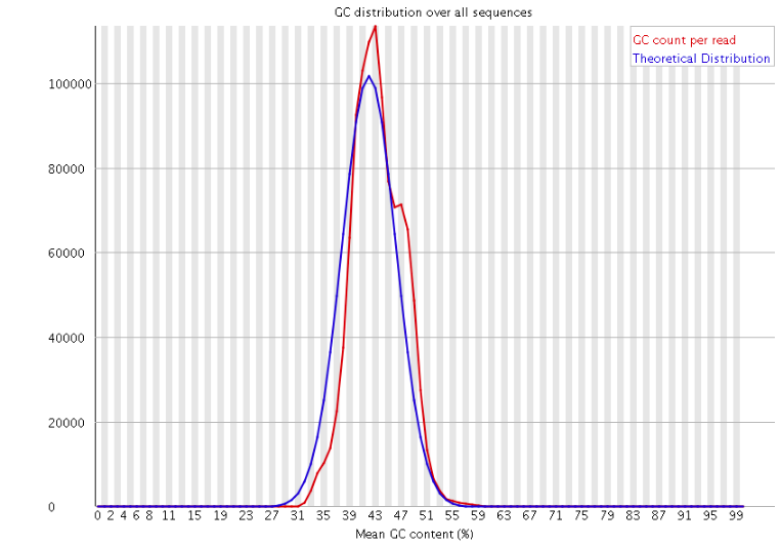
Per base sequence quality



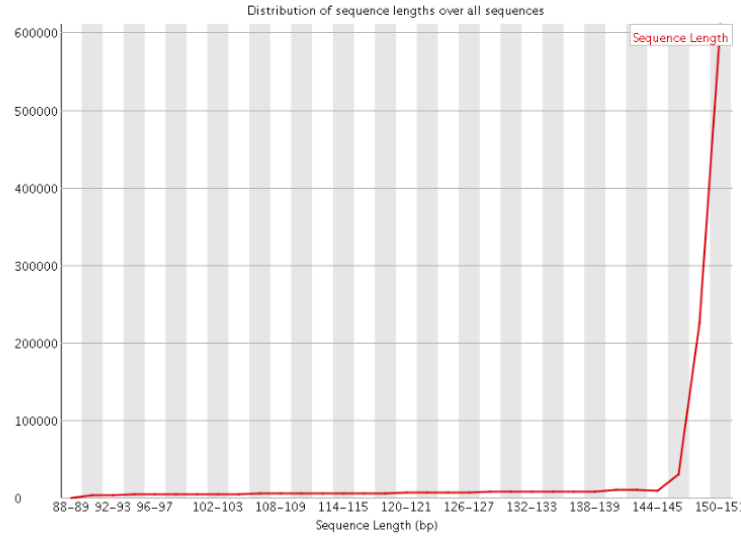
Per sequence quality scores



Per sequence GC content

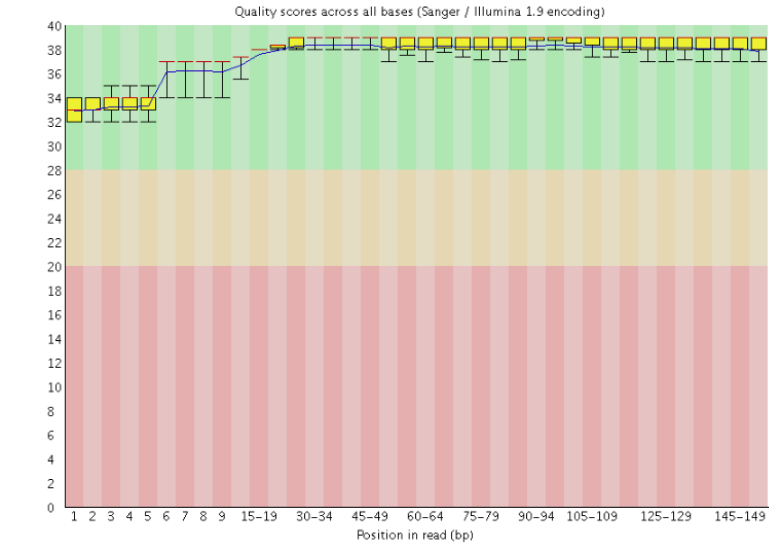


Sequence Length Distribution

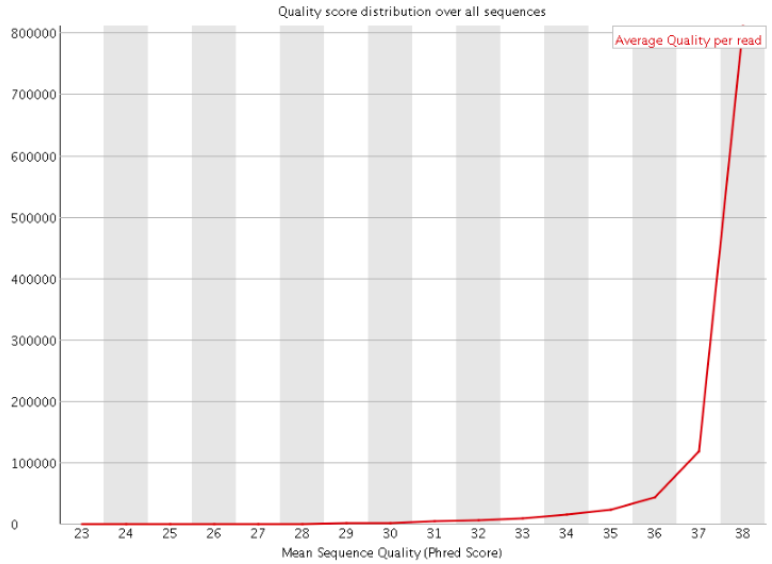


Datos trimados por calidad R1

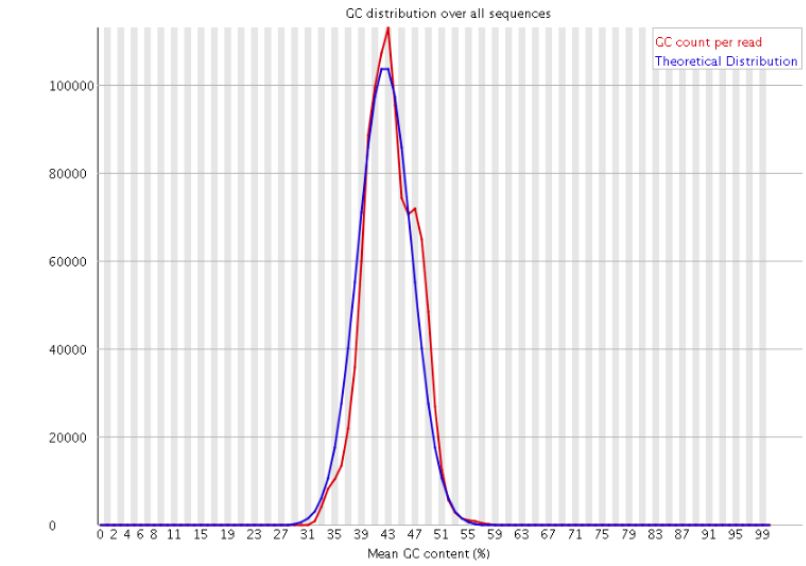
Per base sequence quality



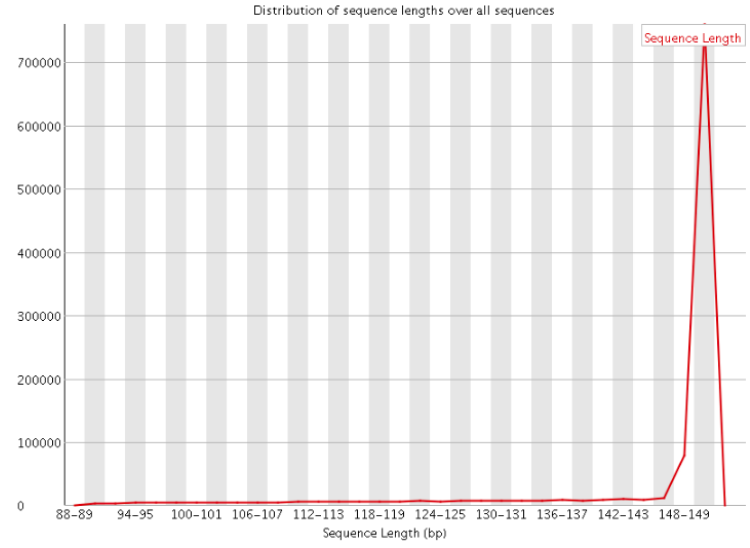
Per sequence quality scores



Per sequence GC content

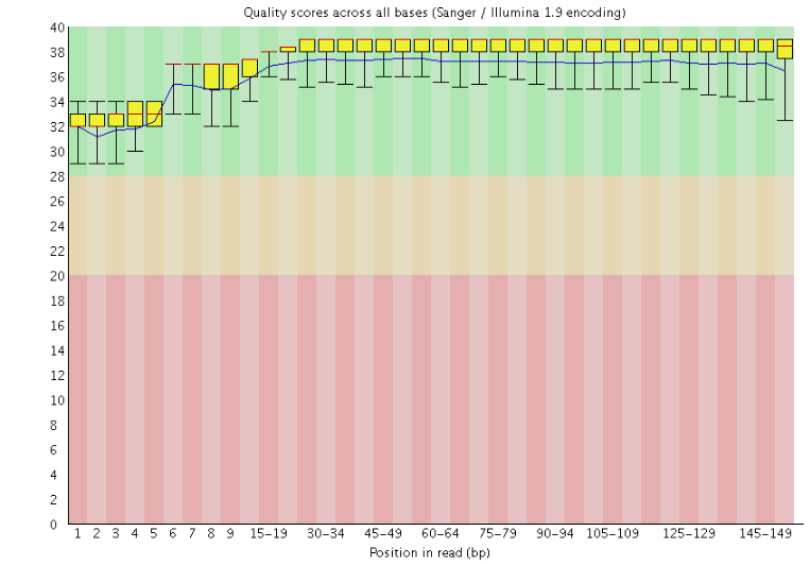


Sequence Length Distribution

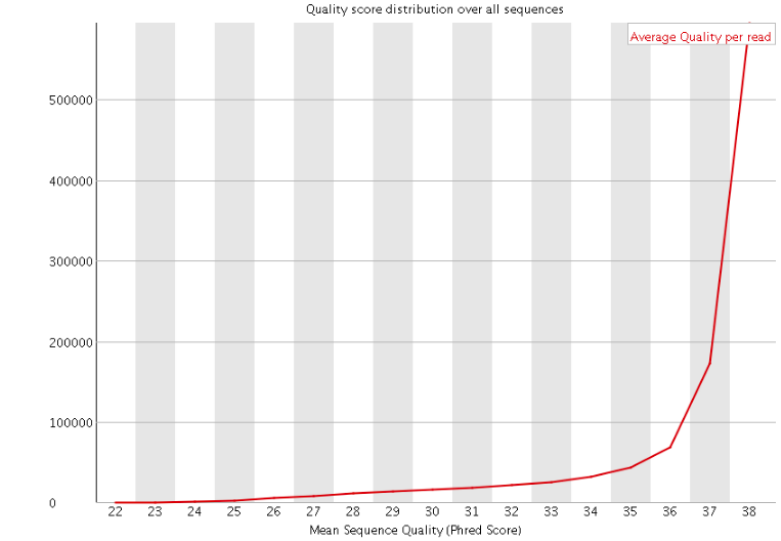


Datos trimados por calidad R2

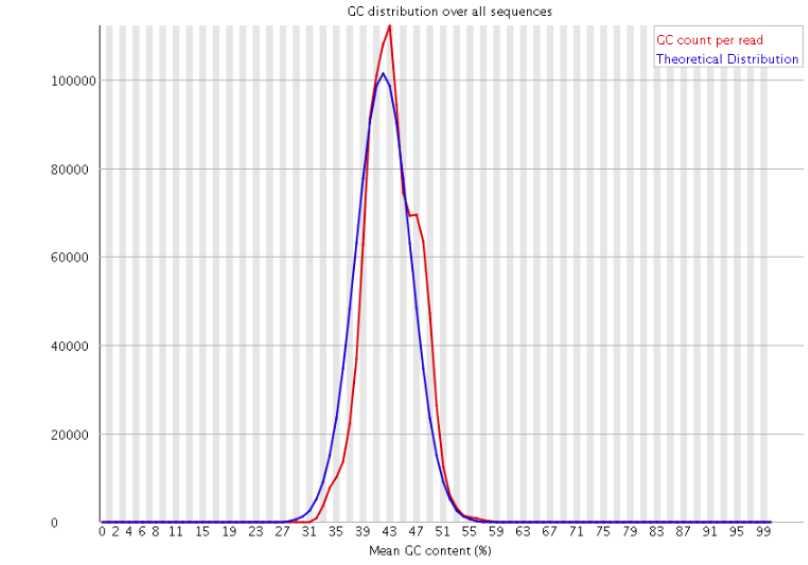
Per base sequence quality



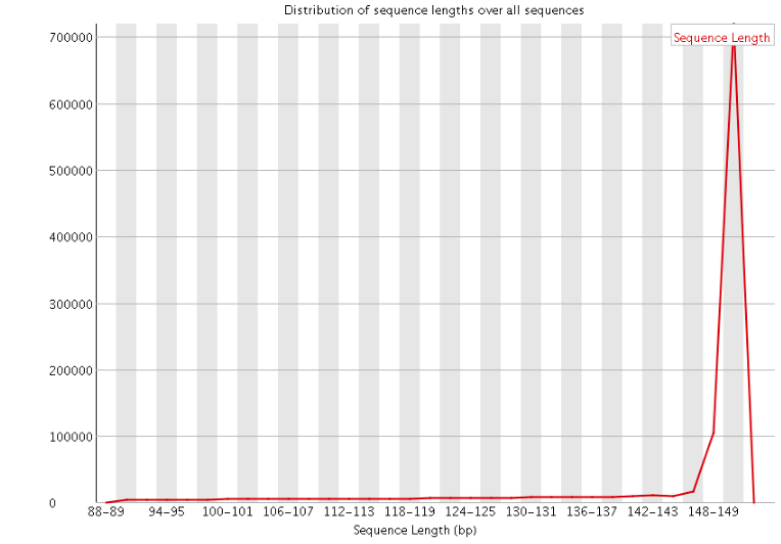
Per sequence quality scores



Per sequence GC content



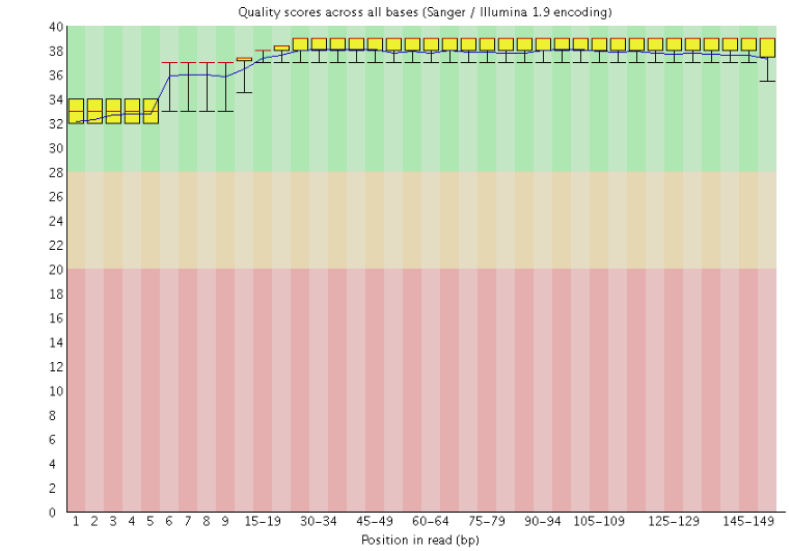
Sequence Length Distribution



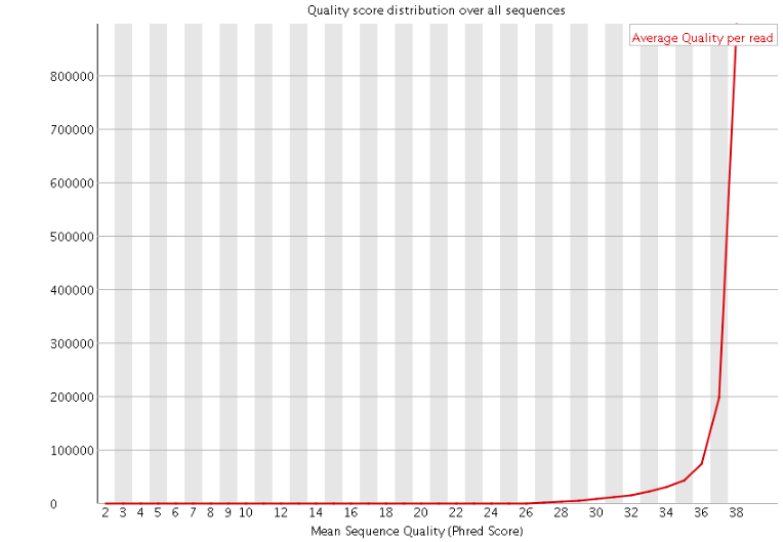
m202:

Datos Crudos R1

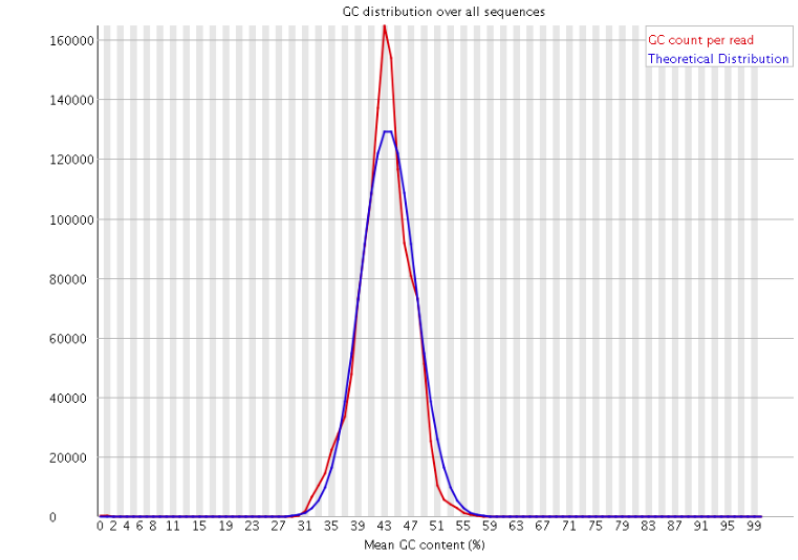
Per base sequence quality



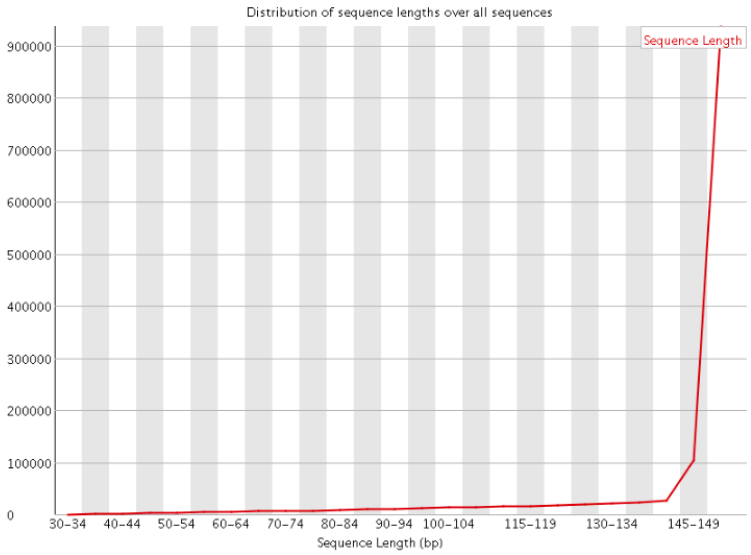
Per sequence quality scores



Per sequence GC content

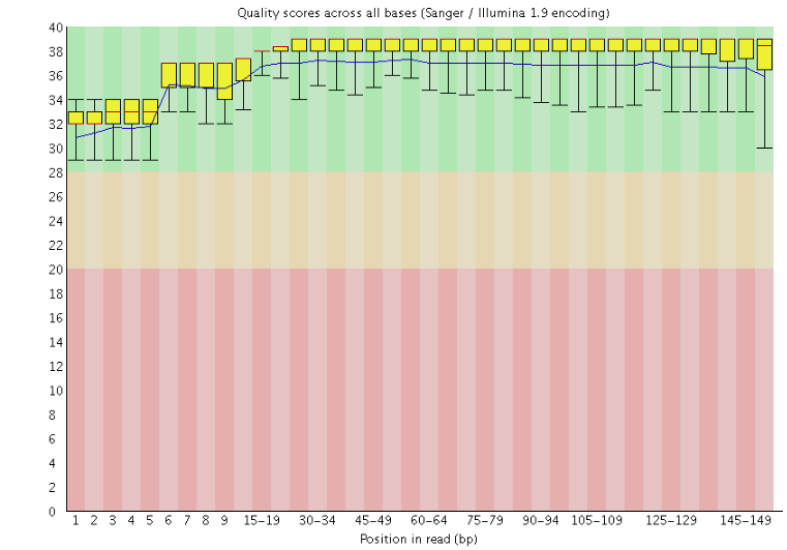


Sequence Length Distribution

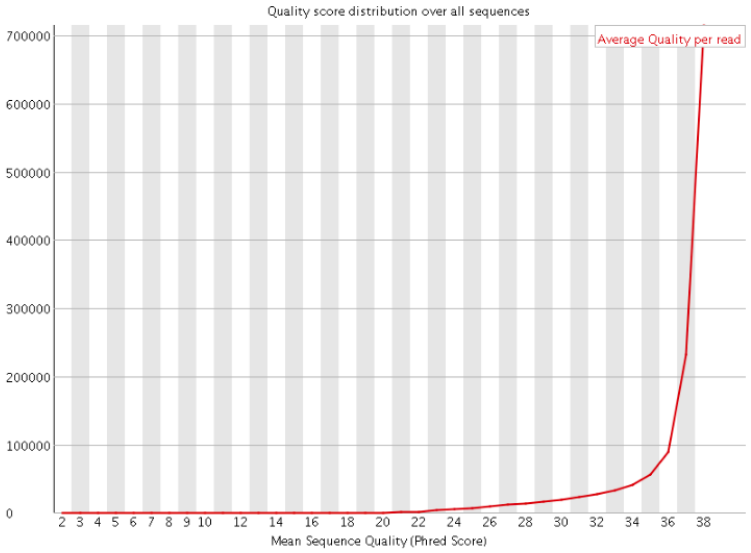


Datos Crudos R2

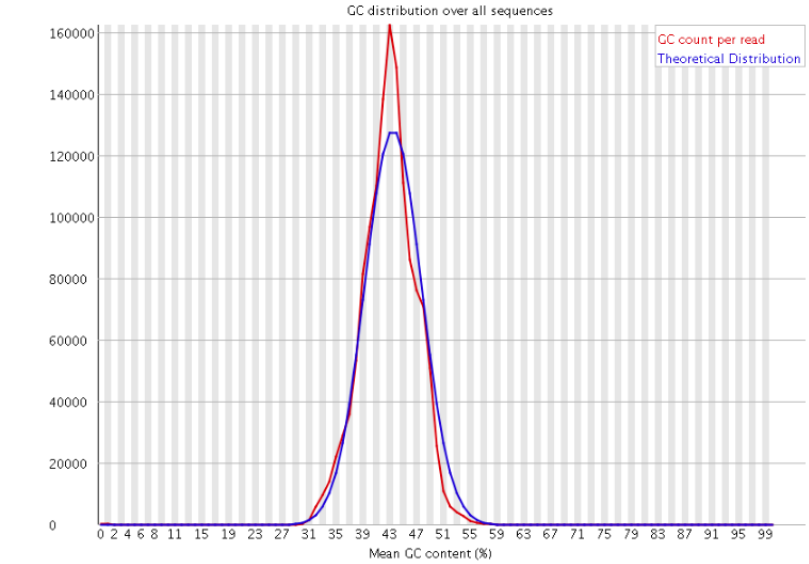
Per base sequence quality



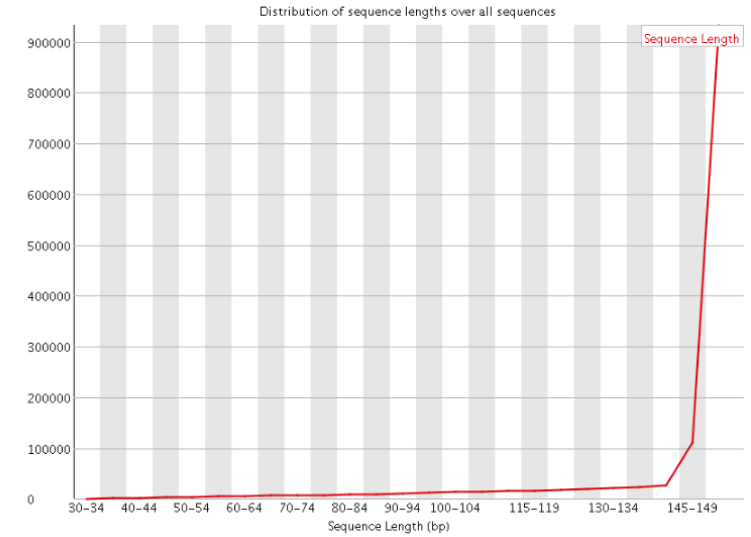
Per sequence quality scores



Per sequence GC content

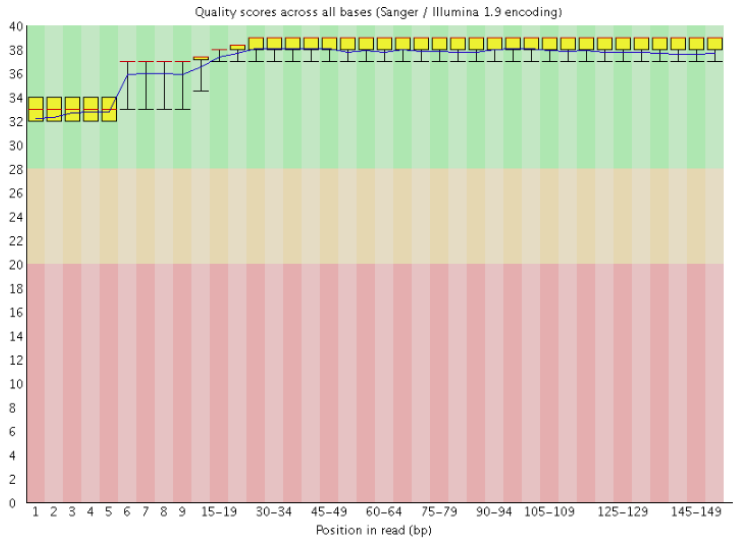


Sequence Length Distribution

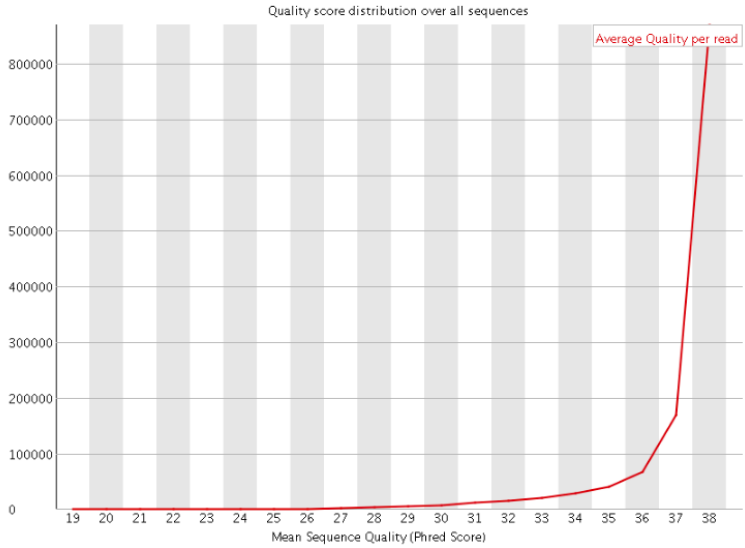


Datos trimados para primers y adaptadores R1

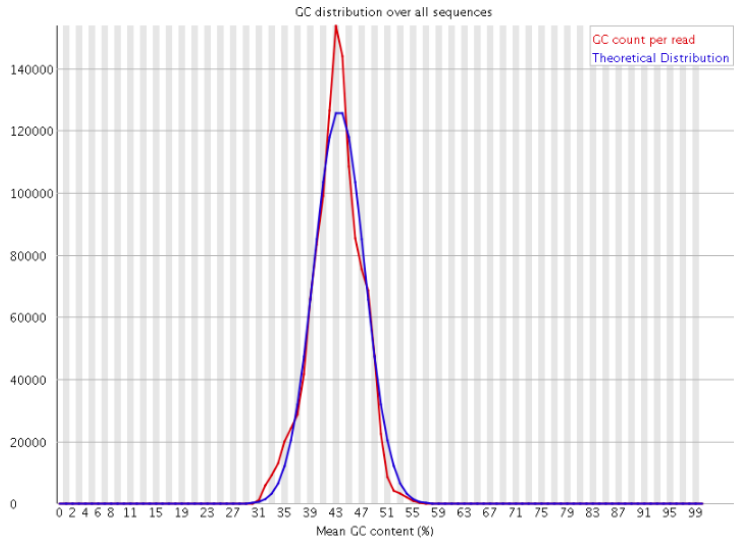
Per base sequence quality



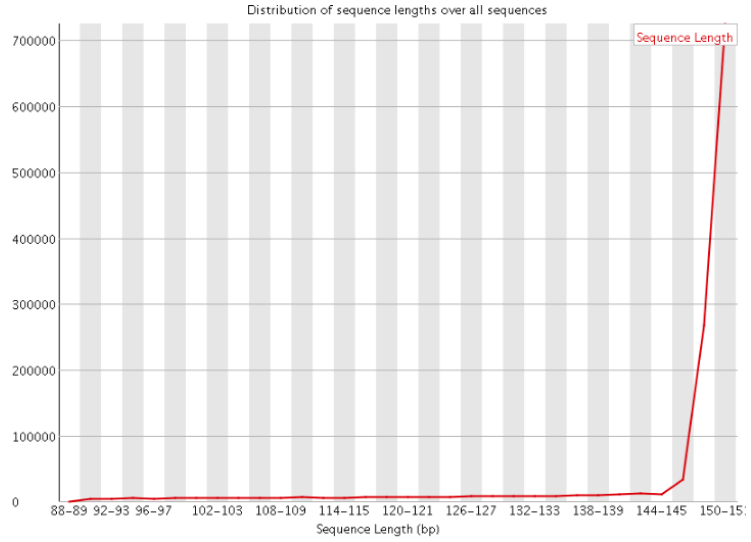
Per sequence quality scores



Per sequence GC content

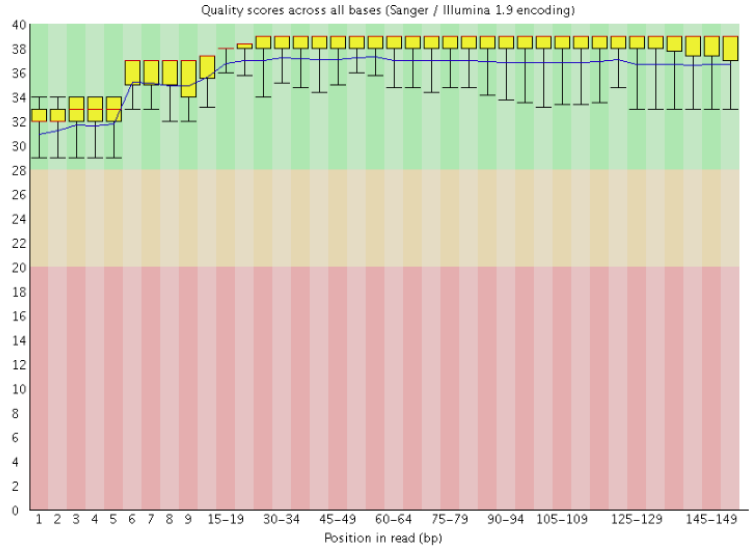


Sequence Length Distribution

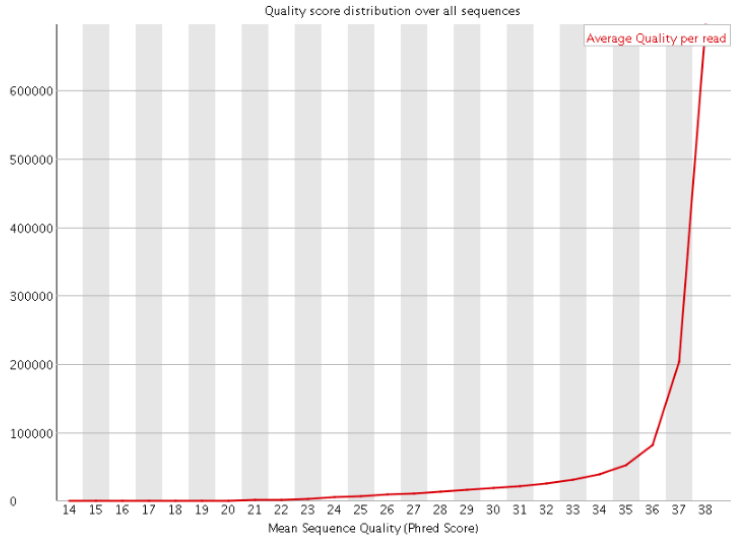


Datos trimados para primers y adaptadores R2

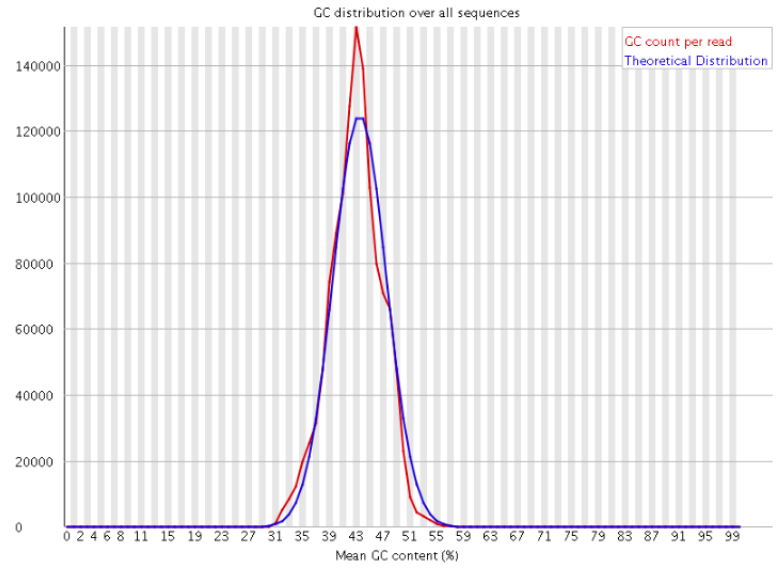
Per base sequence quality



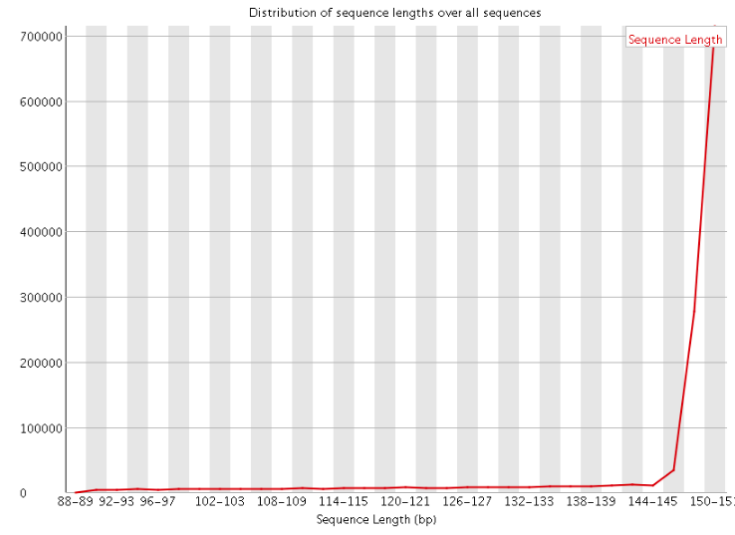
Per sequence quality scores



Per sequence GC content

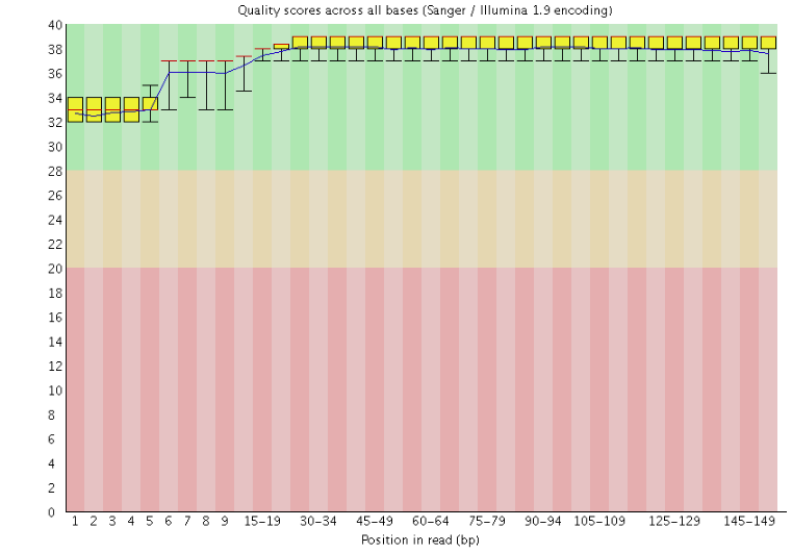


Sequence Length Distribution

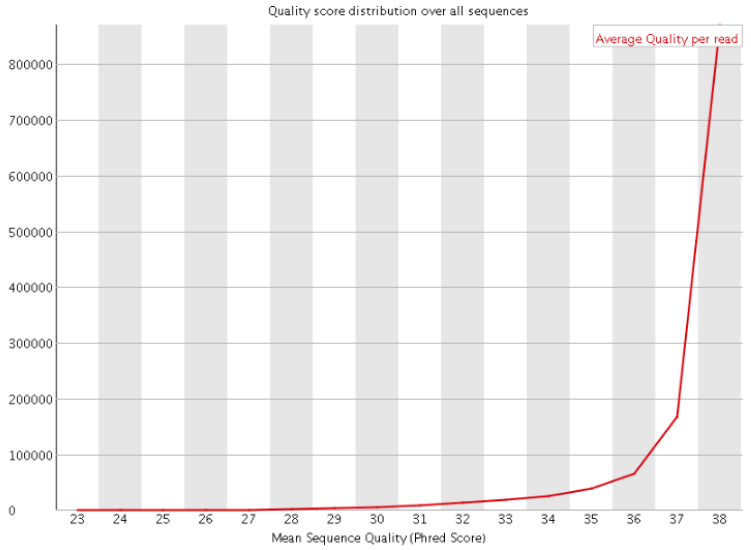


Datos trimados por calidad R1

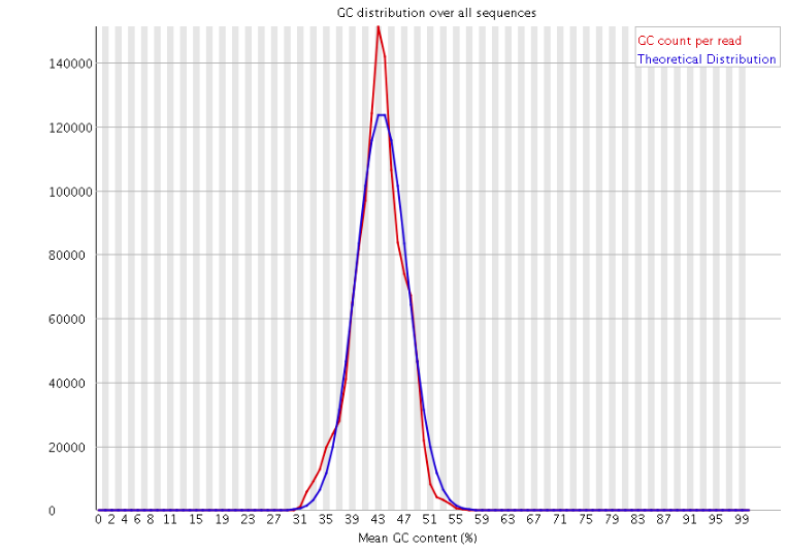
Per base sequence quality



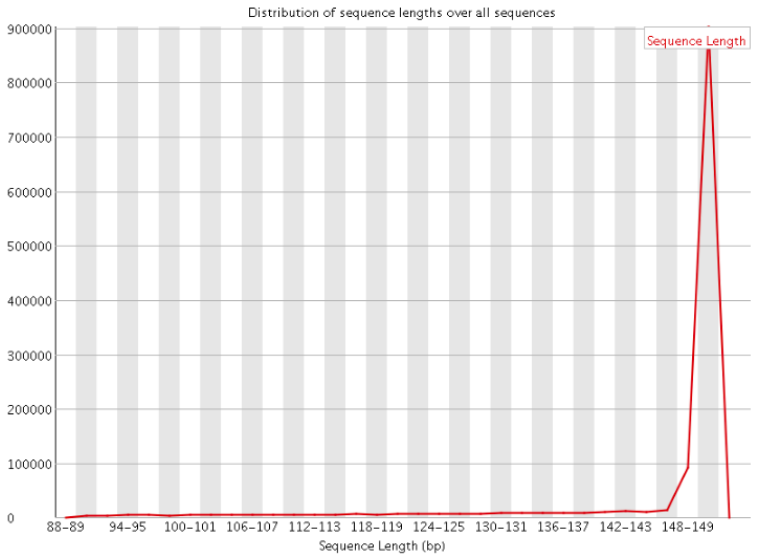
Per sequence quality scores



Per sequence GC content

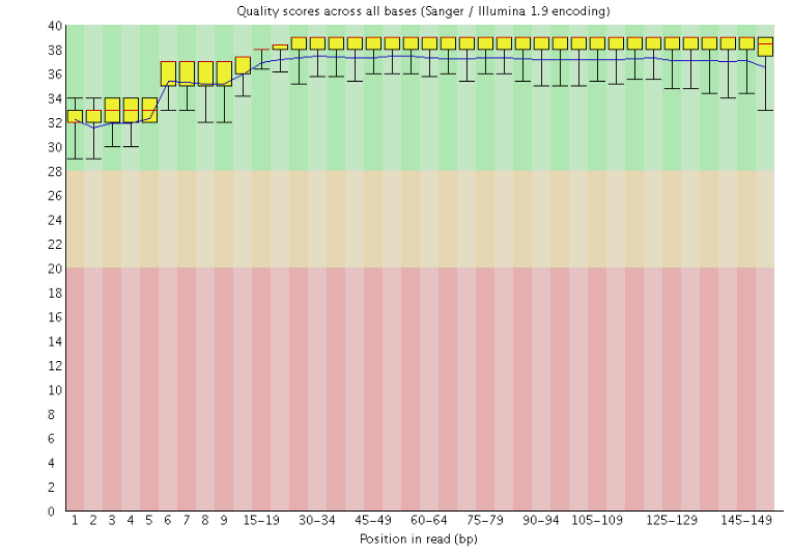


Sequence Length Distribution

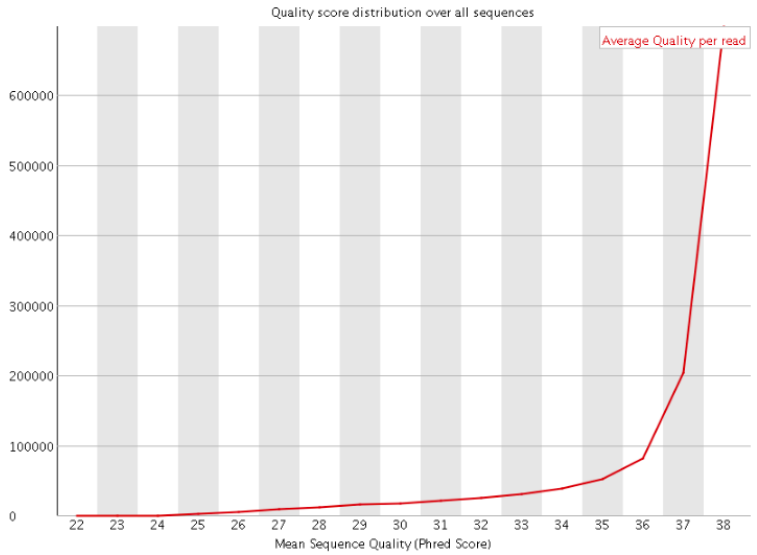


Datos trimados por calidad R2

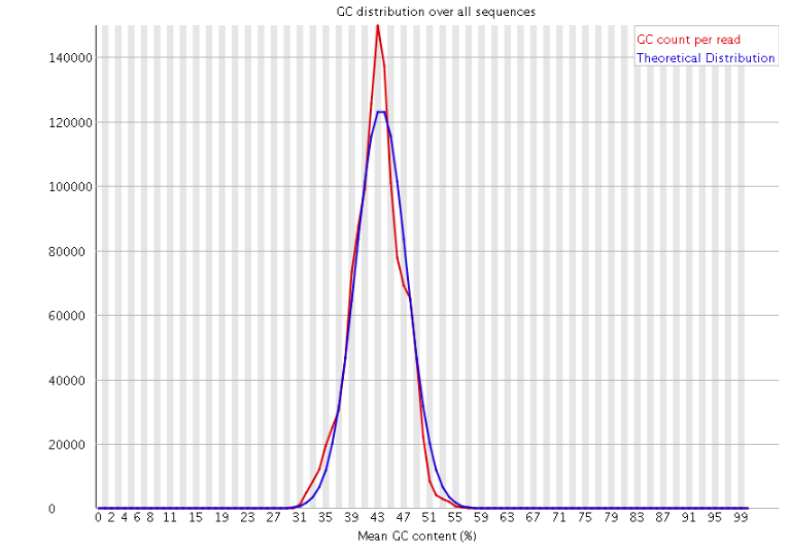
Per base sequence quality



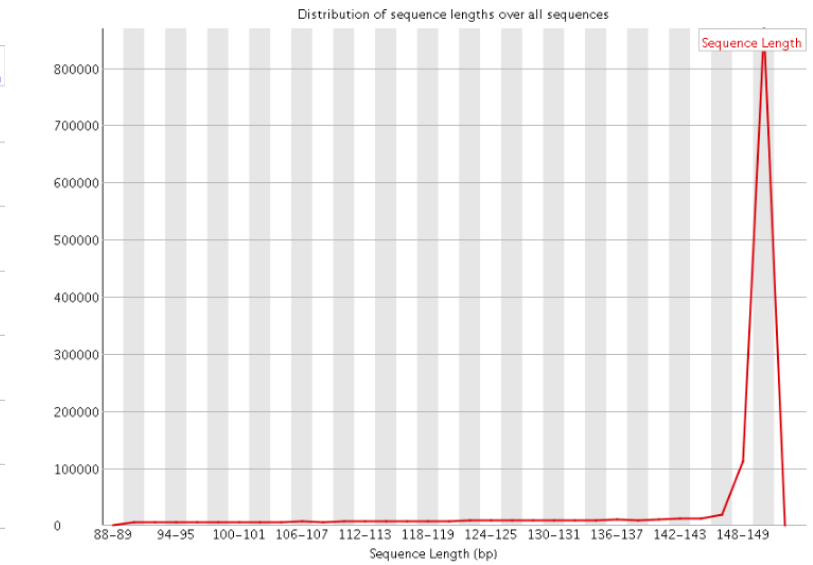
Per sequence quality scores



Per sequence GC content



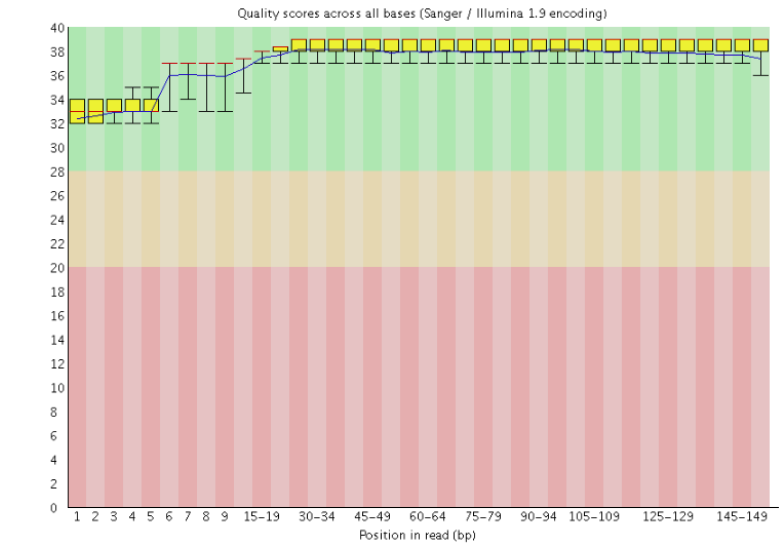
Sequence Length Distribution



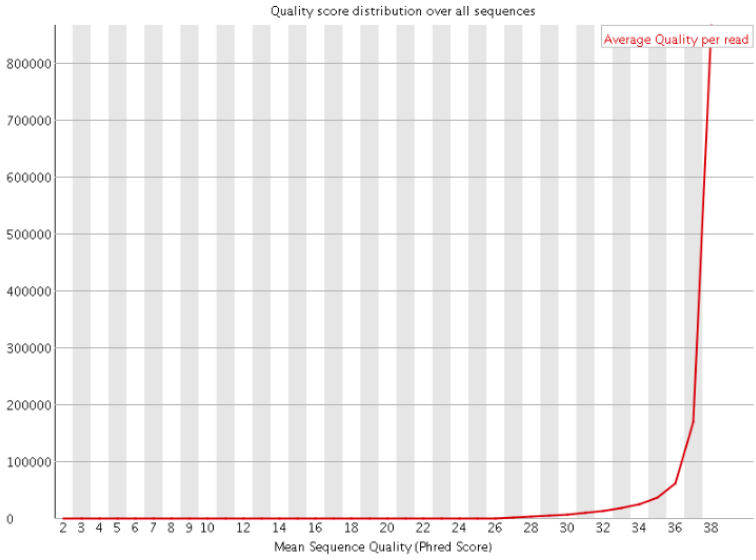
m206:

Datos Crudos R1

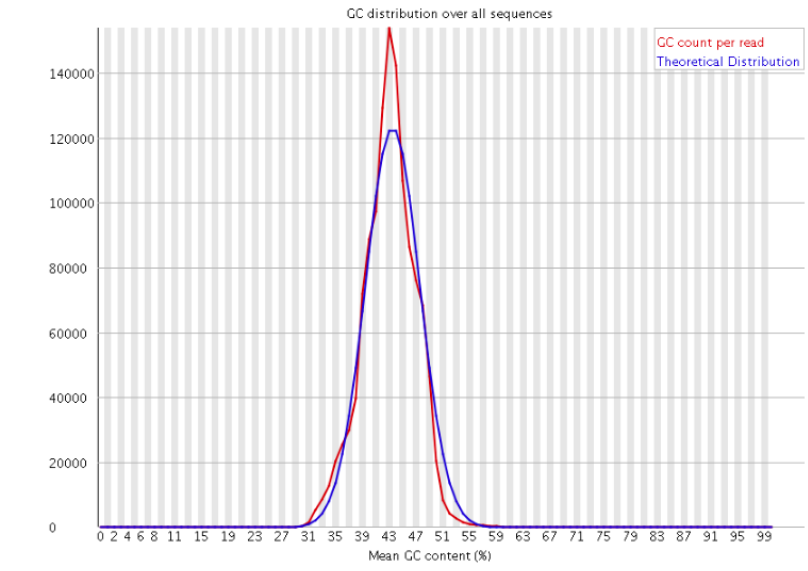
Per base sequence quality



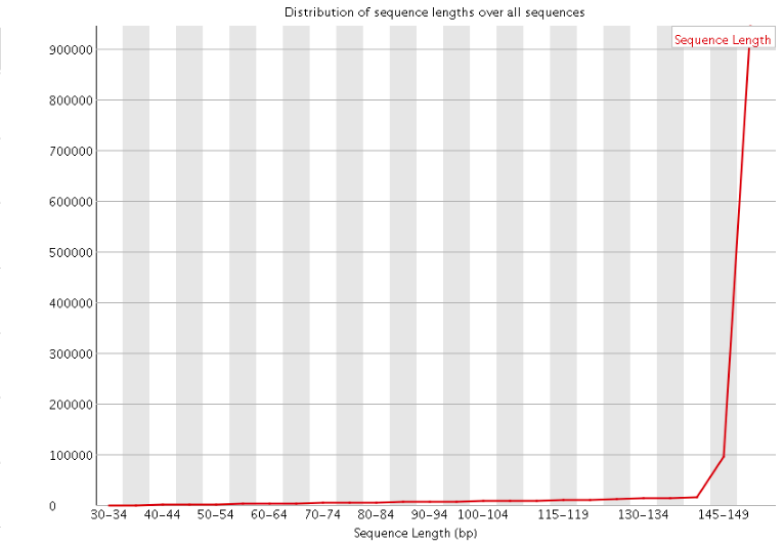
Per sequence quality scores



Per sequence GC content

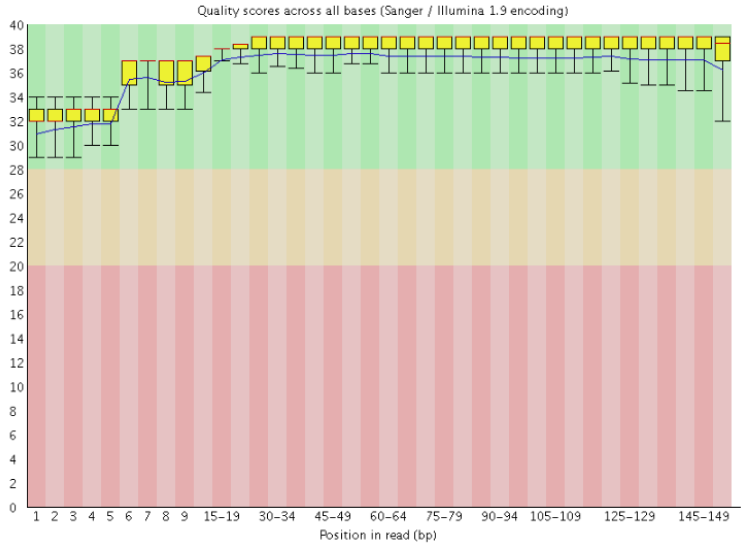


Sequence Length Distribution

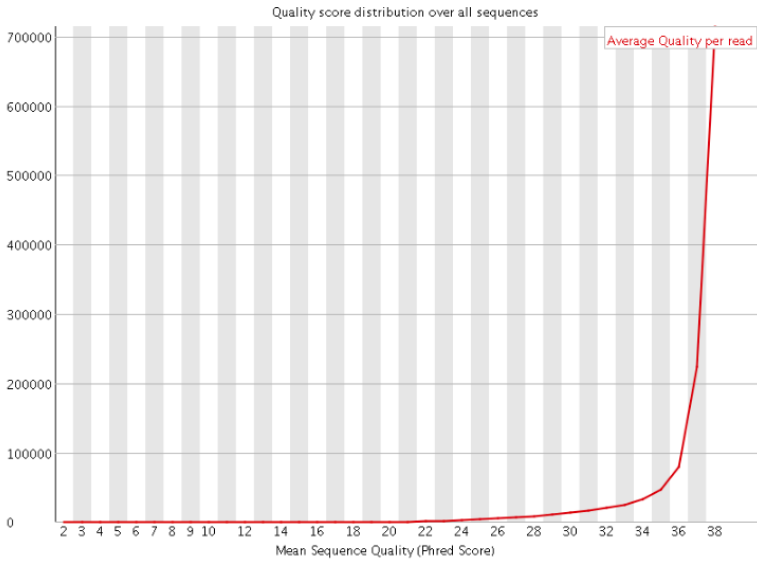


Datos Crudos R2

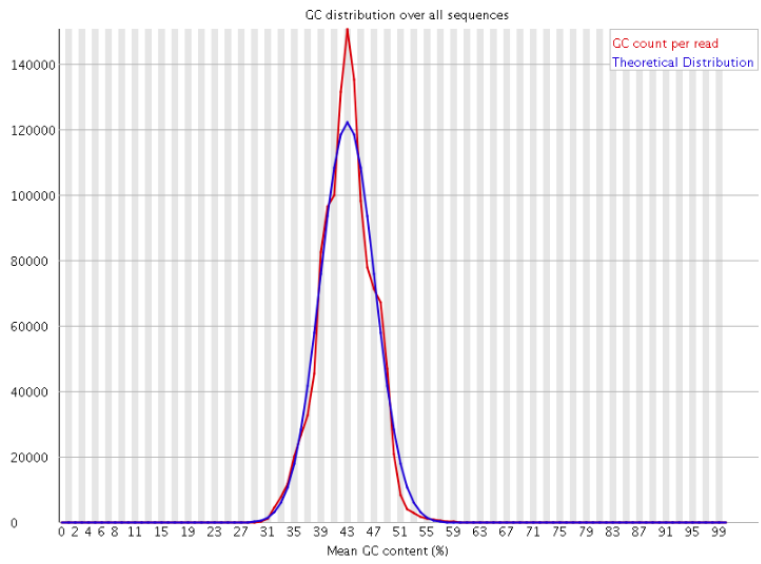
Per base sequence quality



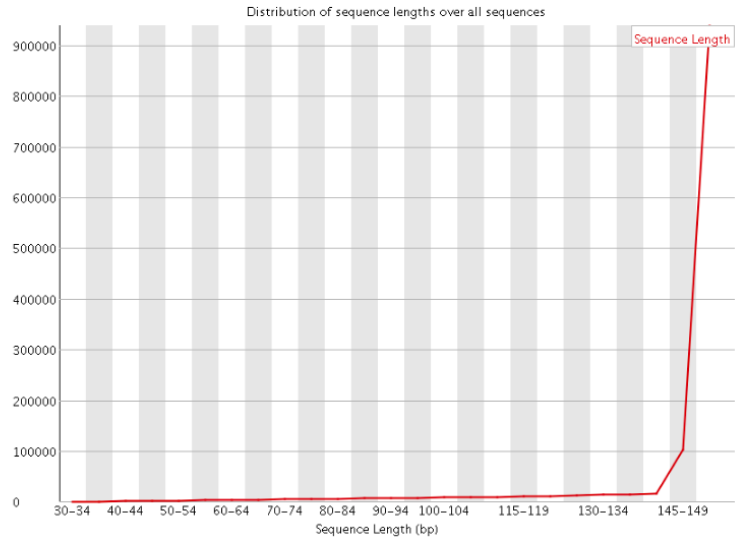
Per sequence quality scores



Per sequence GC content

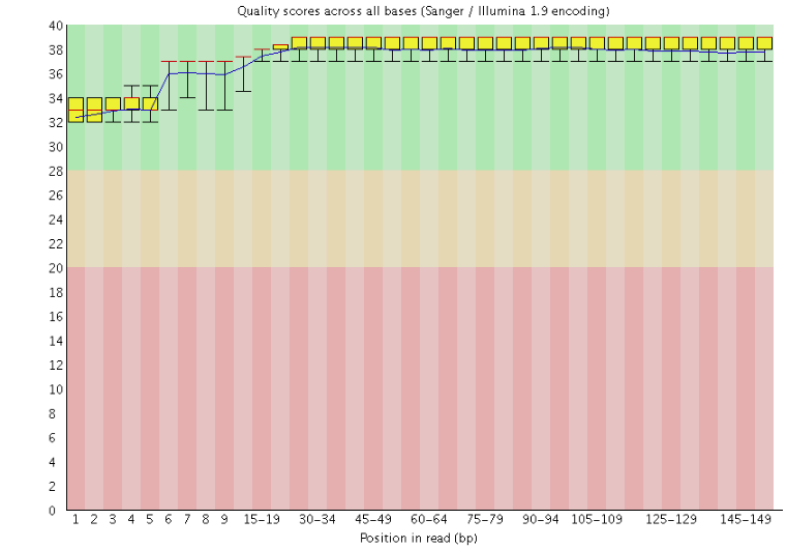


Sequence Length Distribution

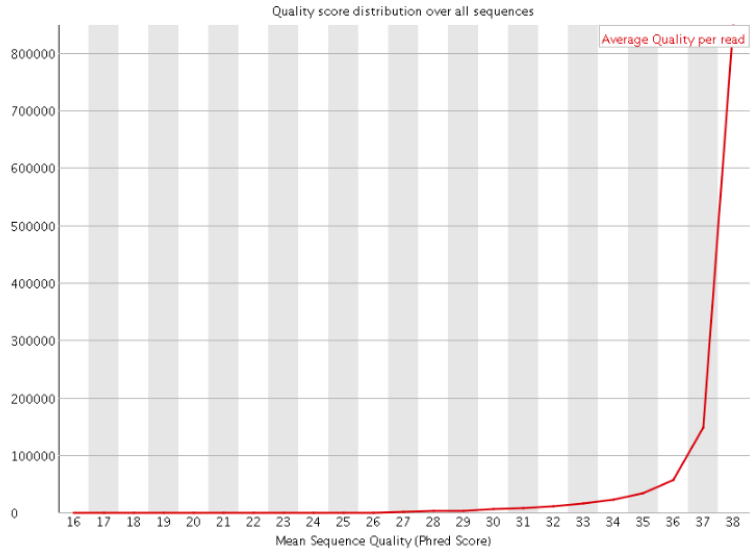


Datos trimados para primers y adaptadores R1

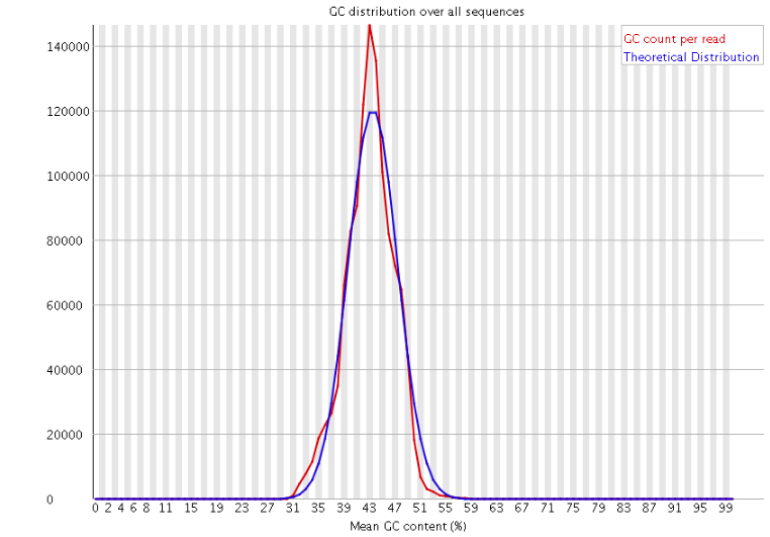
Per base sequence quality



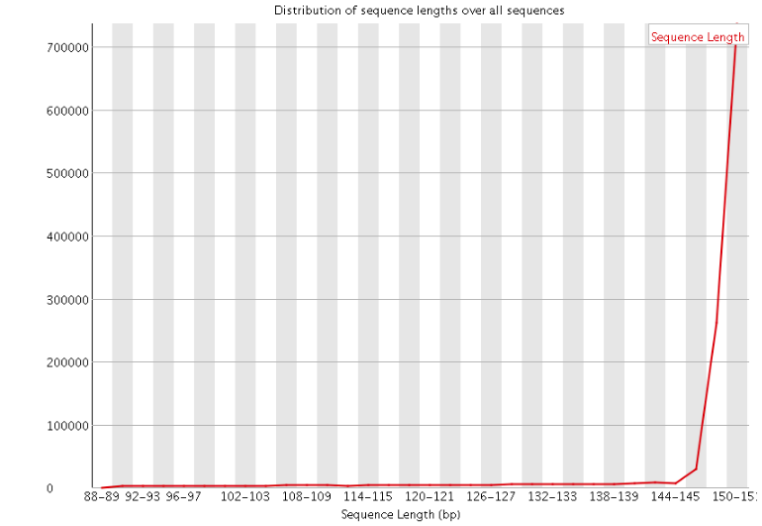
Per sequence quality scores



Per sequence GC content

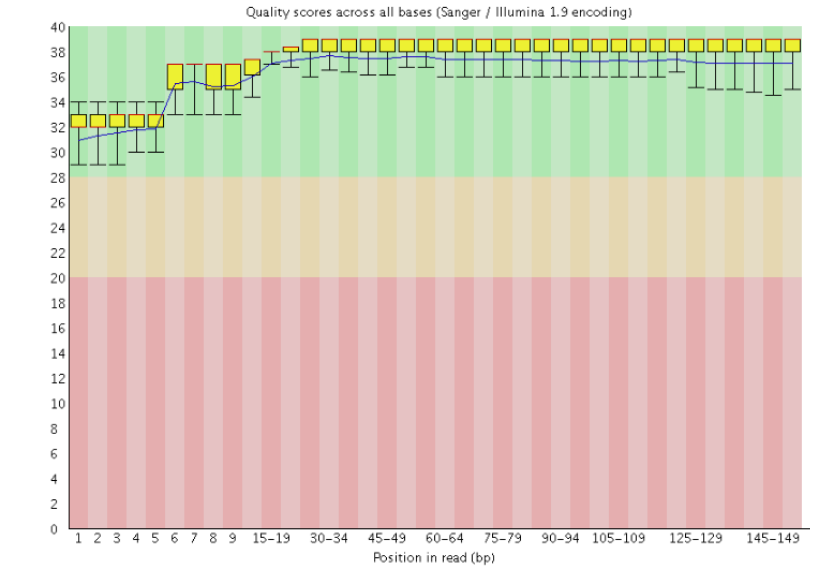


Sequence Length Distribution

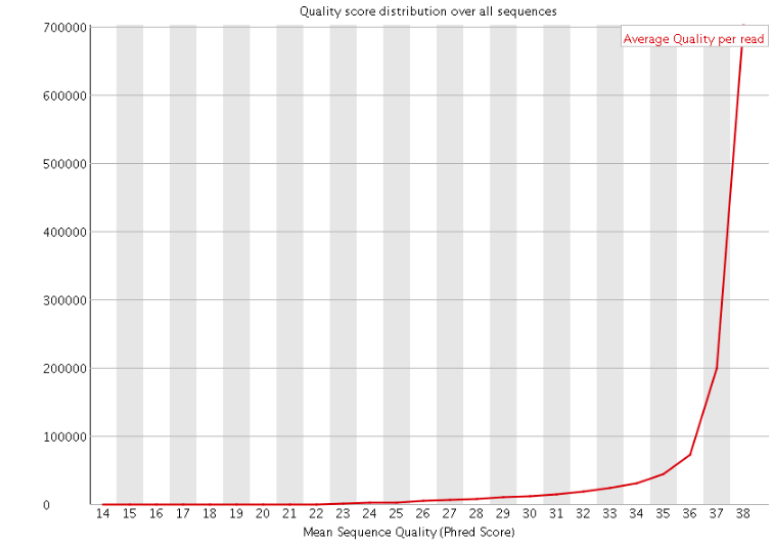


Datos trimados para primers y adaptadores R2

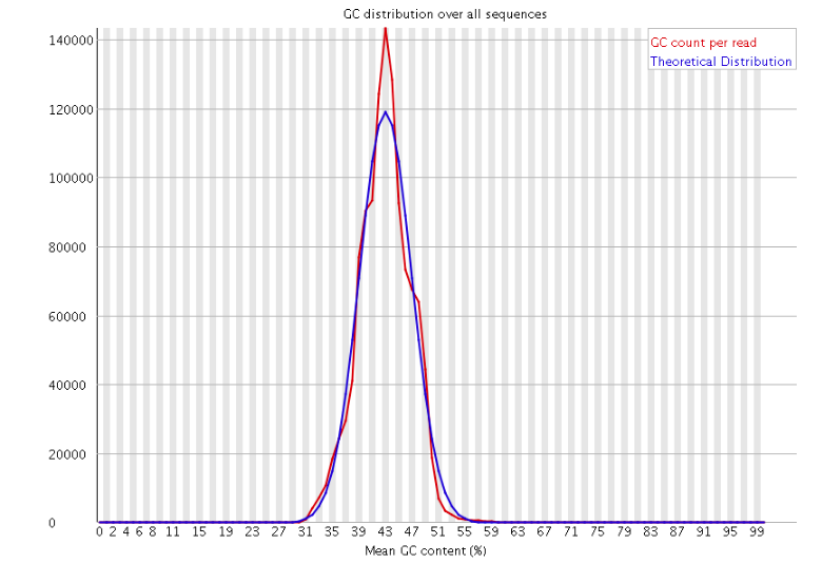
Per base sequence quality



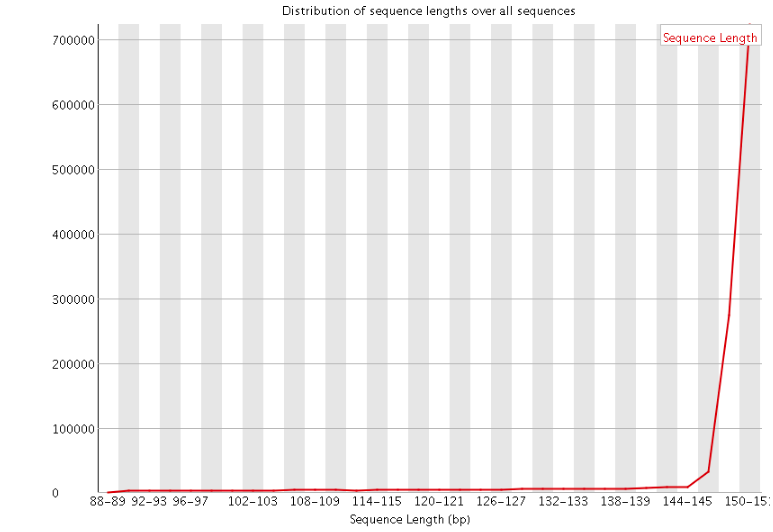
Per sequence quality scores



Per sequence GC content

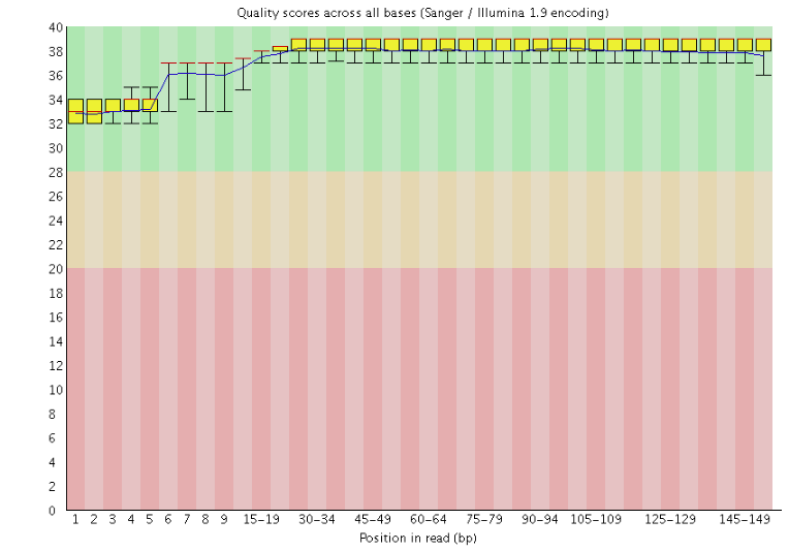


Sequence Length Distribution

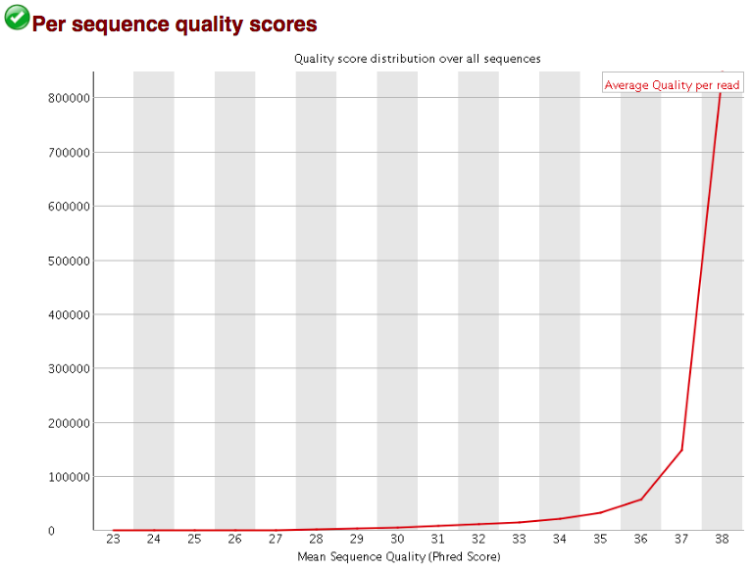


Datos trimados por calidad R1

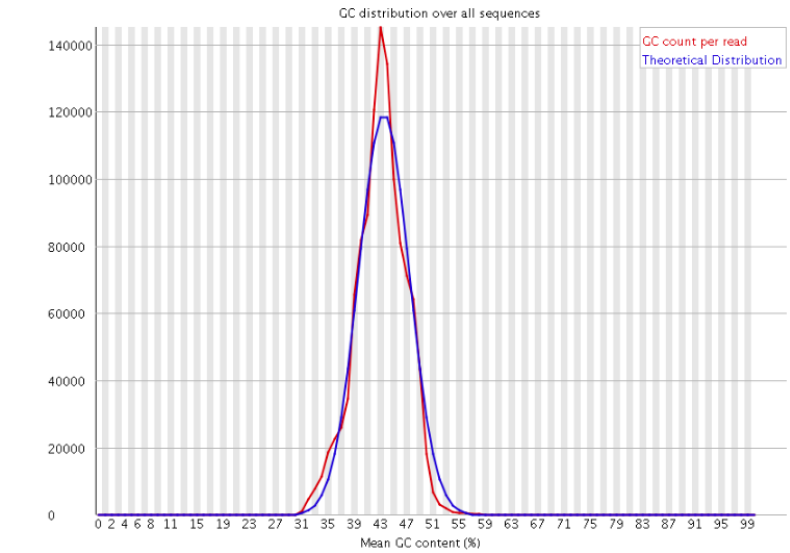
Per base sequence quality



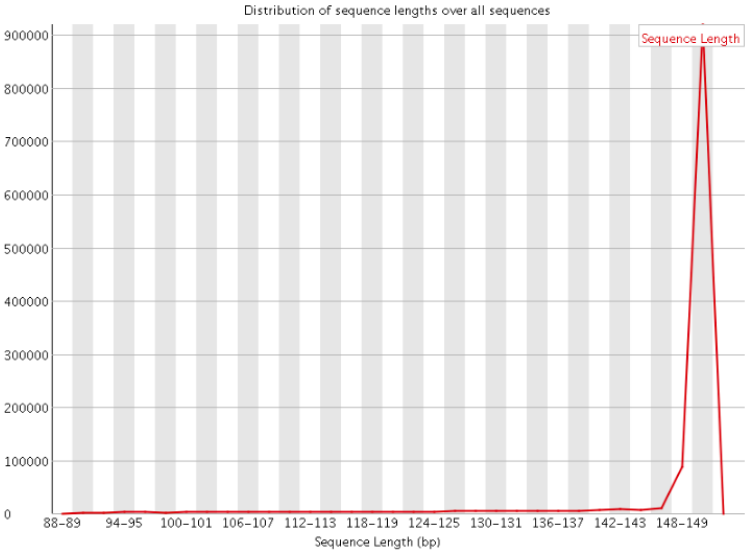
Per sequence quality scores



Per sequence GC content

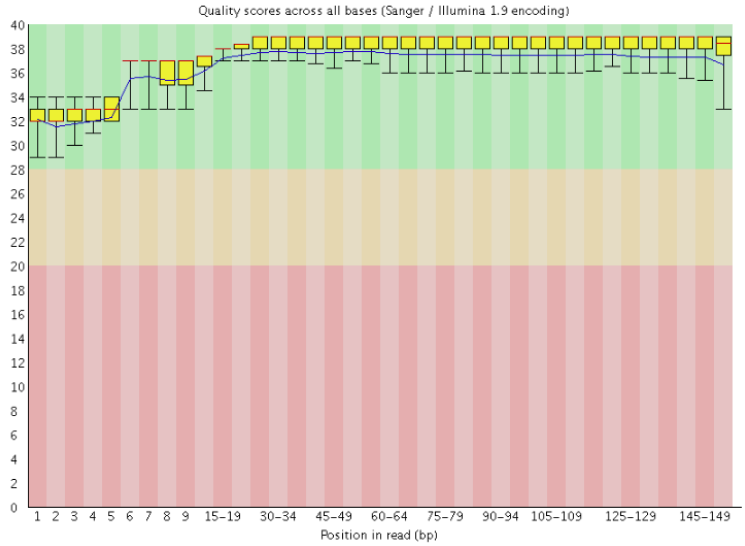


Sequence Length Distribution

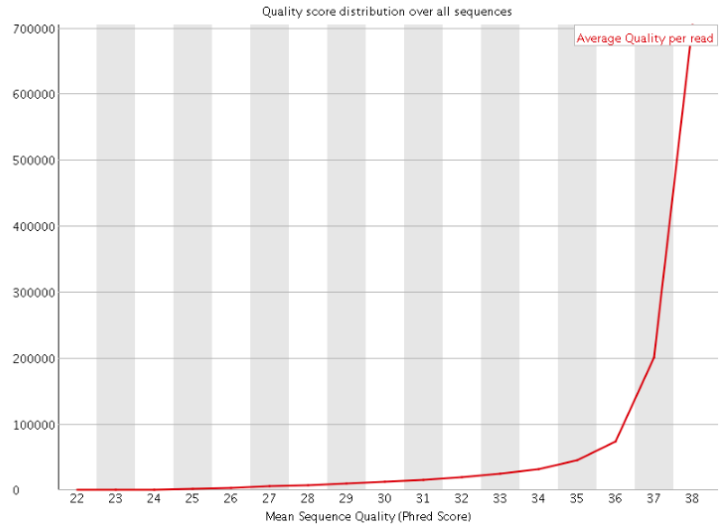


Datos trimados por calidad R2

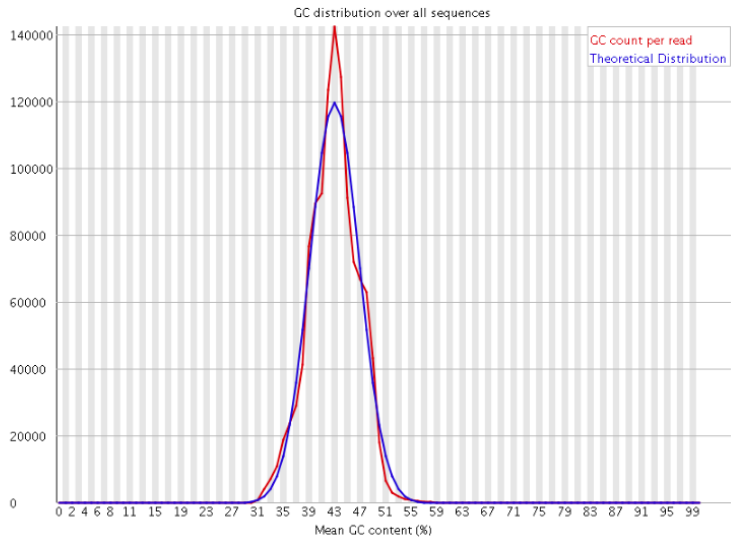
Per base sequence quality



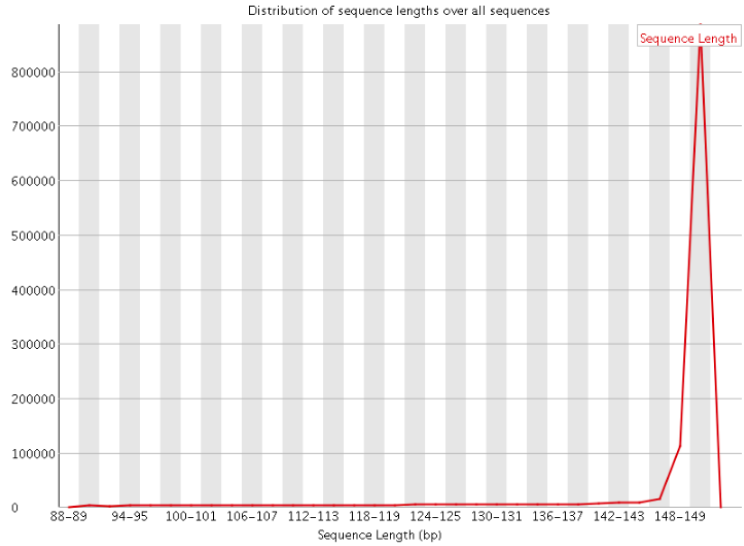
Per sequence quality scores



Per sequence GC content



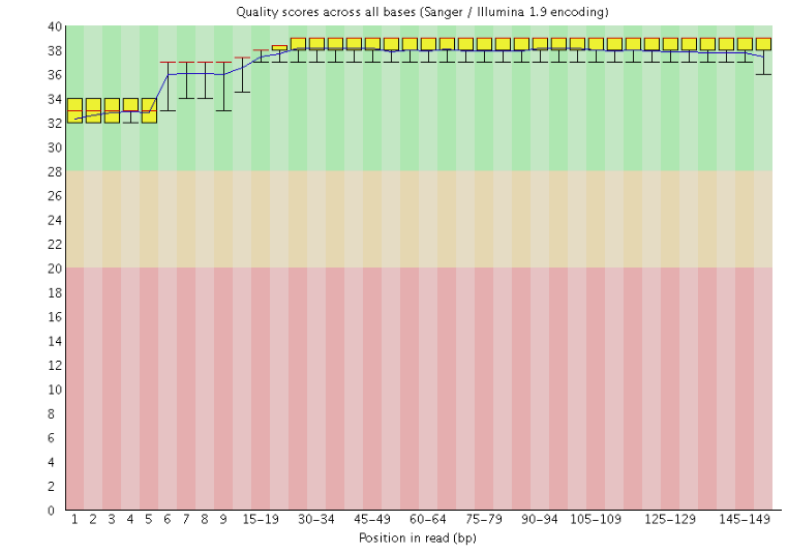
Sequence Length Distribution



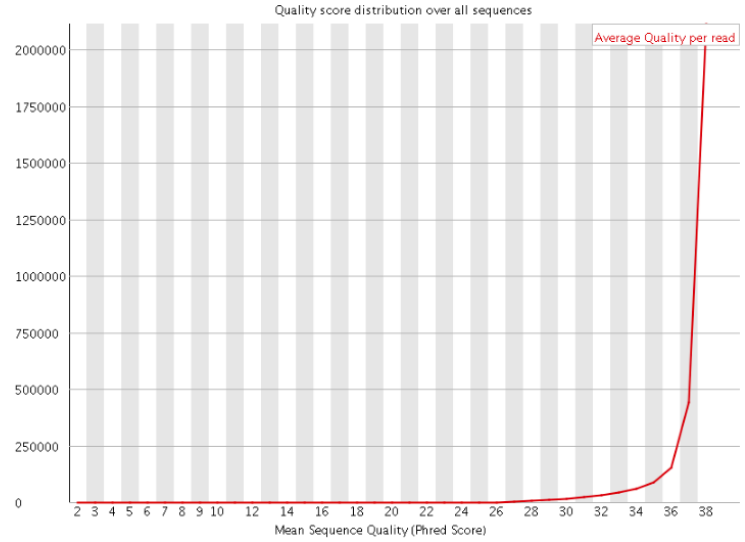
m220:

Datos Crudos R1

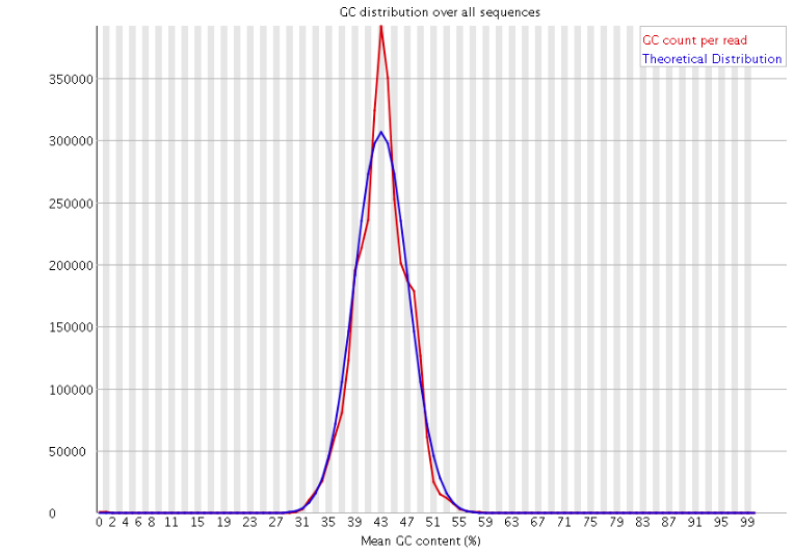
Per base sequence quality



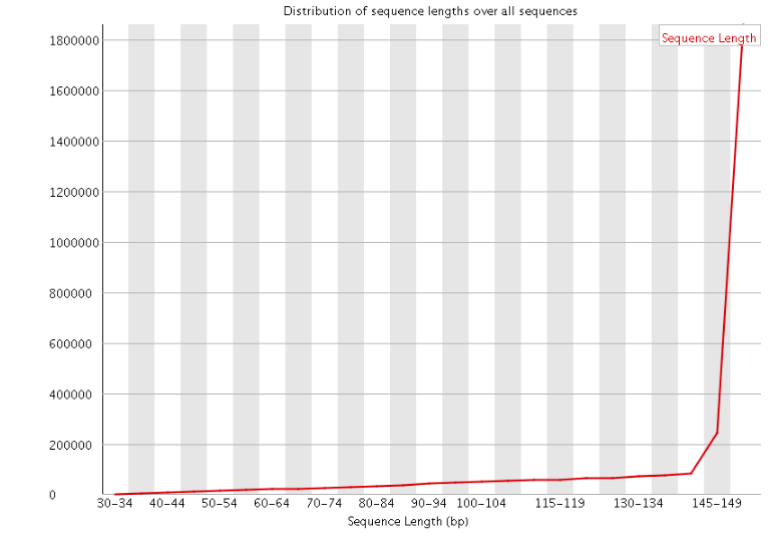
Per sequence quality scores



Per sequence GC content

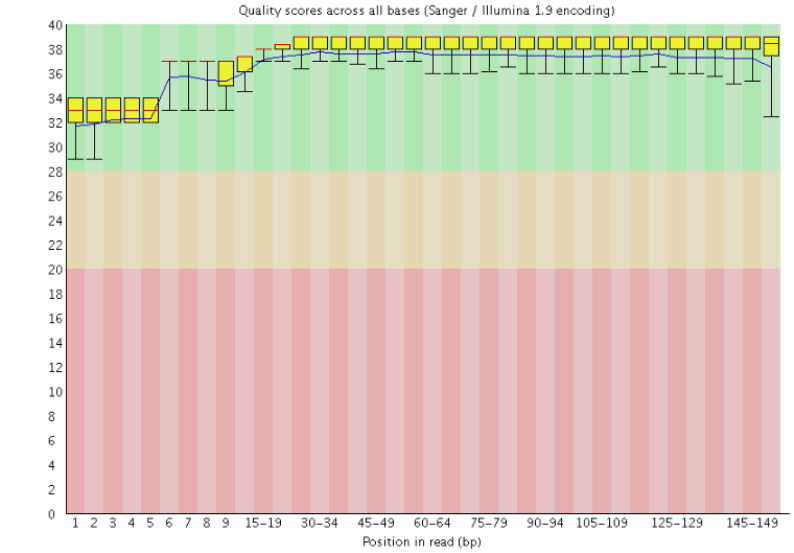


Sequence Length Distribution

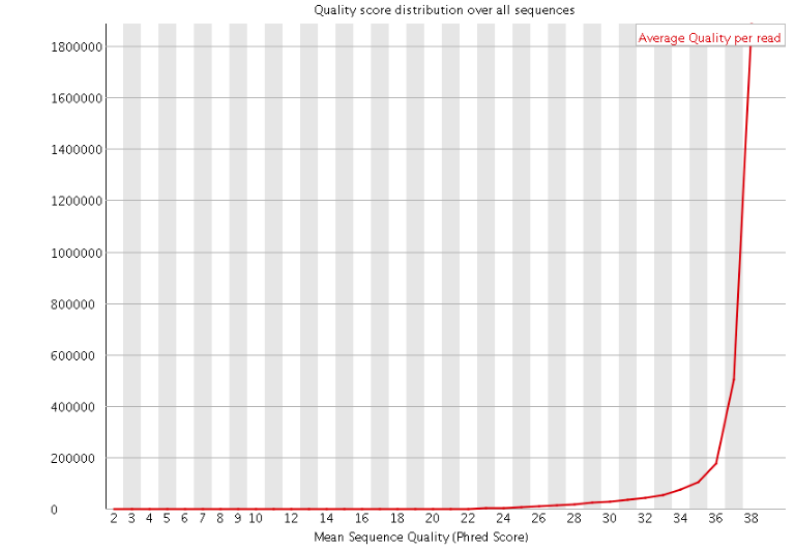


Datos Crudos R2

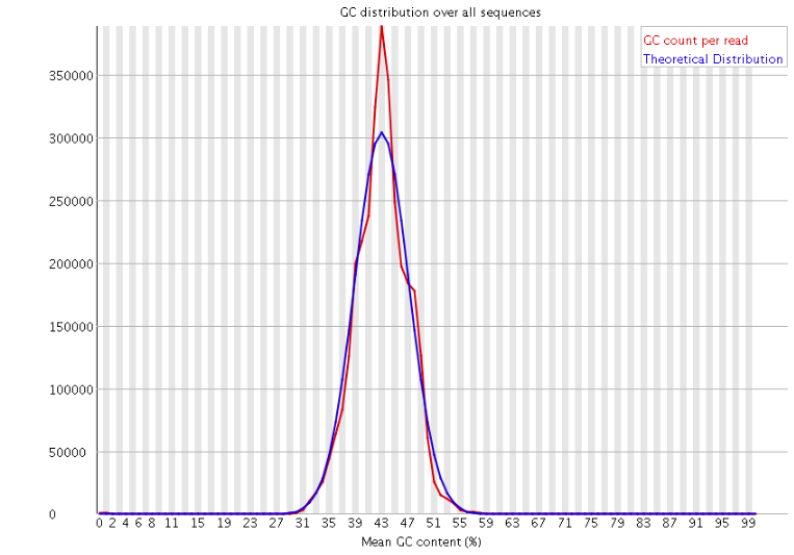
Per base sequence quality



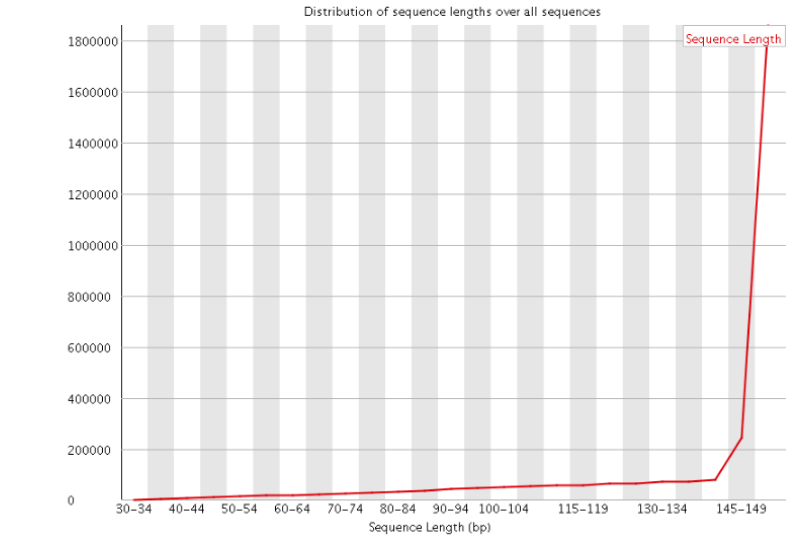
Per sequence quality scores



Per sequence GC content

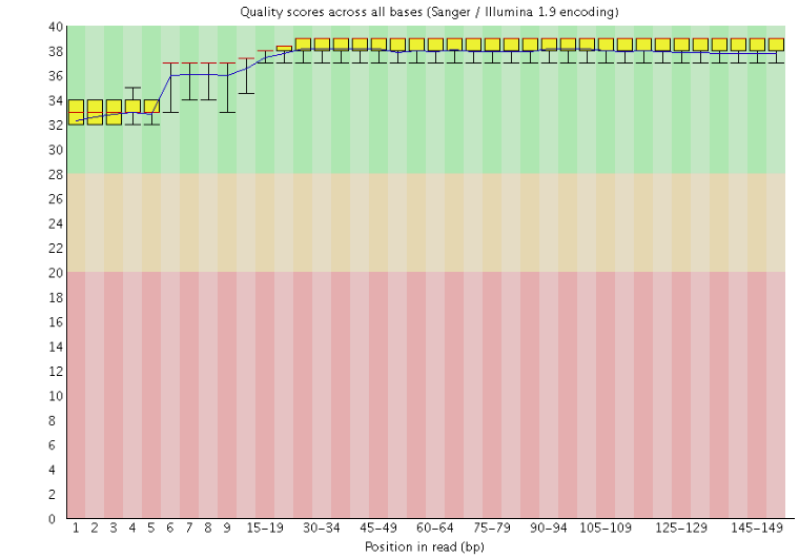


Sequence Length Distribution

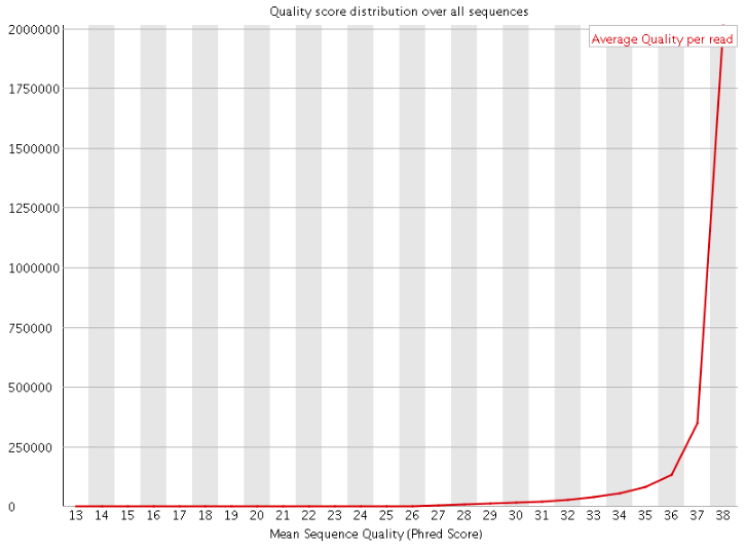


Datos trimados para primers y adaptadores R1

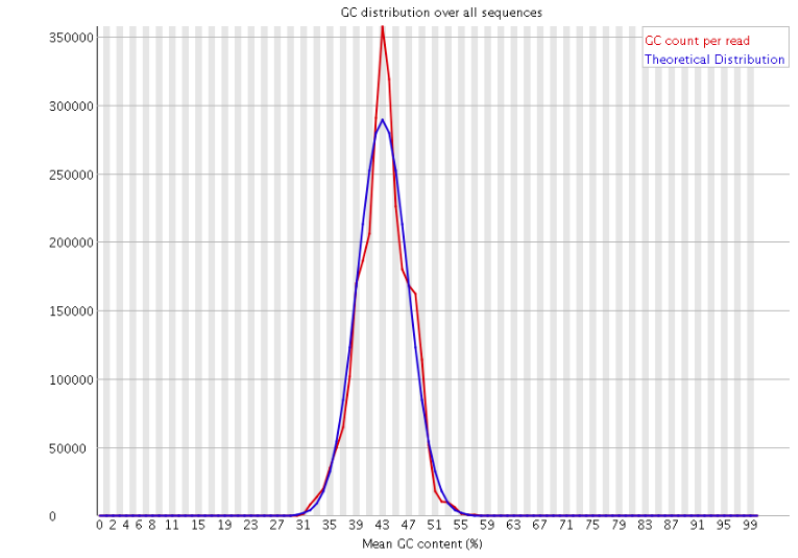
Per base sequence quality



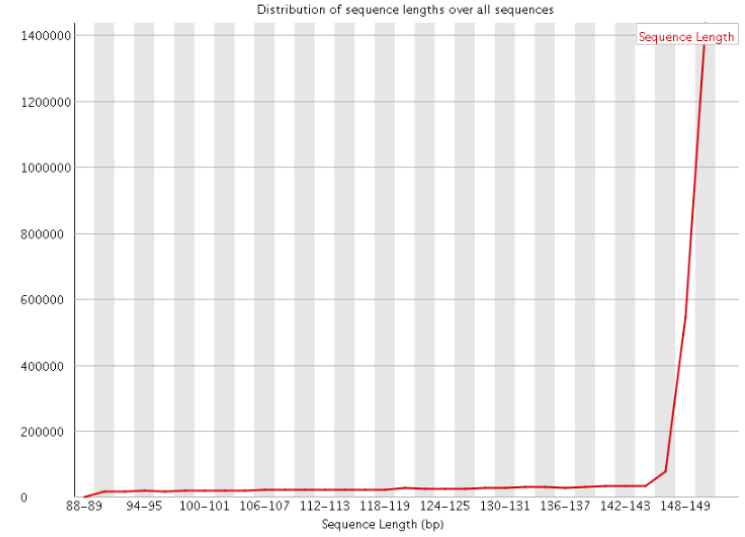
Per sequence quality scores



Per sequence GC content

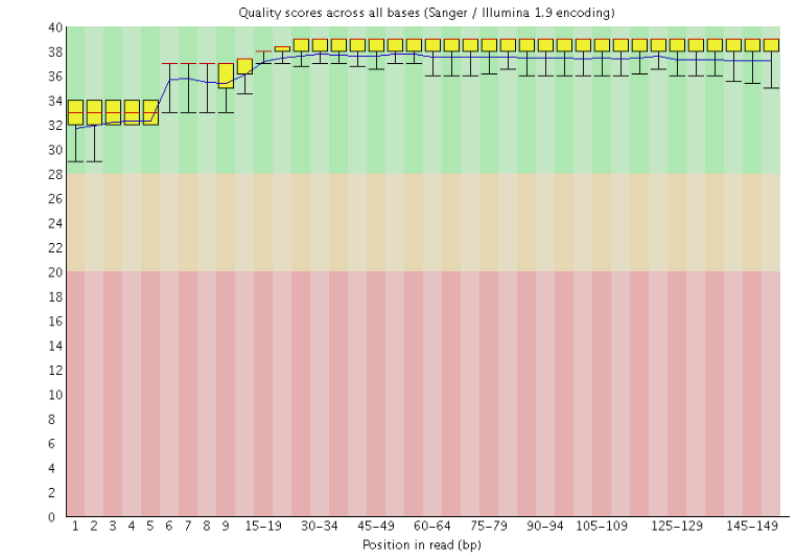


Sequence Length Distribution

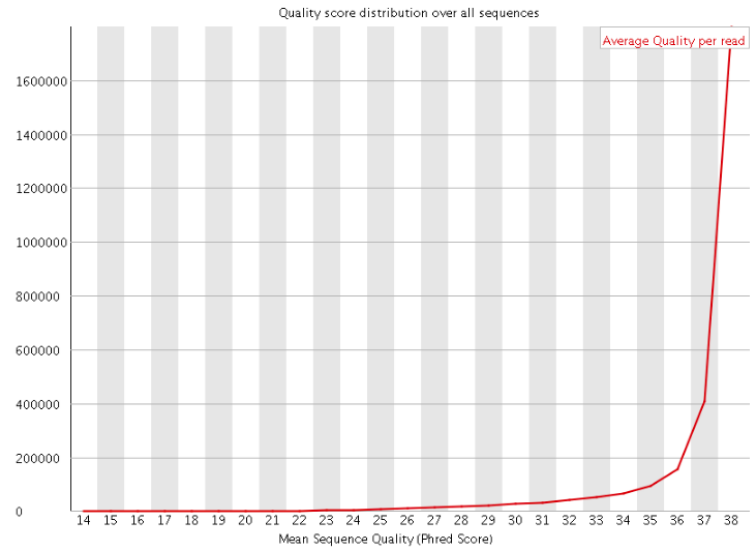


Datos trimados para primers y adaptadores R2

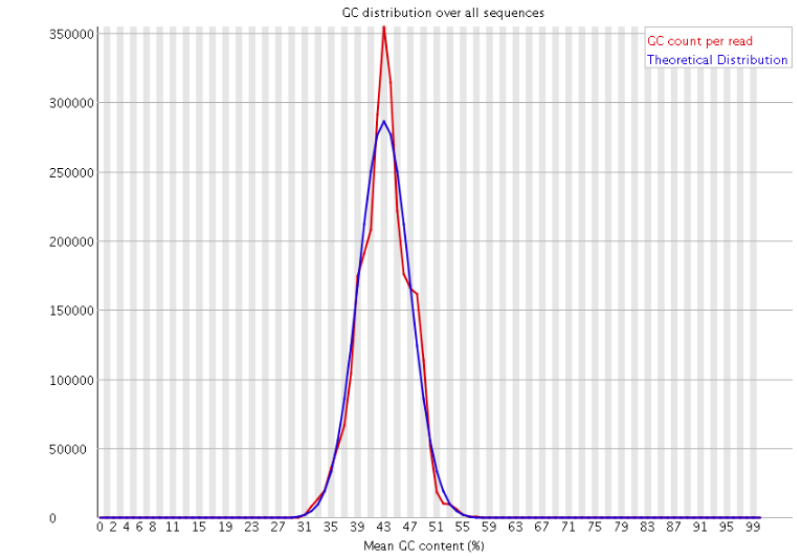
Per base sequence quality



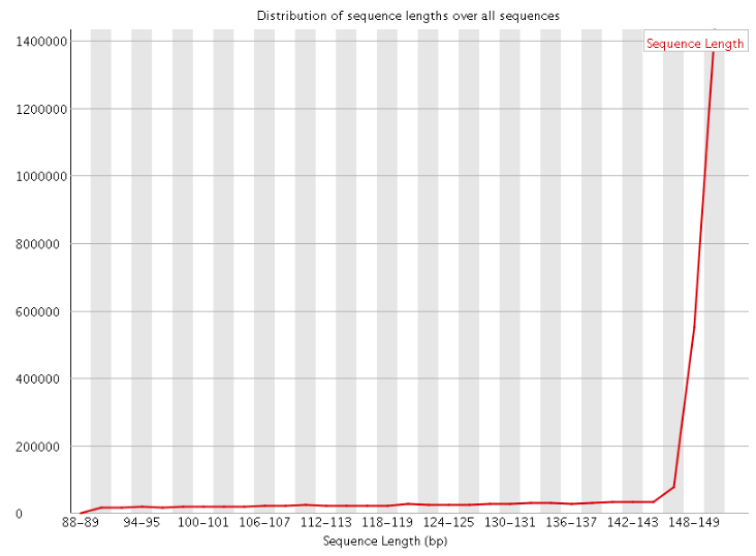
Per sequence quality scores



Per sequence GC content

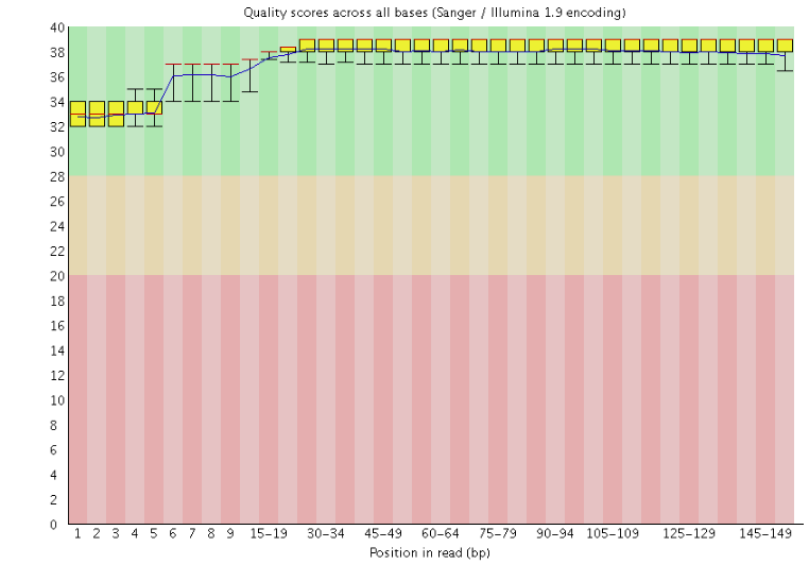


Sequence Length Distribution

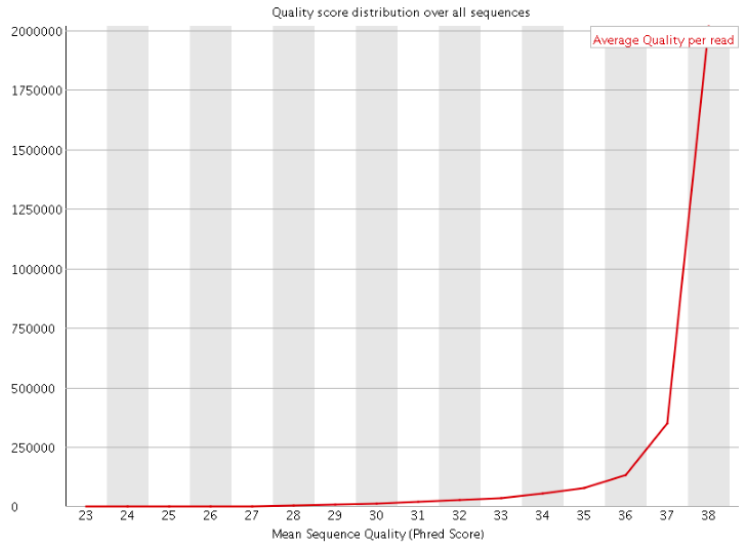


Datos trimados por calidad R1

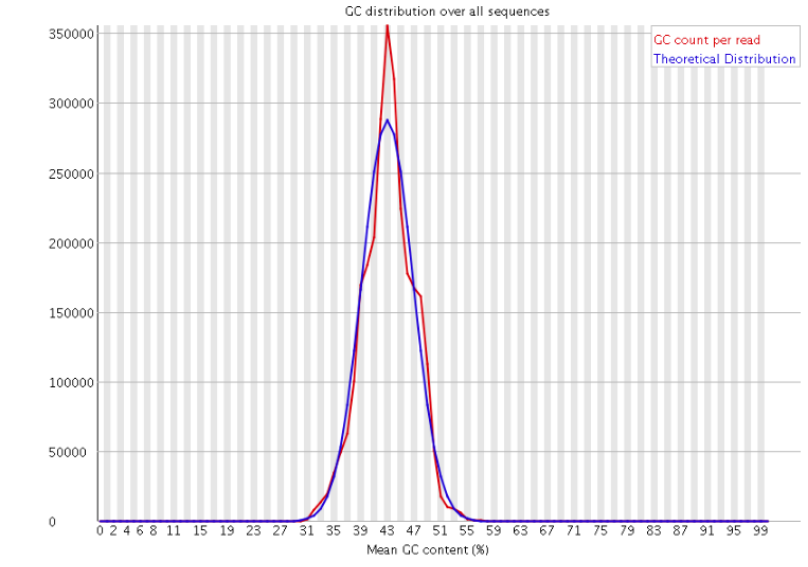
Per base sequence quality



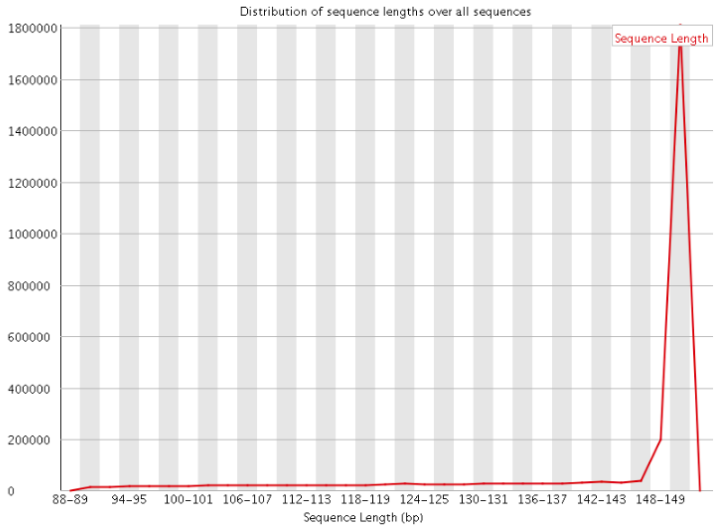
Per sequence quality scores



Per sequence GC content

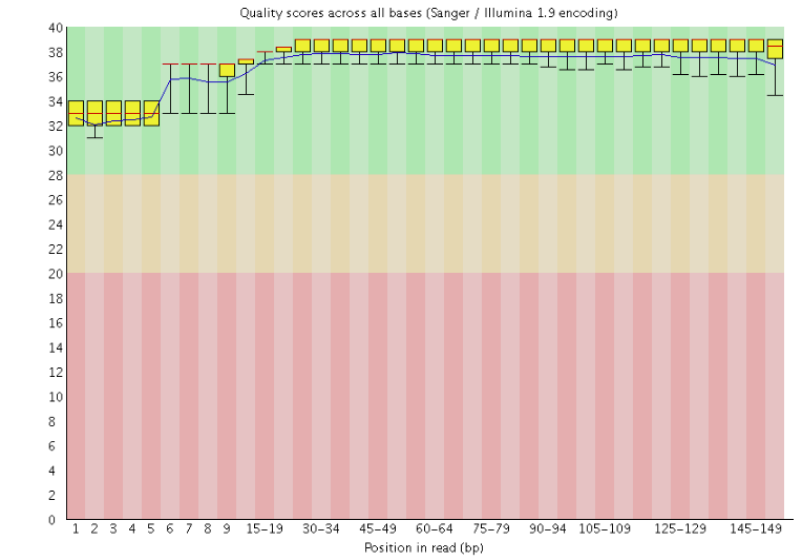


Sequence Length Distribution

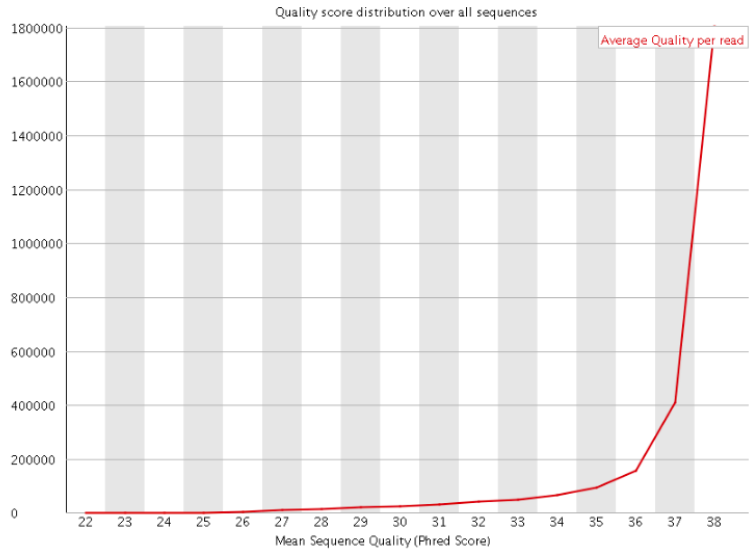


Datos trimados por calidad R2

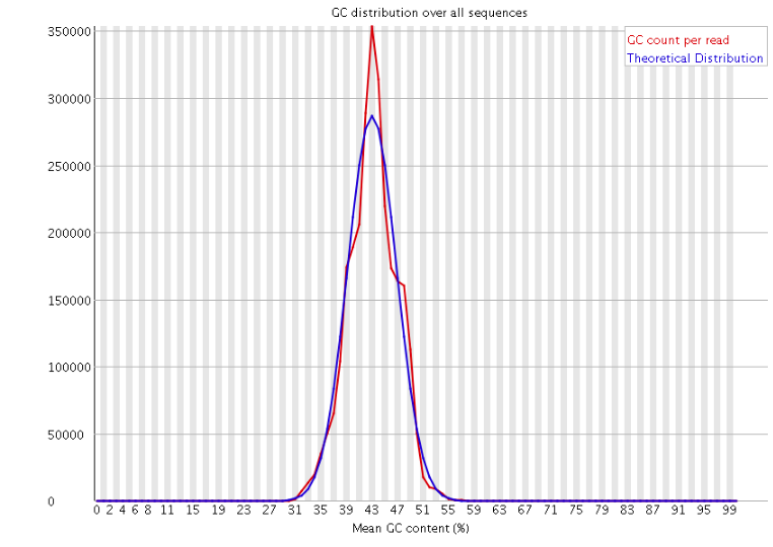
Per base sequence quality



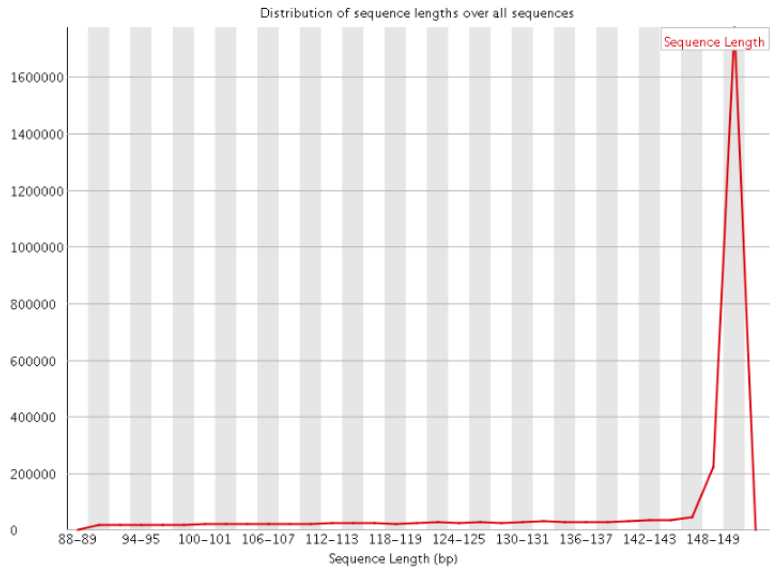
Per sequence quality scores



Per sequence GC content



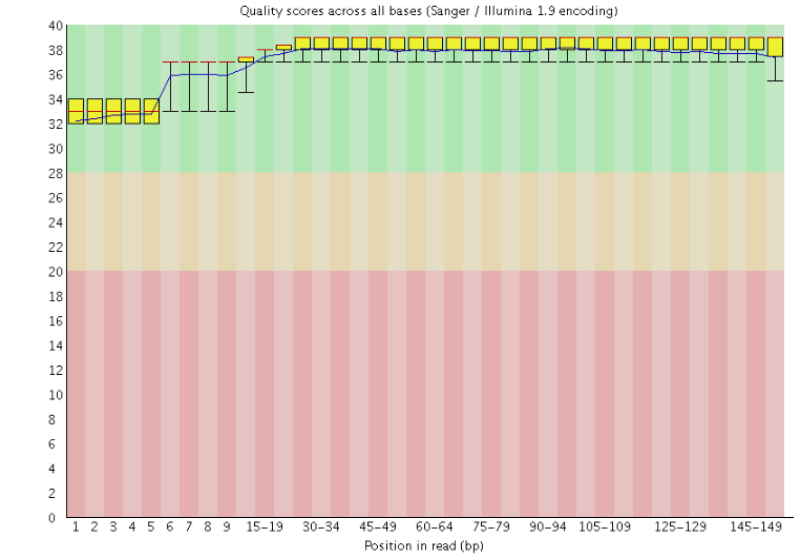
Sequence Length Distribution



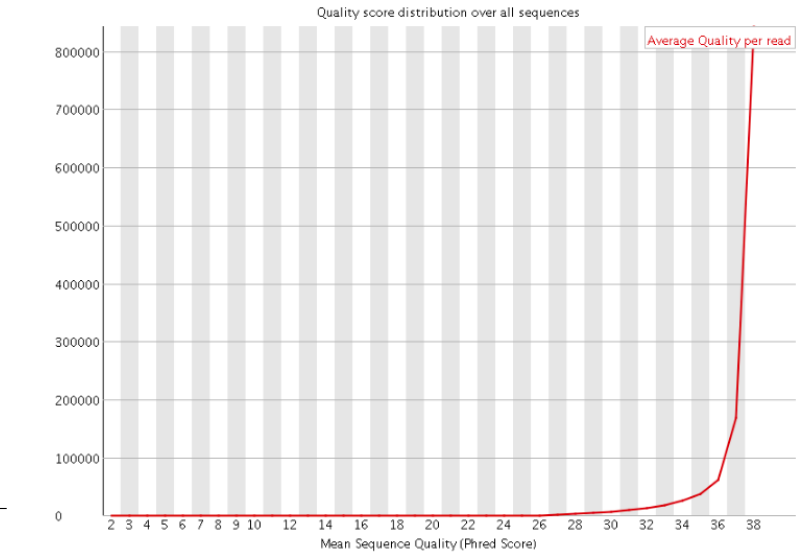
m224:

Datos Crudos R1

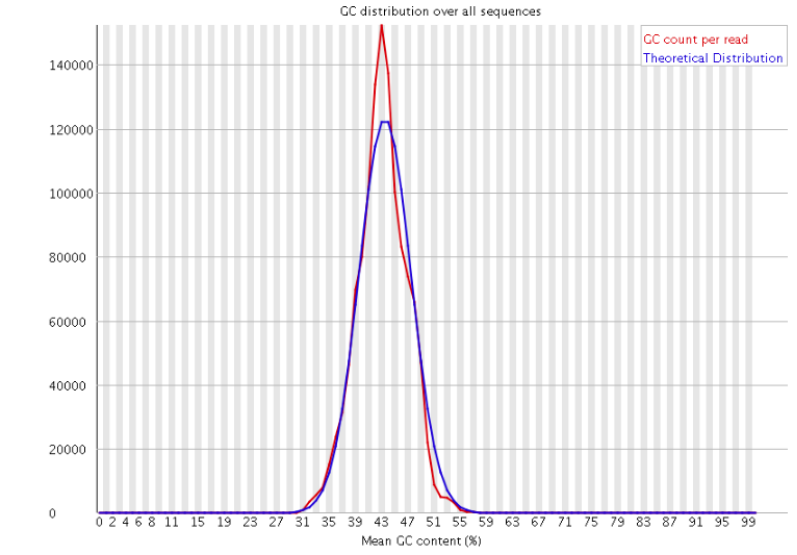
Per base sequence quality



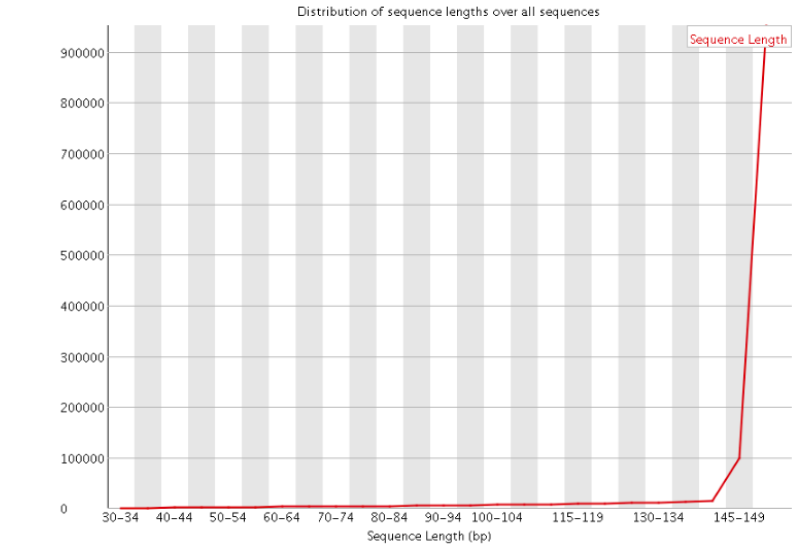
Per sequence quality scores



Per sequence GC content

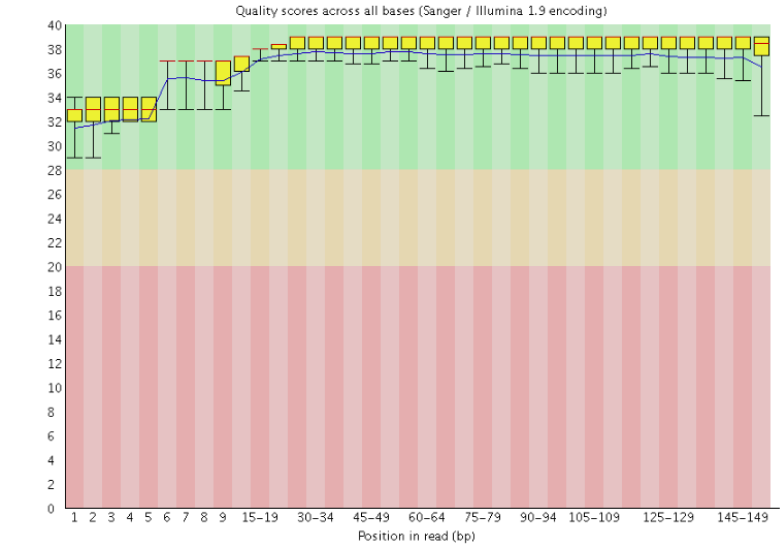


Sequence Length Distribution

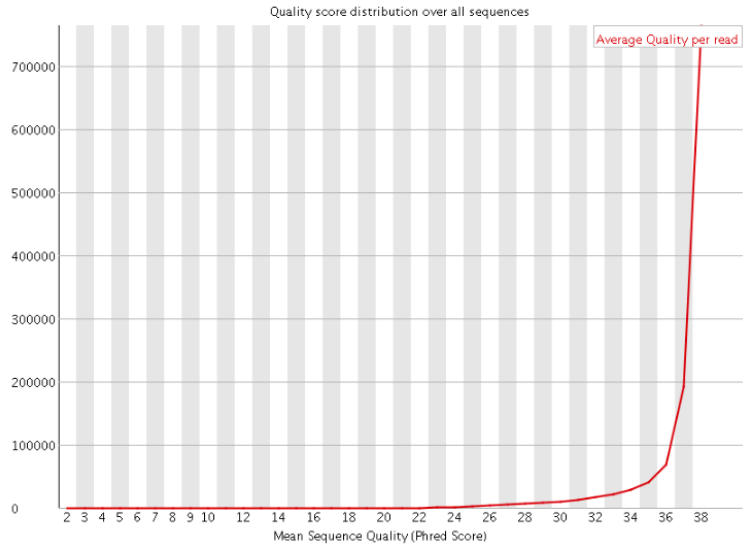


Datos Crudos R2

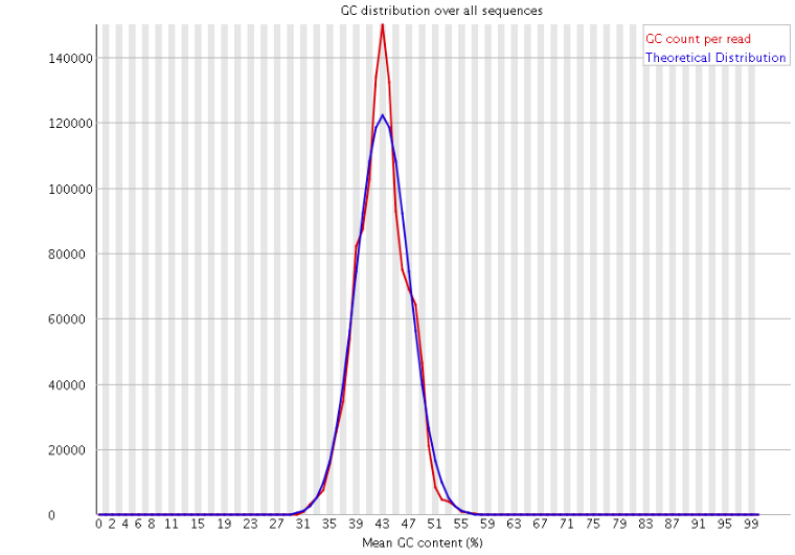
Per base sequence quality



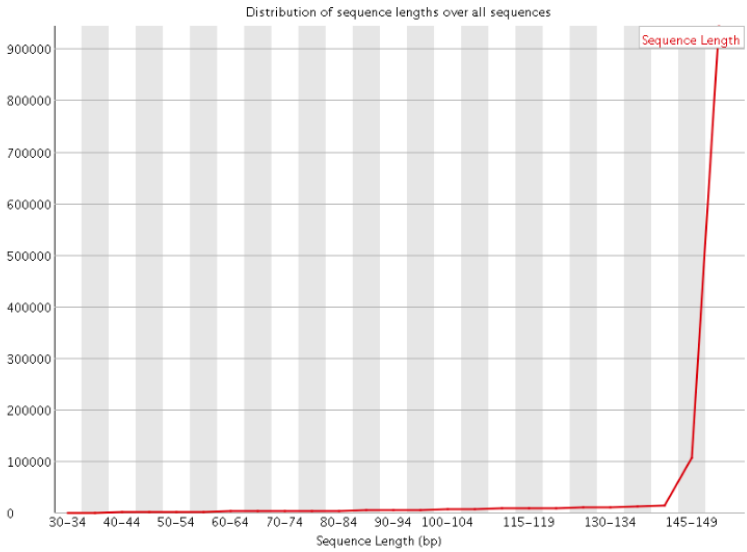
Per sequence quality scores



Per sequence GC content

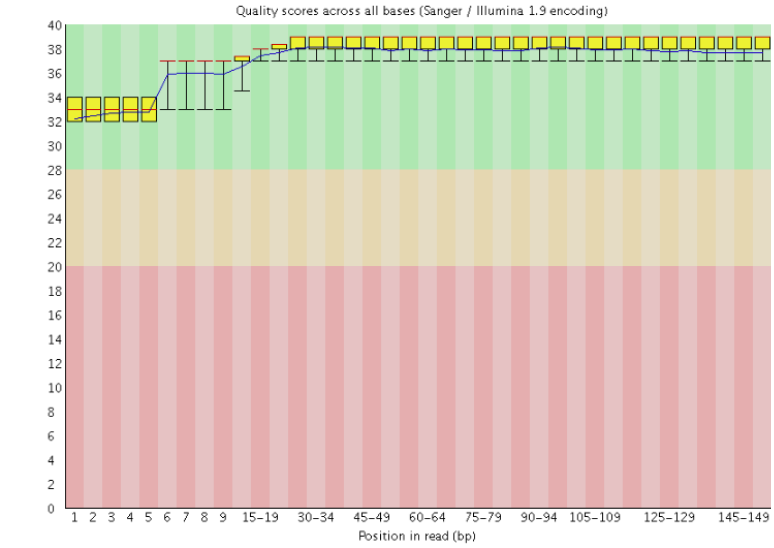


Sequence Length Distribution

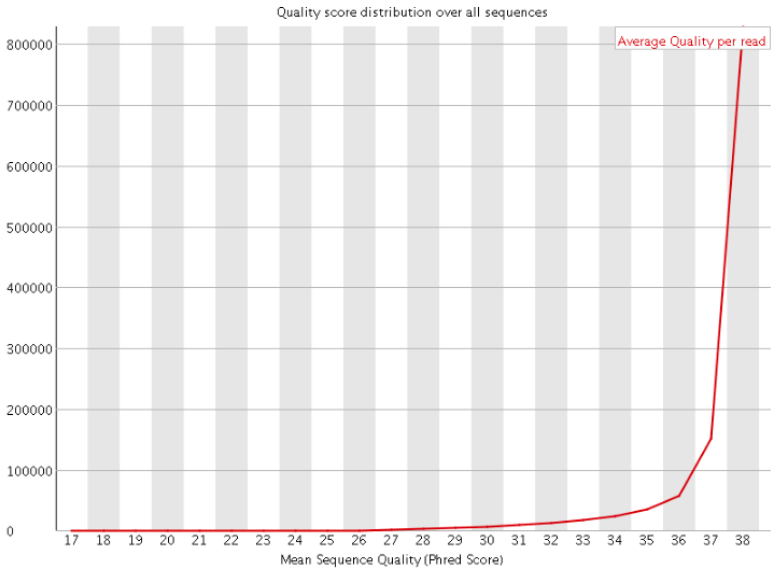


Datos trimados para primers y adaptadores R1

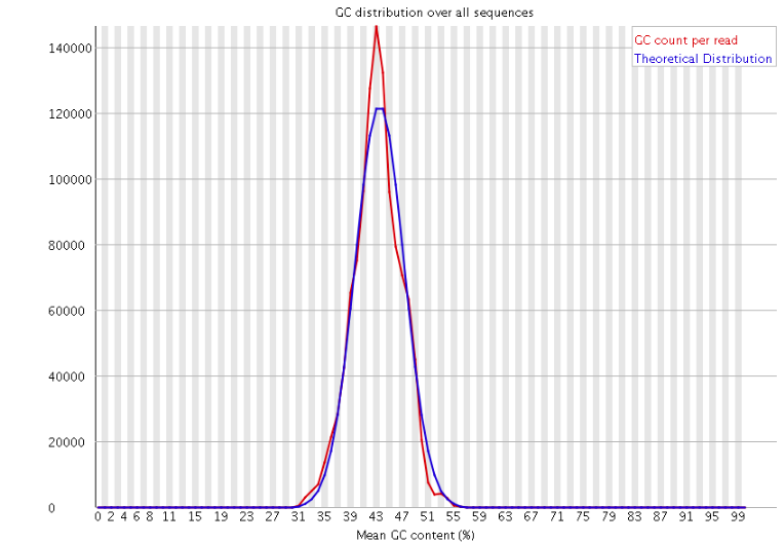
Per base sequence quality



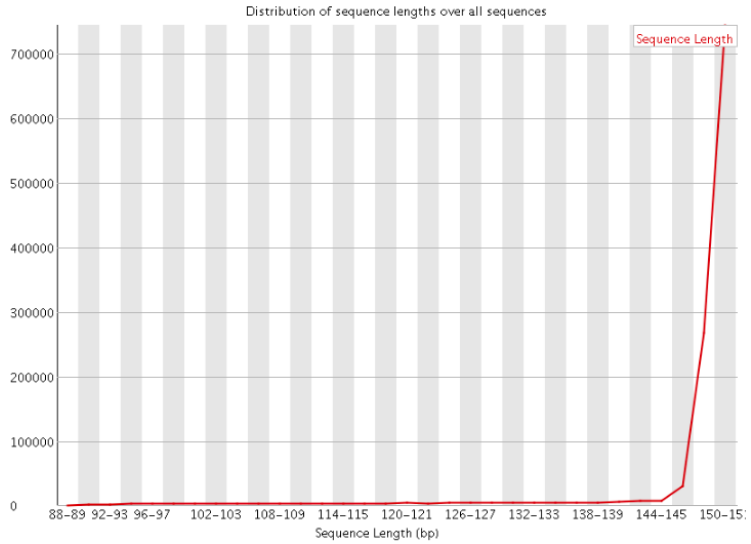
Per sequence quality scores



Per sequence GC content

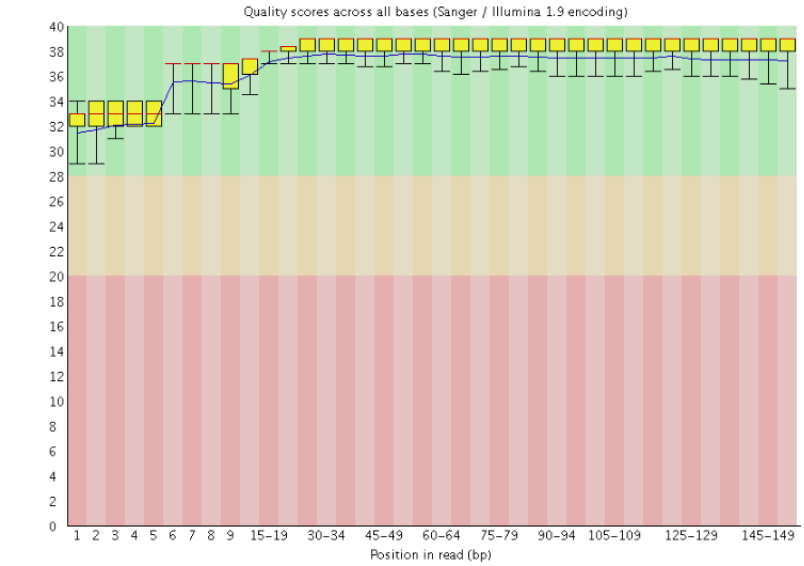


Sequence Length Distribution

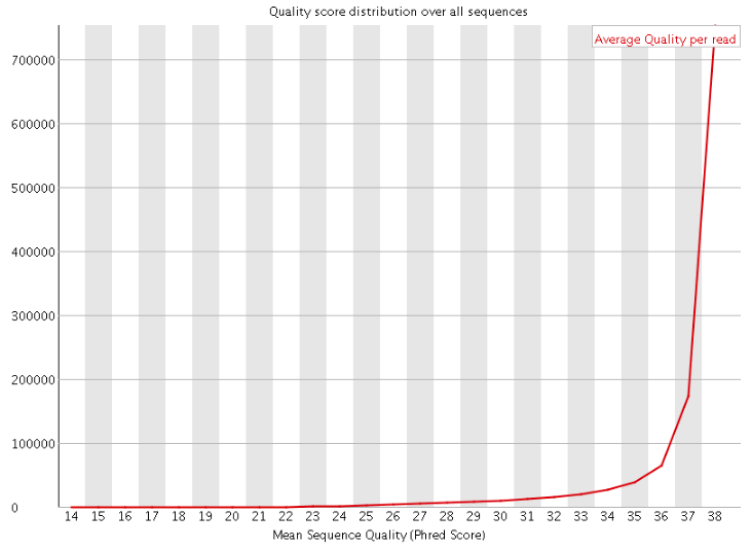


Datos trimados para primers y adaptadores R2

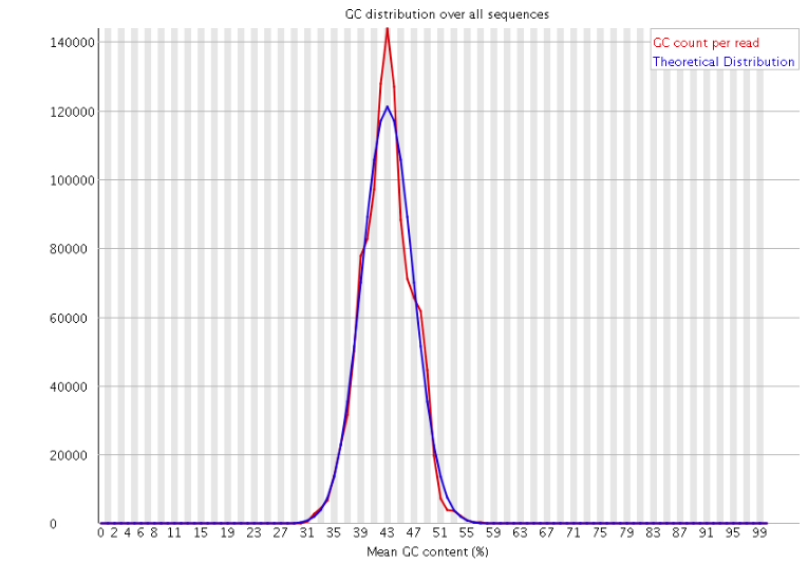
Per base sequence quality



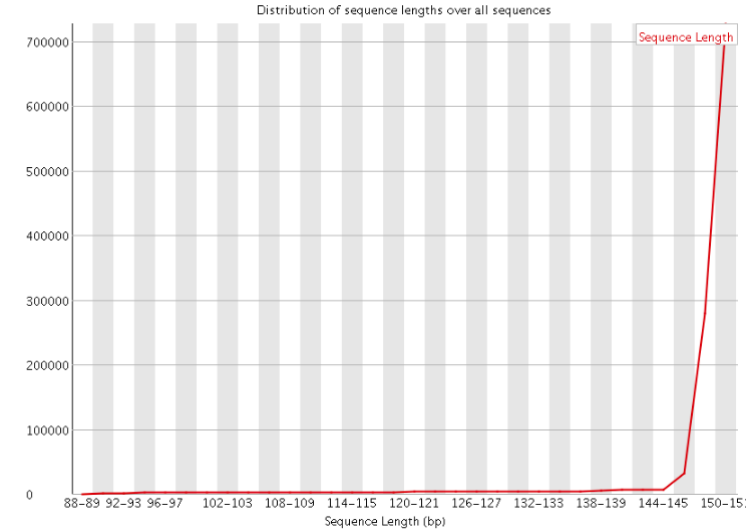
Per sequence quality scores



Per sequence GC content

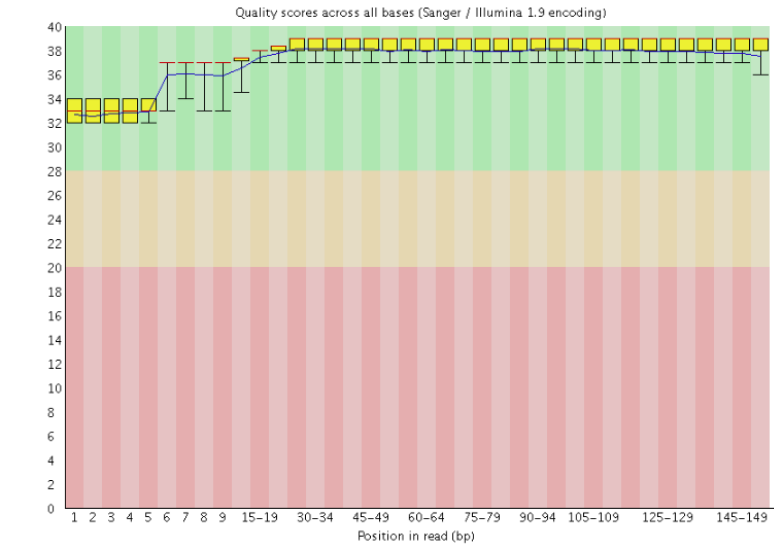


Sequence Length Distribution

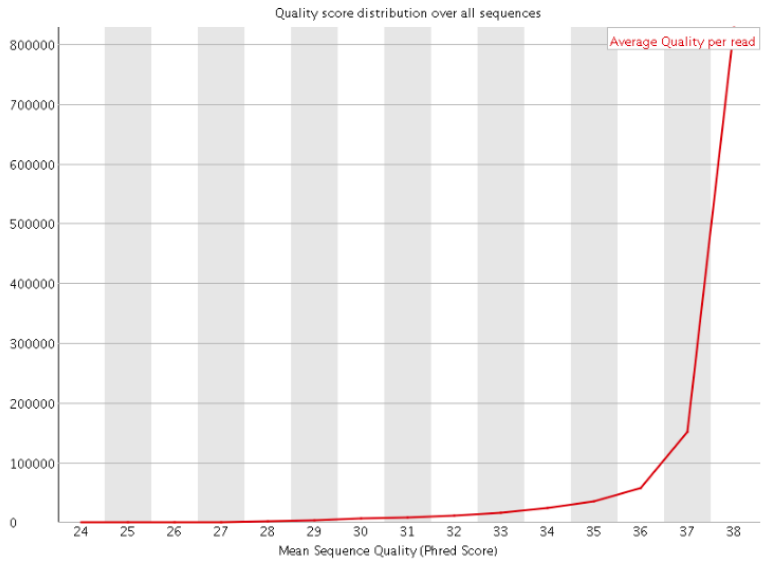


Datos trimados por calidad R1

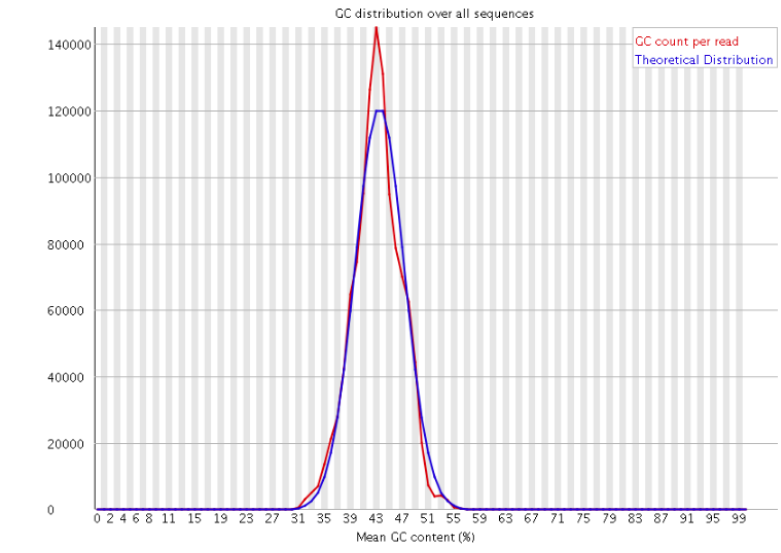
Per base sequence quality



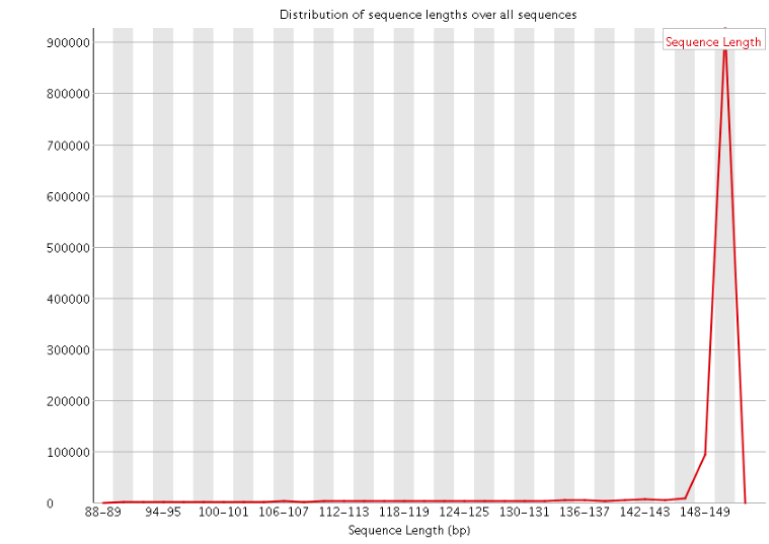
Per sequence quality scores



Per sequence GC content

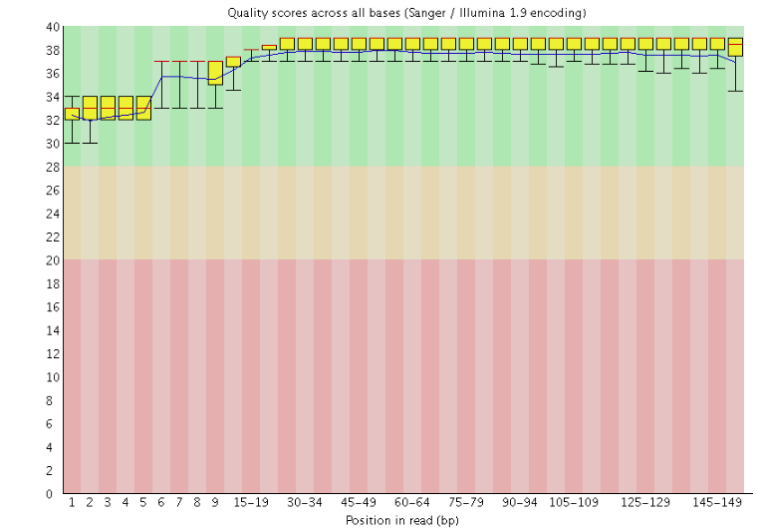


Sequence Length Distribution

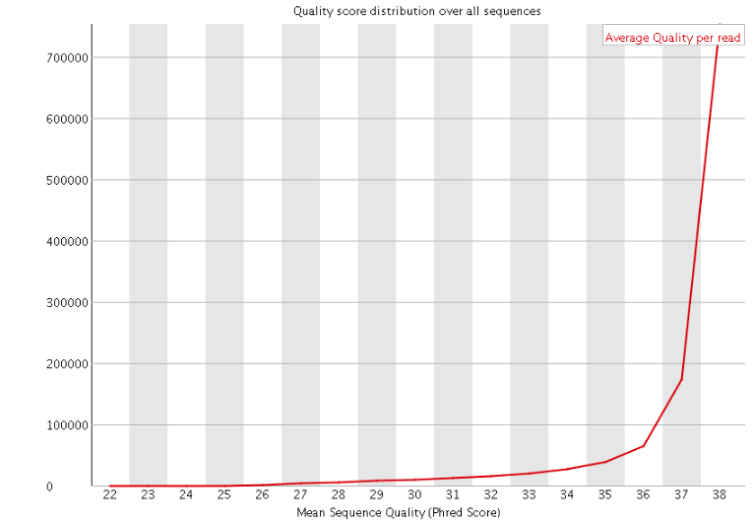


Datos trimados por calidad R2

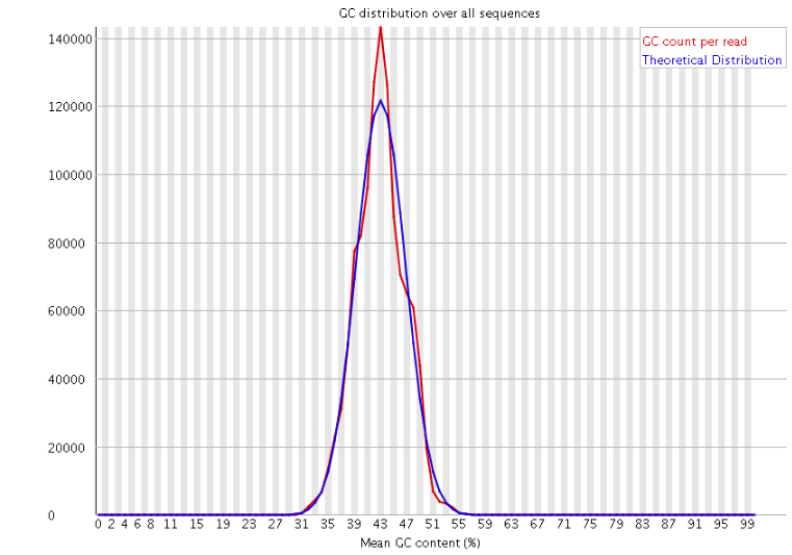
Per base sequence quality



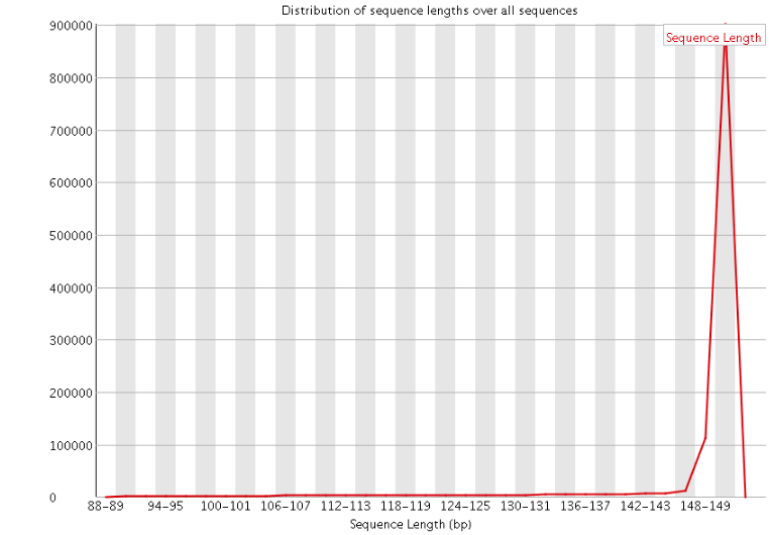
Per sequence quality scores



Per sequence GC content



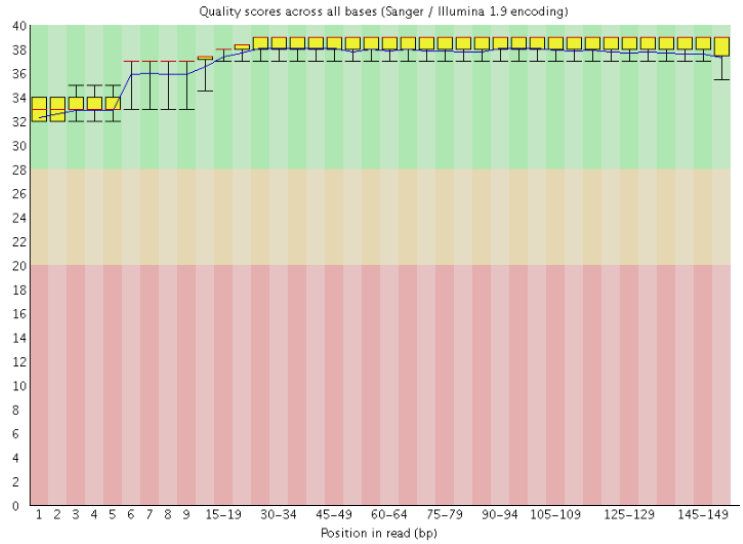
Sequence Length Distribution



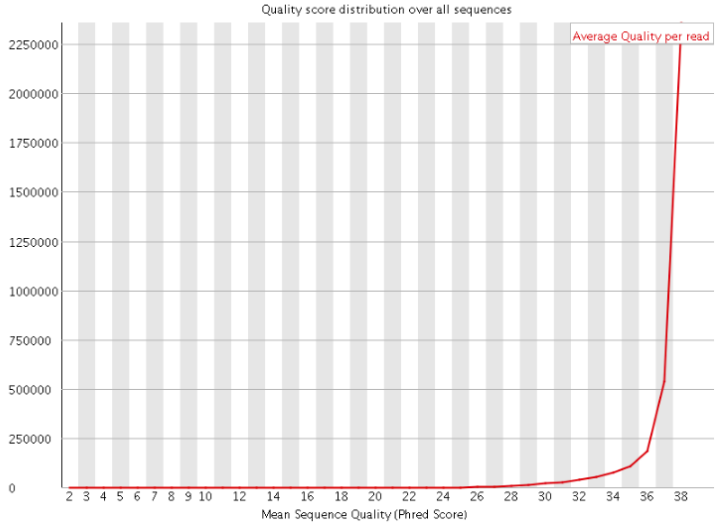
m250:

Datos crudos R1

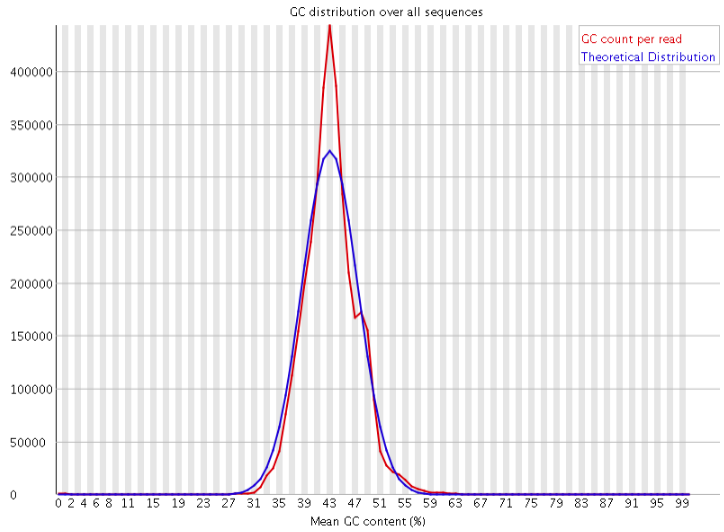
Per base sequence quality



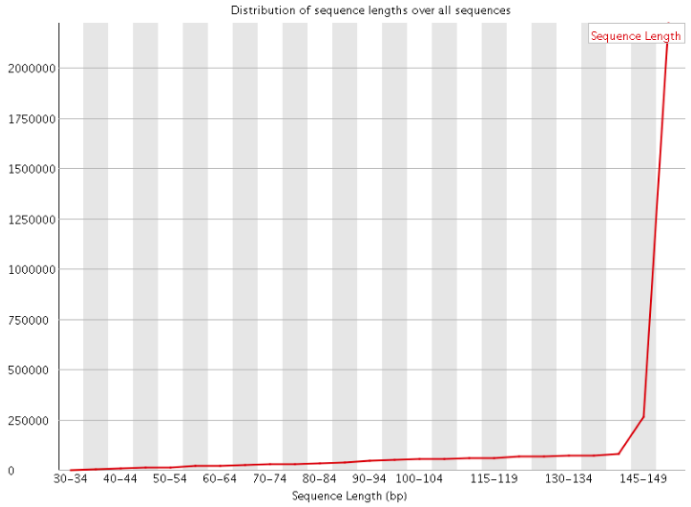
Per sequence quality scores



Per sequence GC content

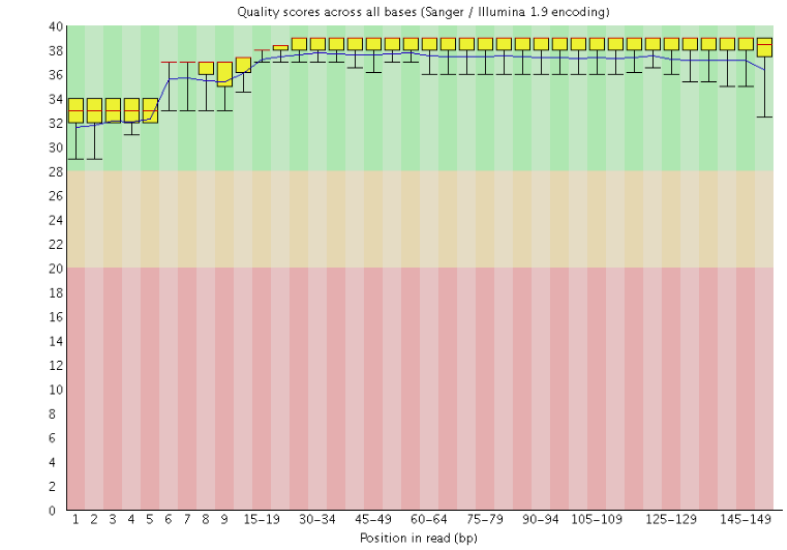


Sequence Length Distribution

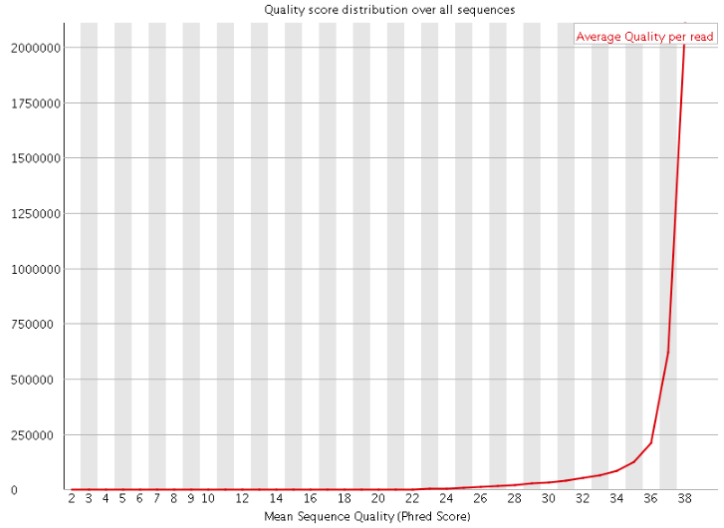


Datos crudos R2

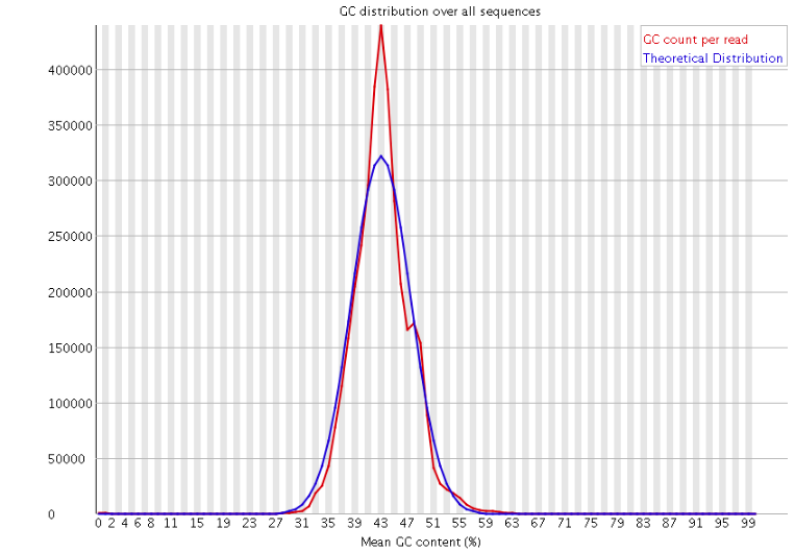
Per base sequence quality



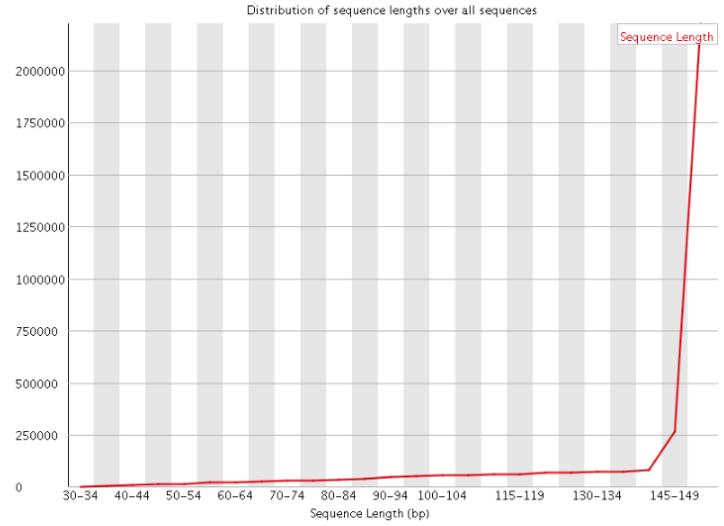
Per sequence quality scores



Per sequence GC content

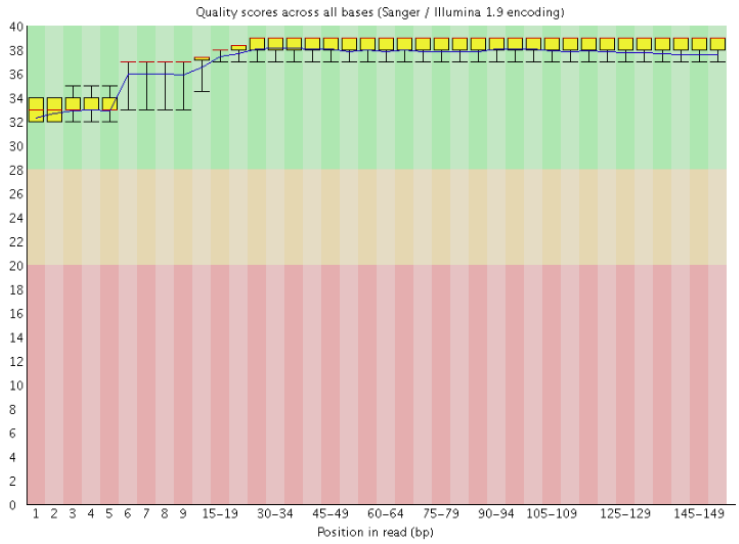


Sequence Length Distribution

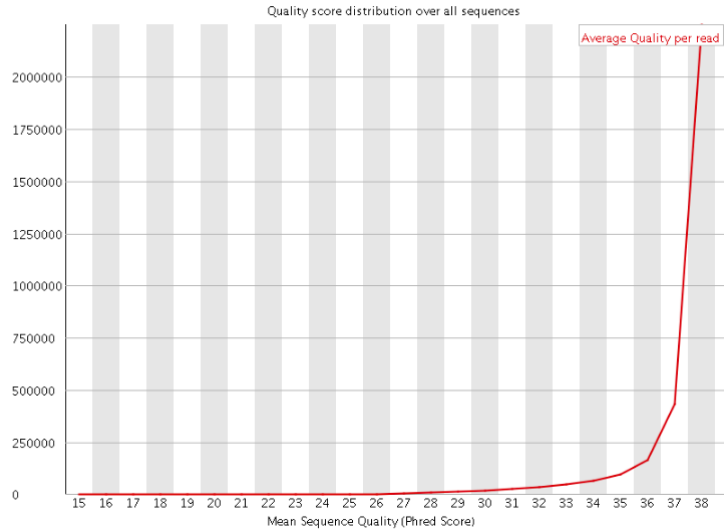


Datos trimados para primers y adaptadores R1

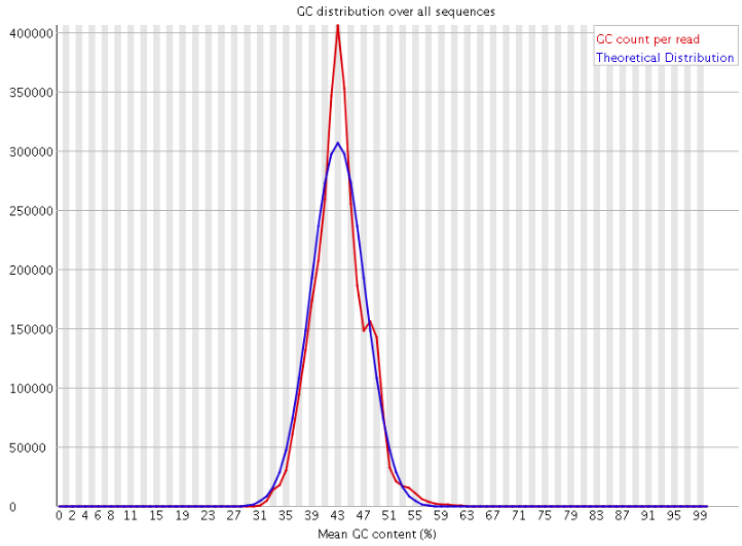
Per base sequence quality



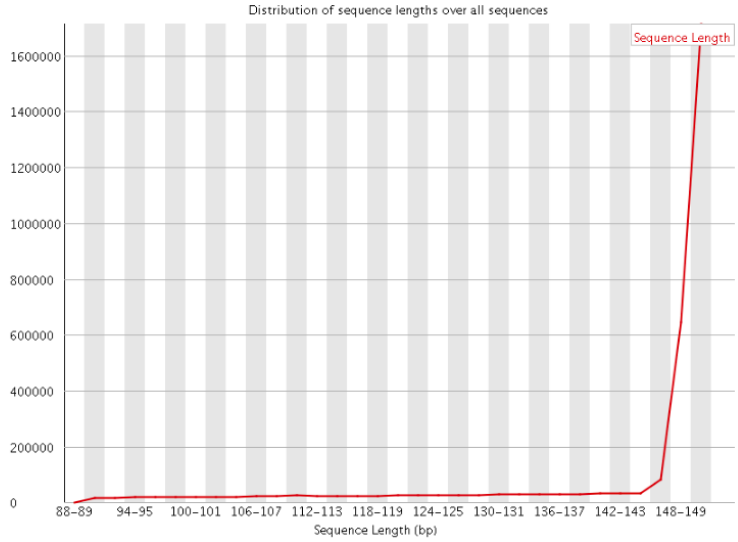
Per sequence quality scores



Per sequence GC content

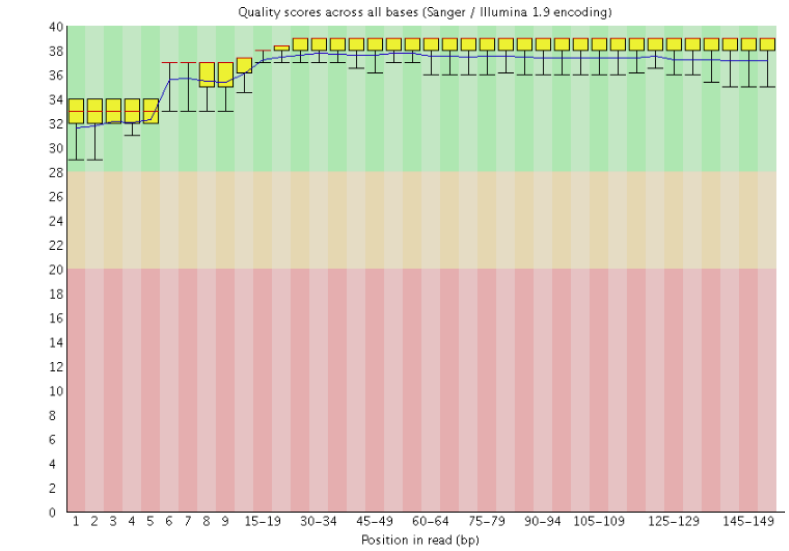


Sequence Length Distribution

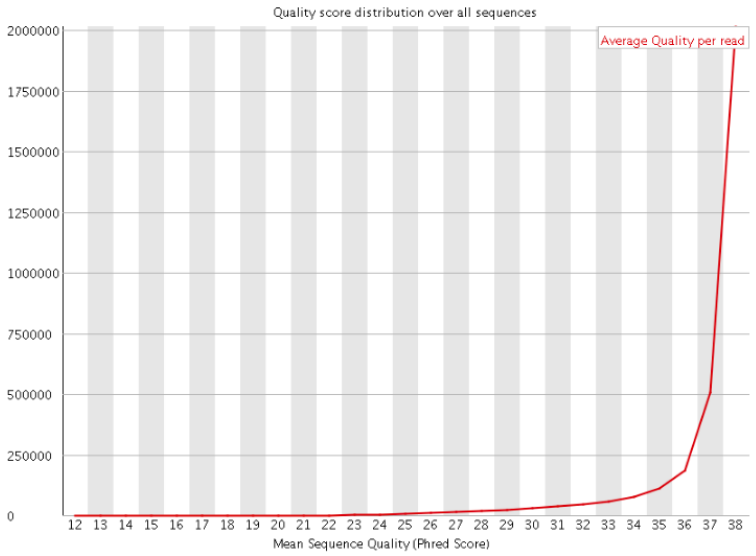


Datos trimados para primers y adaptadores R2

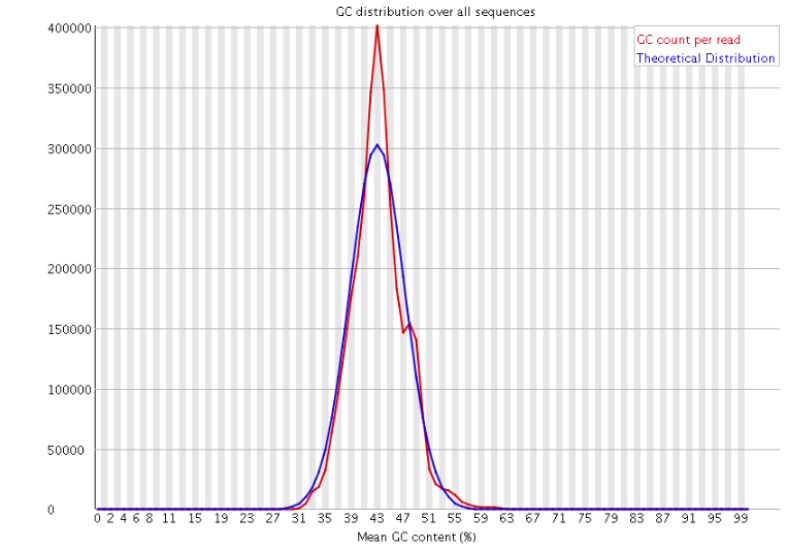
Per base sequence quality



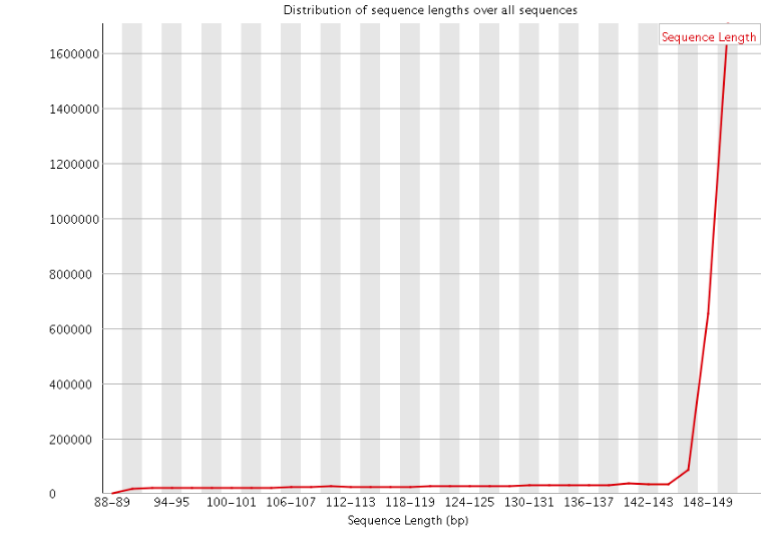
Per sequence quality scores



Per sequence GC content

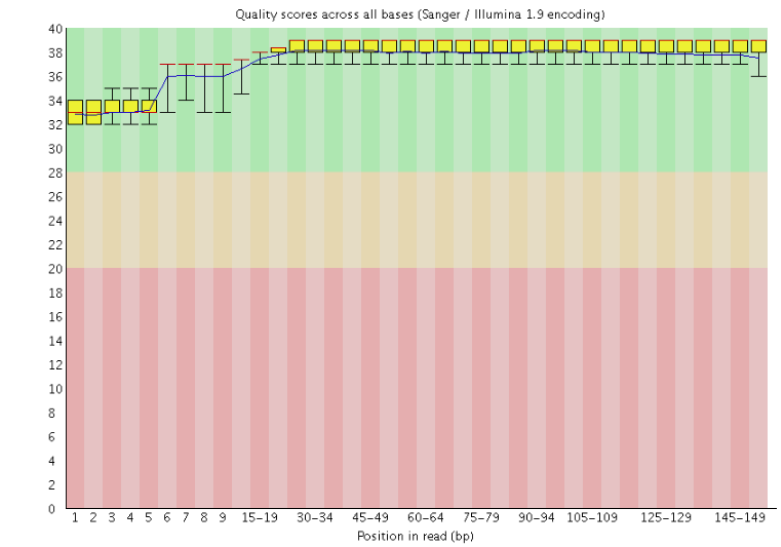


Sequence Length Distribution

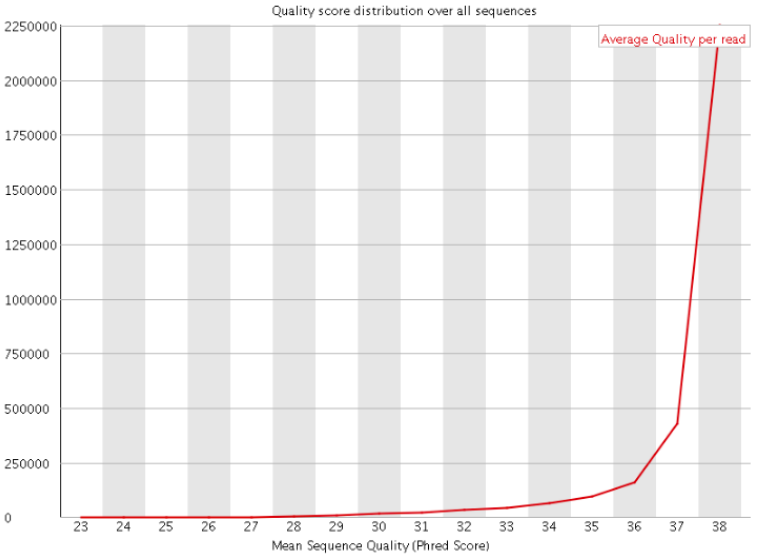


Datos trimados por calidad R1

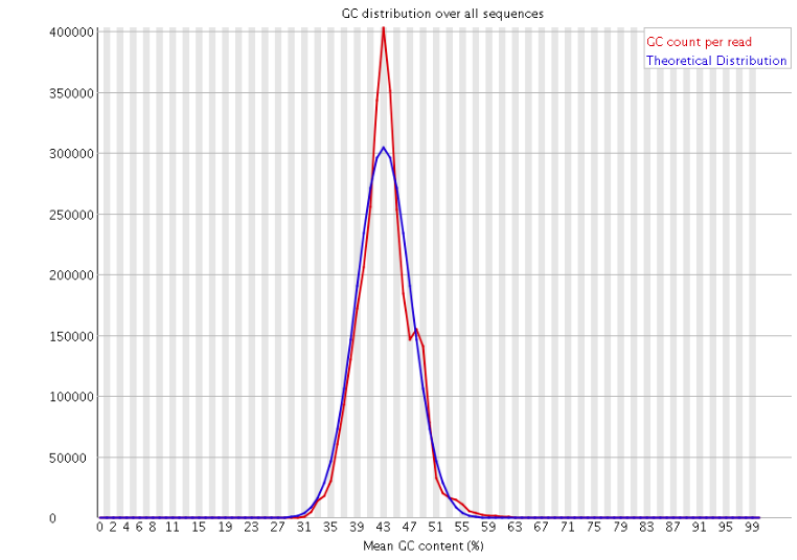
Per base sequence quality



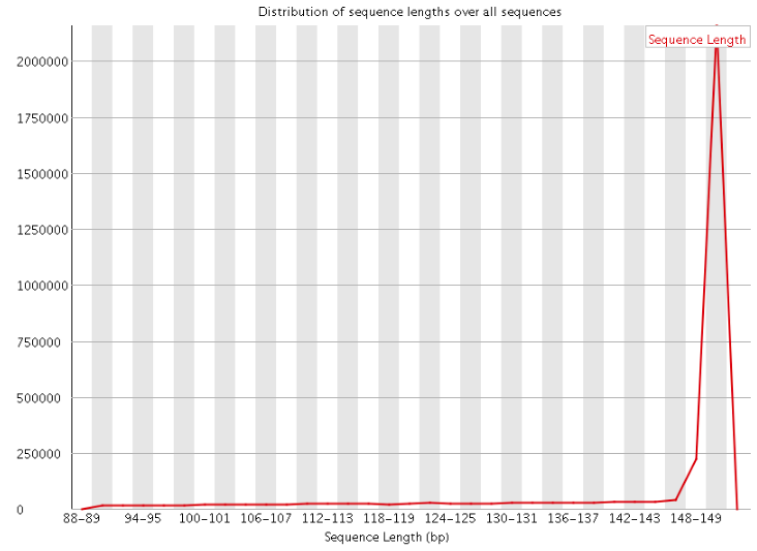
Per sequence quality scores



Per sequence GC content

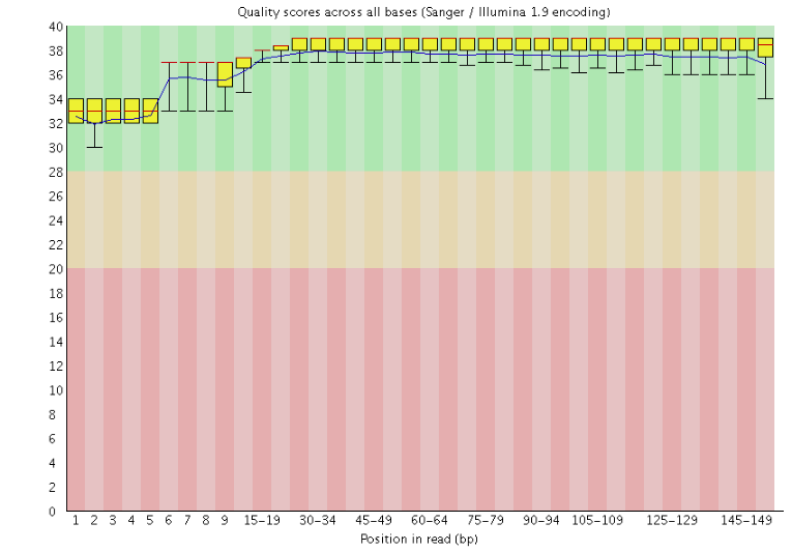


Sequence Length Distribution

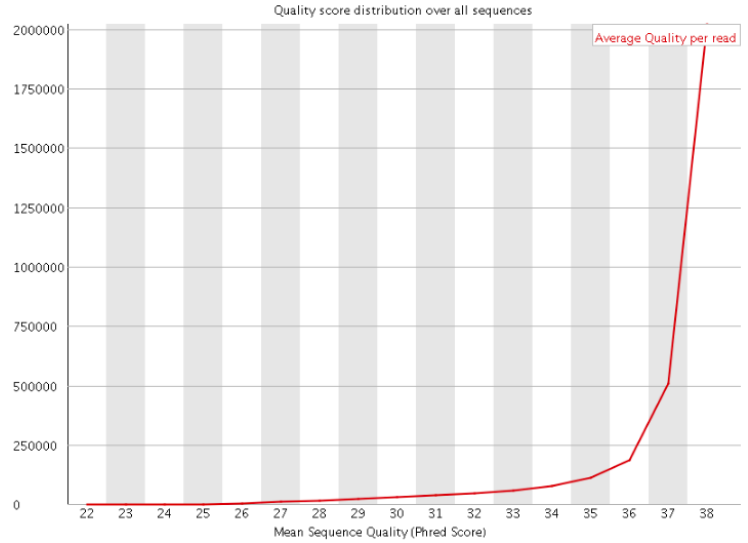


Datos trimados por calidad R2

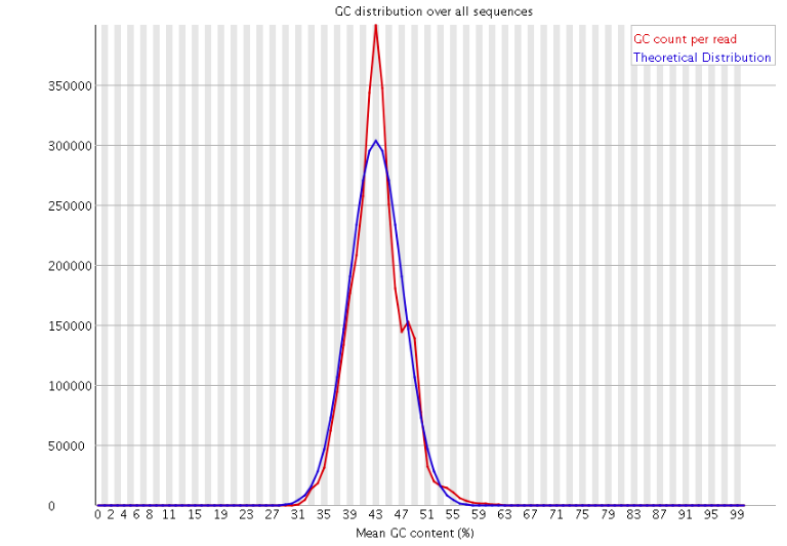
Per base sequence quality



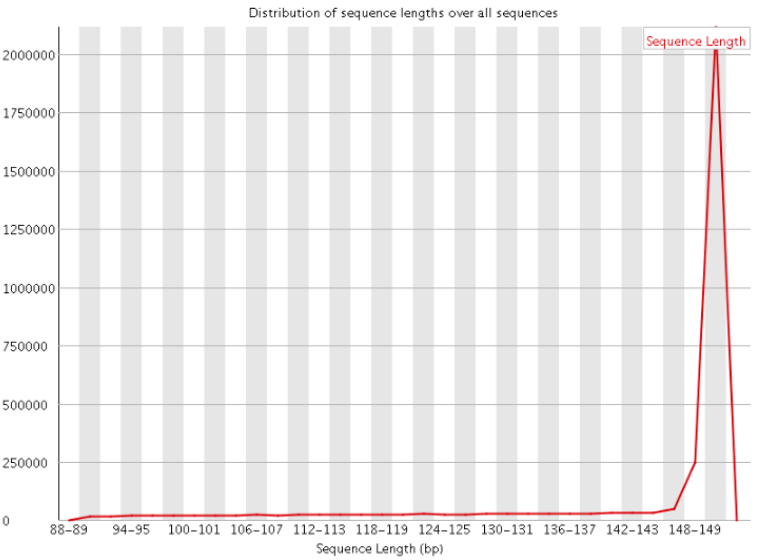
Per sequence quality scores



Per sequence GC content



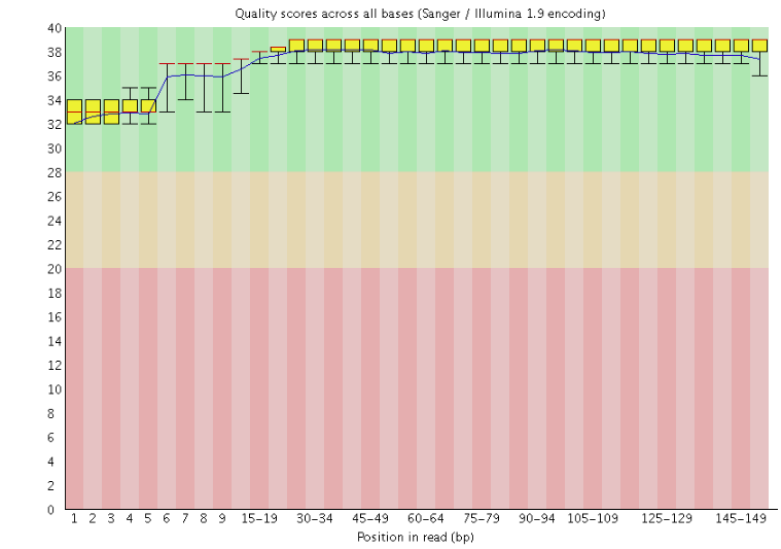
Sequence Length Distribution



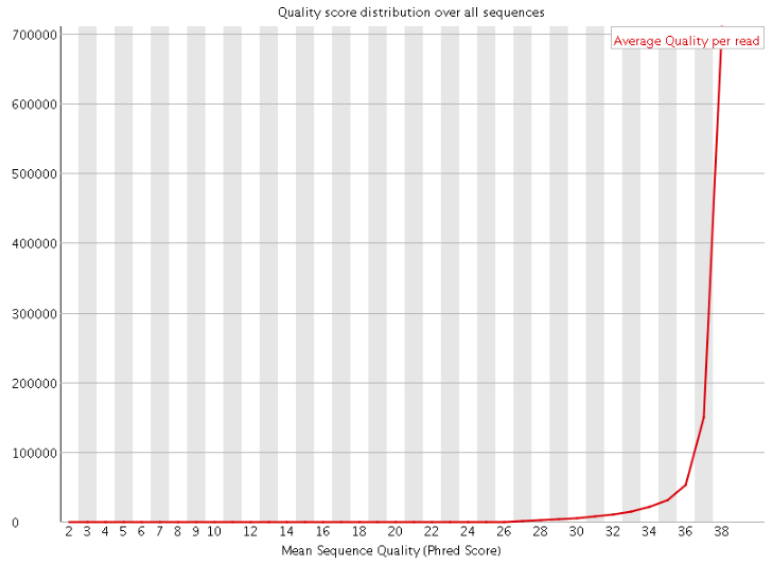
m264:

Datos Crudos R1

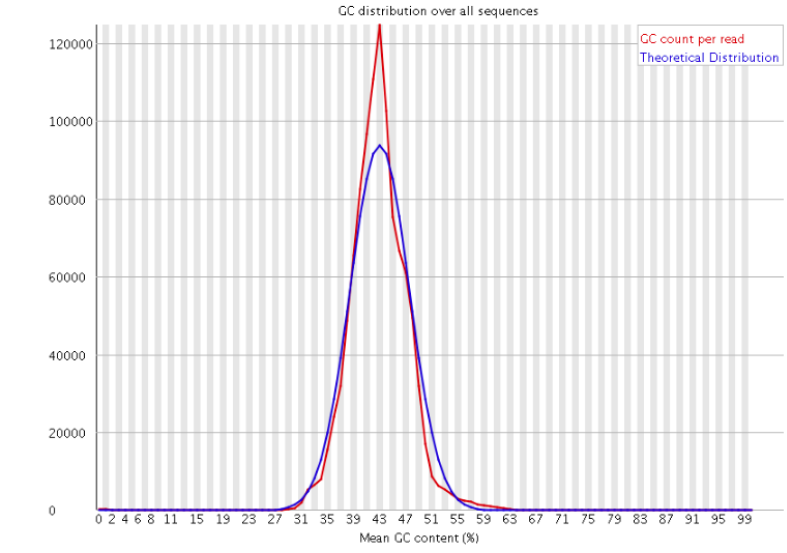
Per base sequence quality



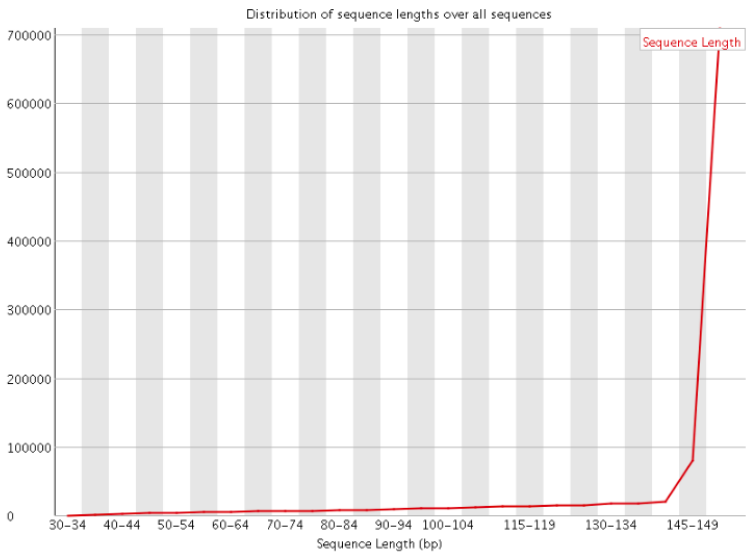
Per sequence quality scores



Per sequence GC content

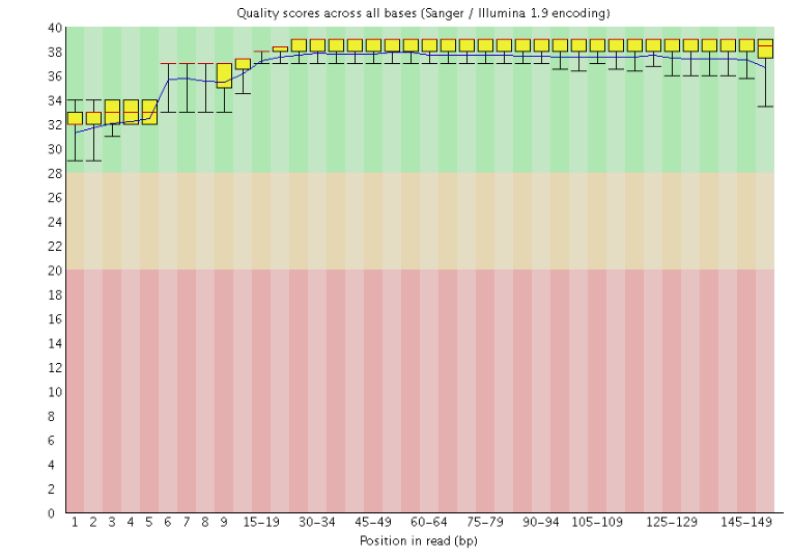


Sequence Length Distribution

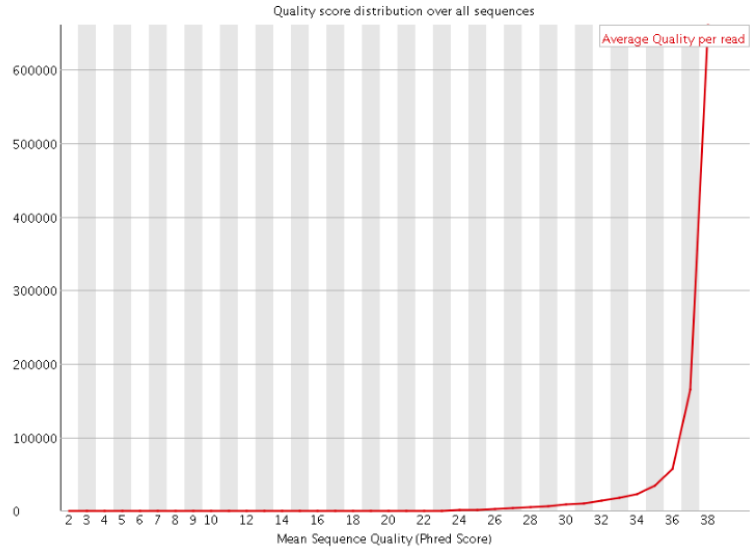


Datos Crudos R2

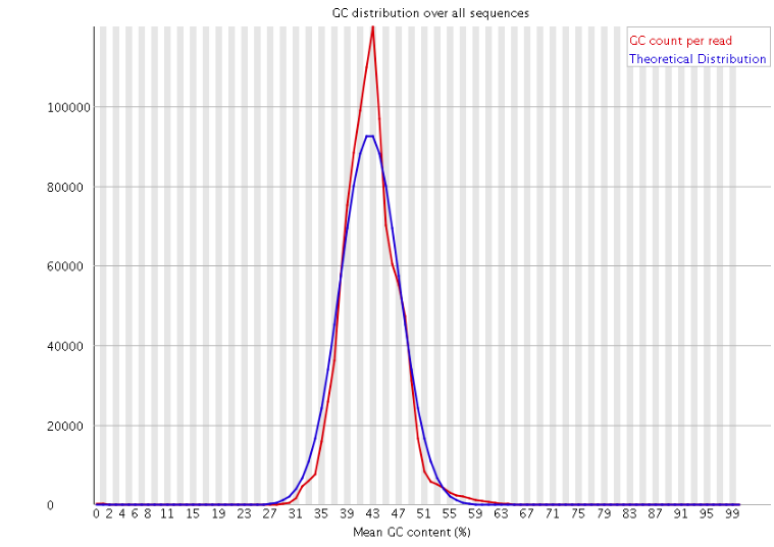
Per base sequence quality



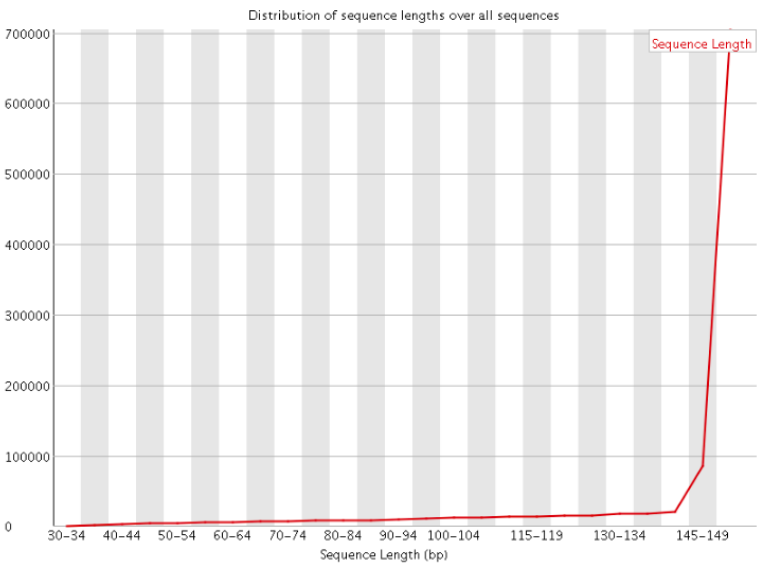
Per sequence quality scores



Per sequence GC content

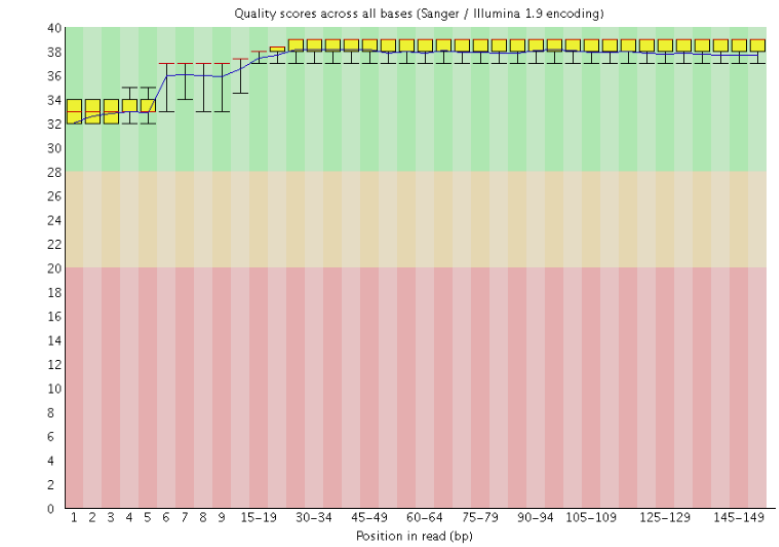


Sequence Length Distribution

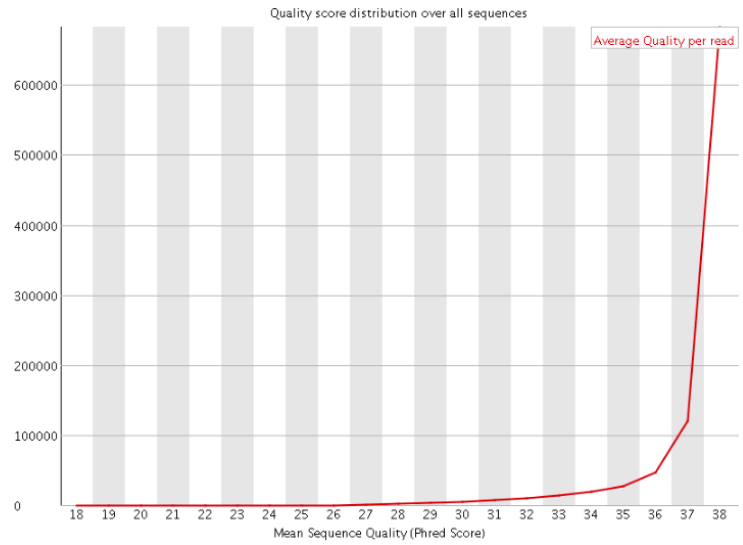


Datos trimados para primers y adaptadores R1

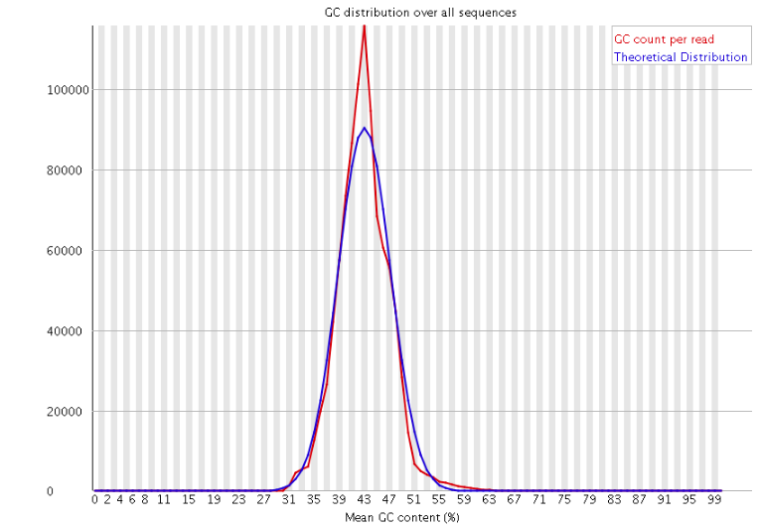
Per base sequence quality



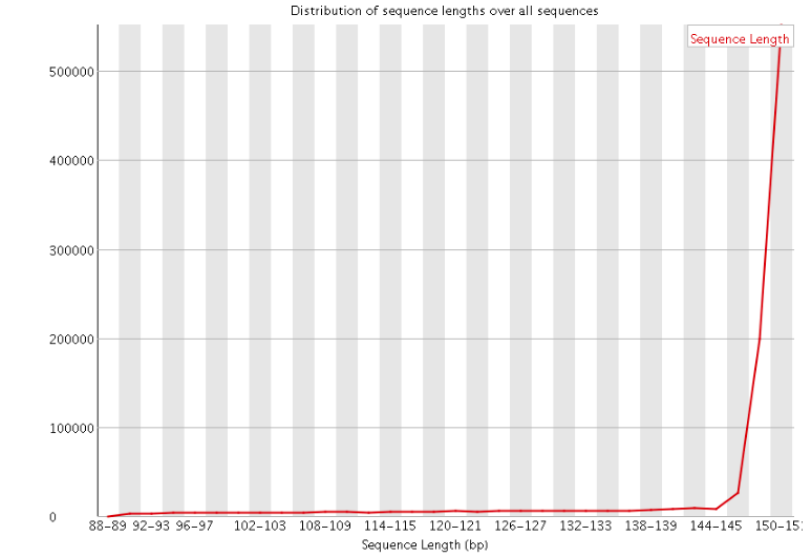
Per sequence quality scores



Per sequence GC content

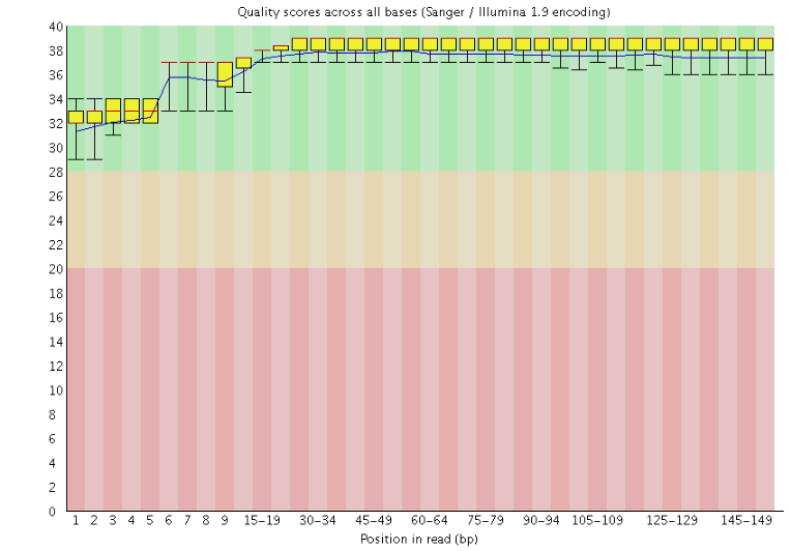


Sequence Length Distribution

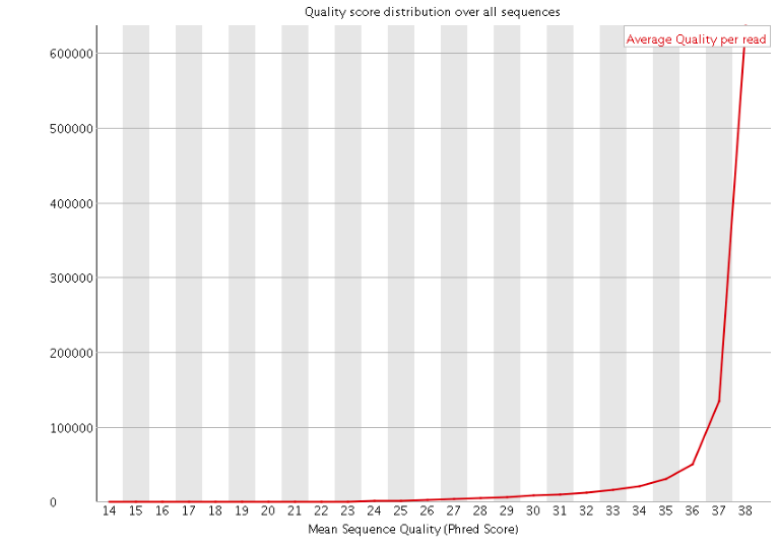


Datos trimados para primers y adaptadores R2

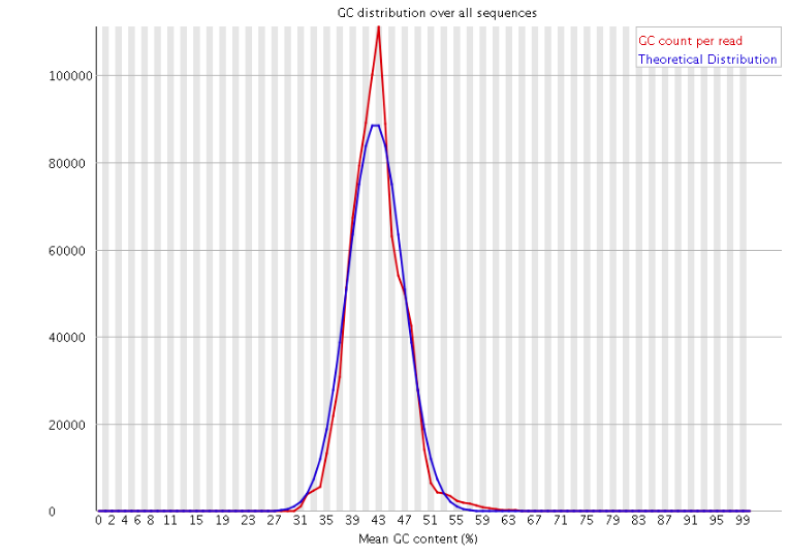
Per base sequence quality



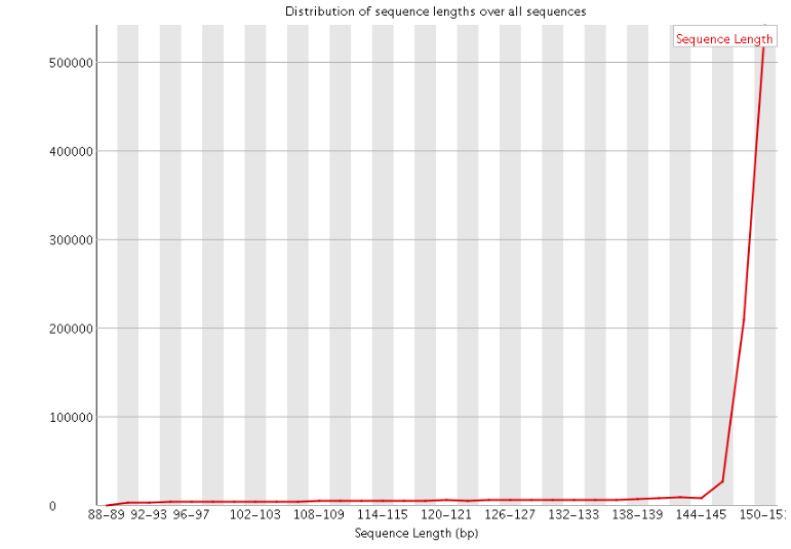
Per sequence quality scores



Per sequence GC content

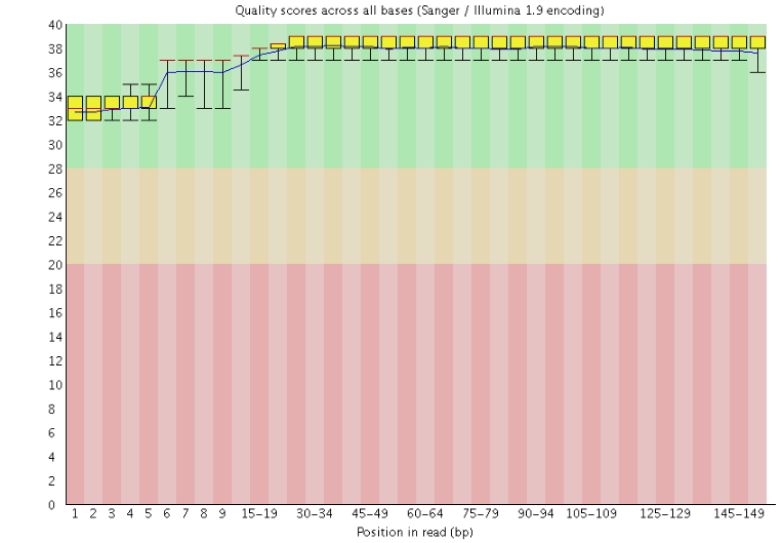


Sequence Length Distribution

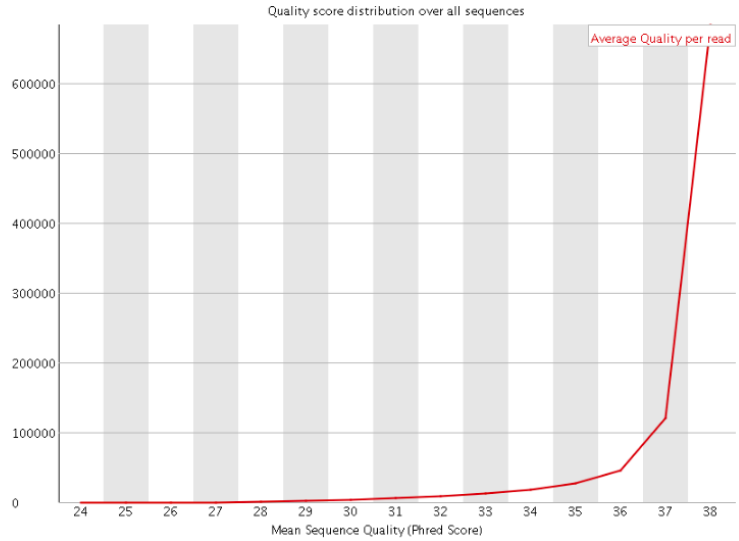


Datos trimados por calidad R1

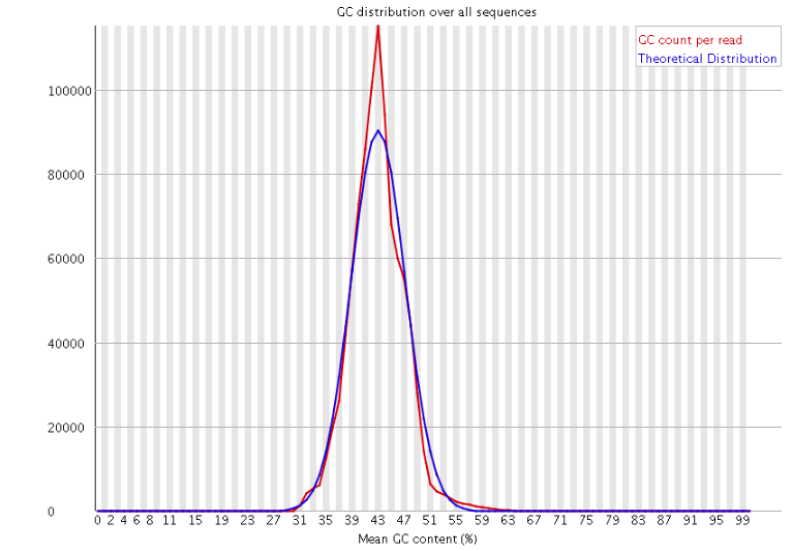
Per base sequence quality



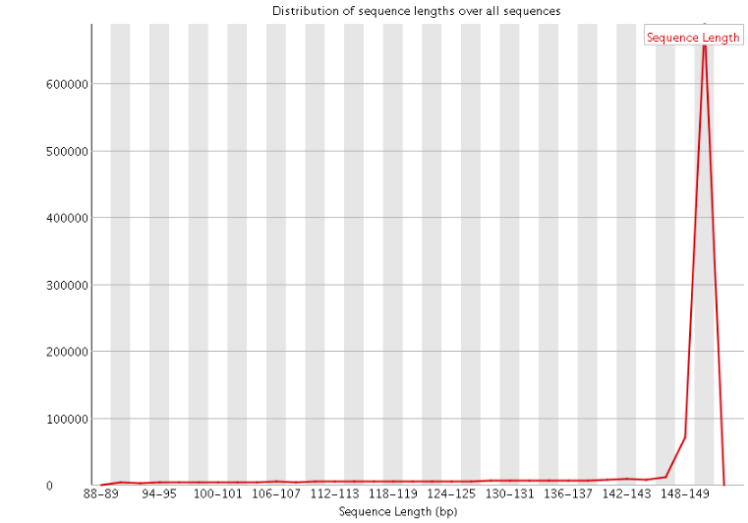
Per sequence quality scores



Per sequence GC content

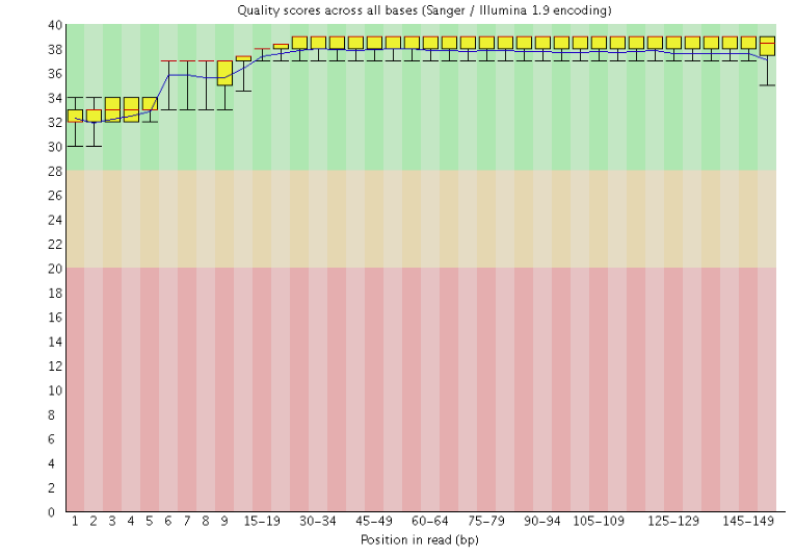


Sequence Length Distribution

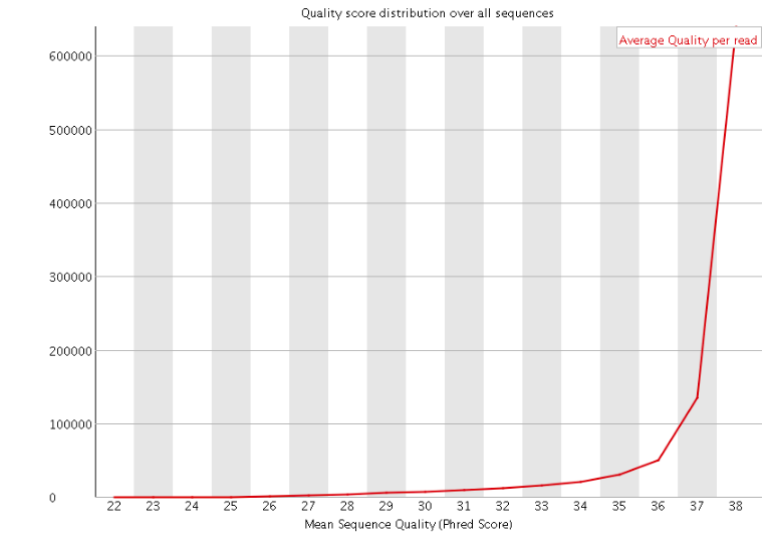


Datos trimados por calidad R2

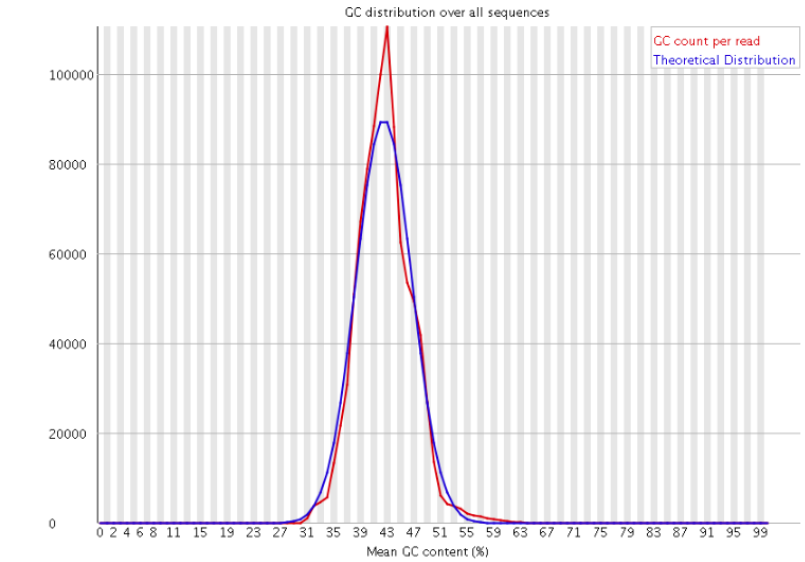
Per base sequence quality



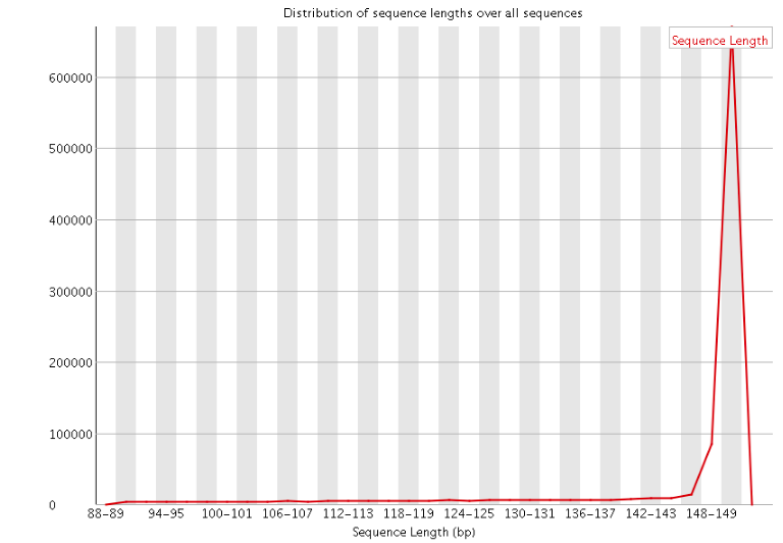
Per sequence quality scores



Per sequence GC content

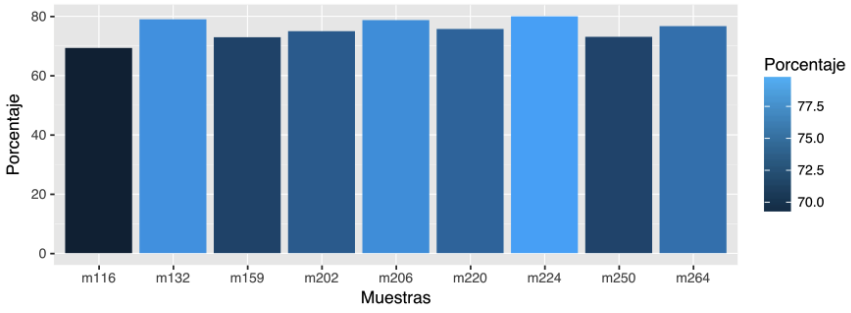


Sequence Length Distribution



Procesamiento de datos masivos:

Figura 2. Pre-Procesamiento: porcentaje de reads luego del trimado de calidad:

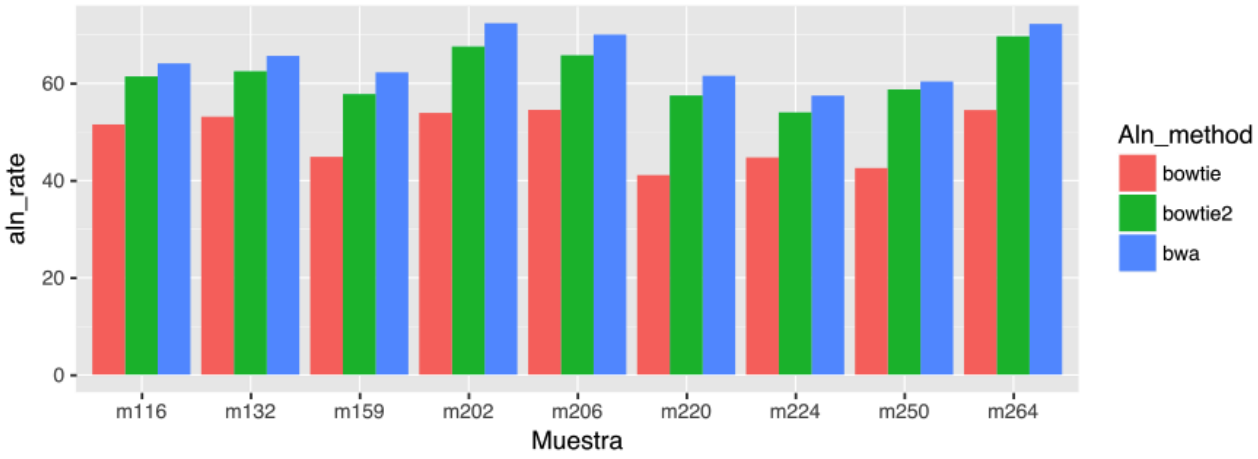


Alineamientos:

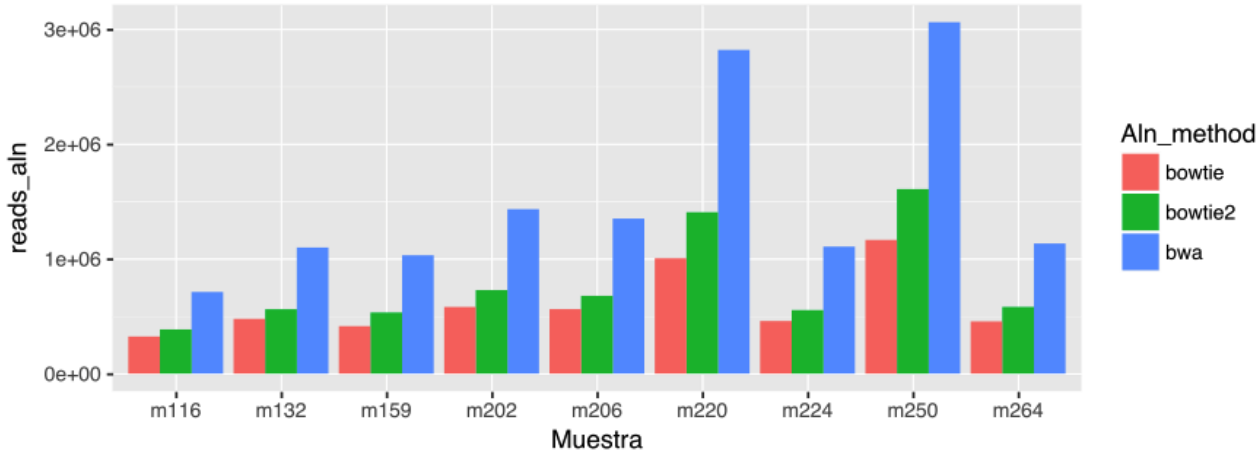
Figura 3. Conteo y porcentaje de reads específico de cada muestra y gen analizados

Resumen Alineamientos HA:

Porcentaje

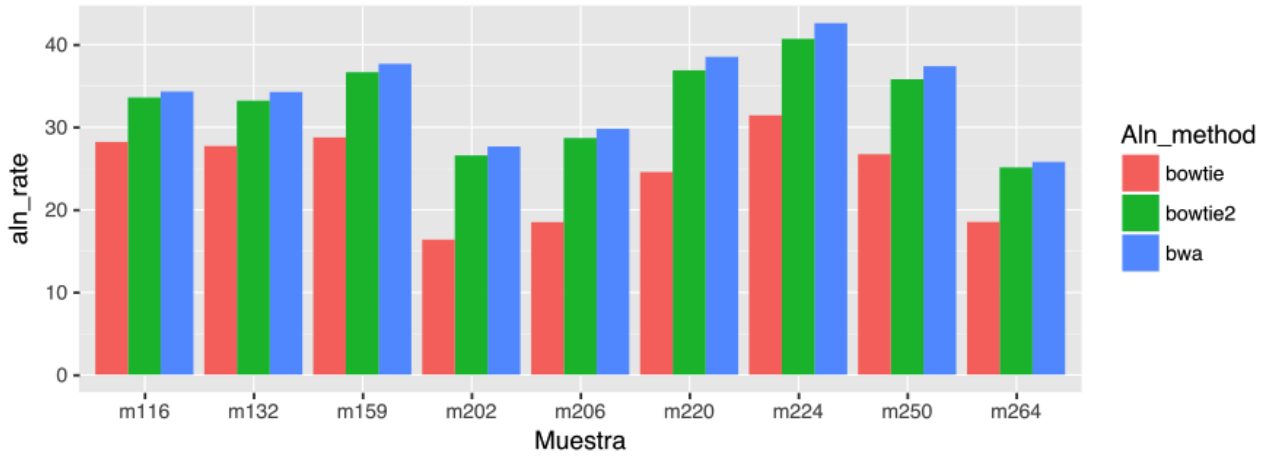


Conteo:



Resumen Alineamientos NA:

Porcentaje:



Conteo:

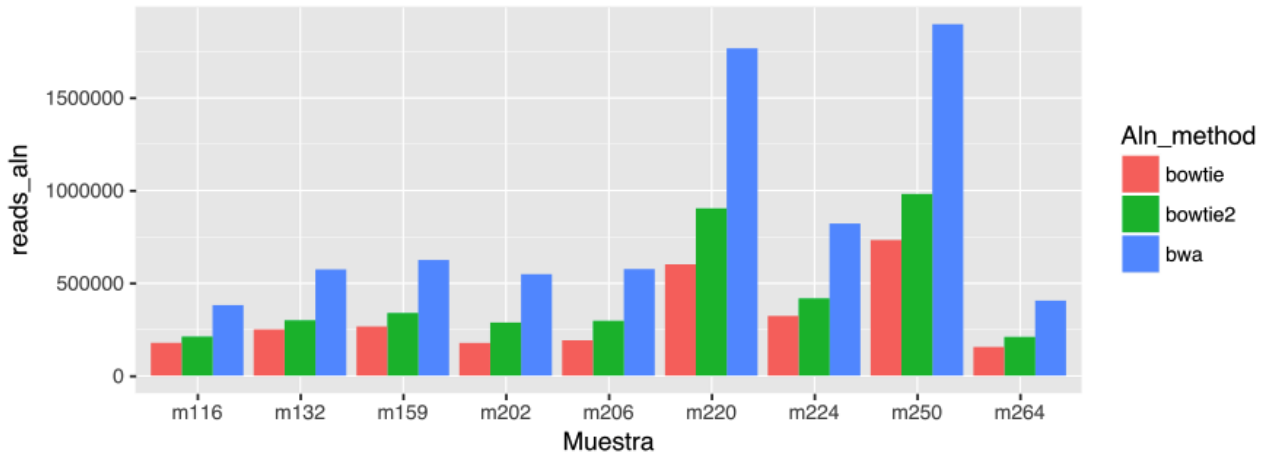
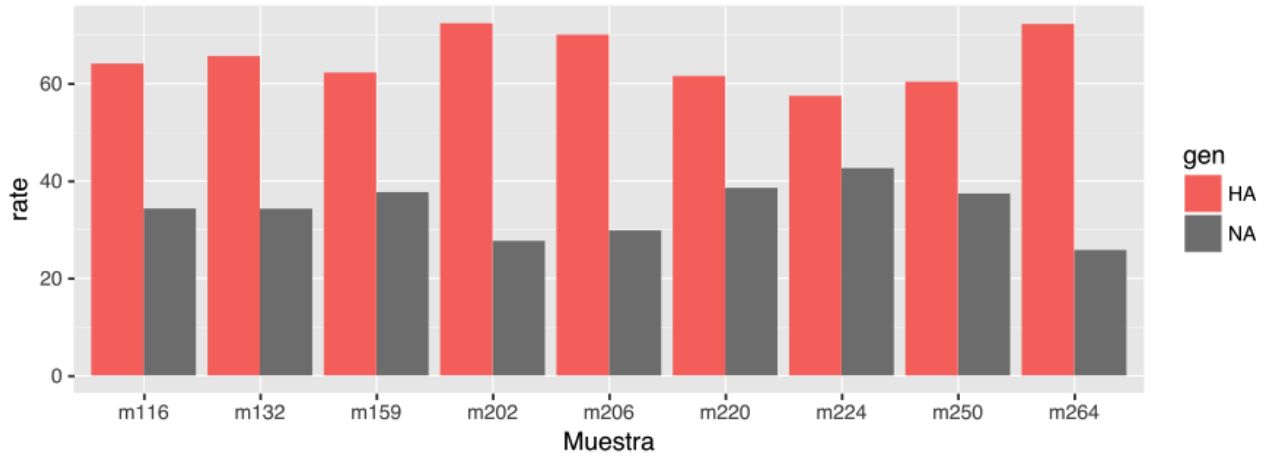


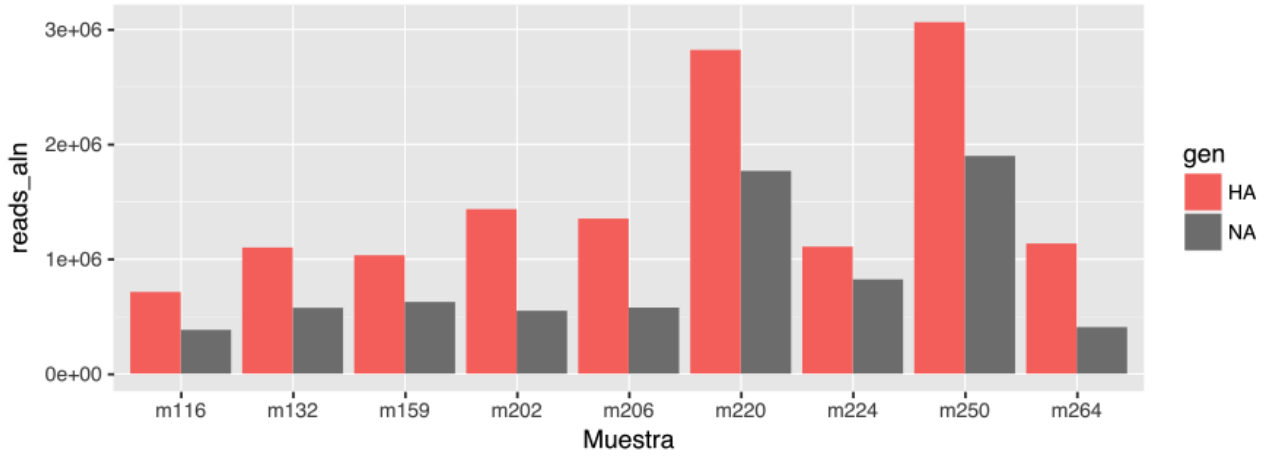
Figura 4. Comparación de los alineadores en función de cada muestra analizada.

Alineamiento con BWA

Porcentaje:

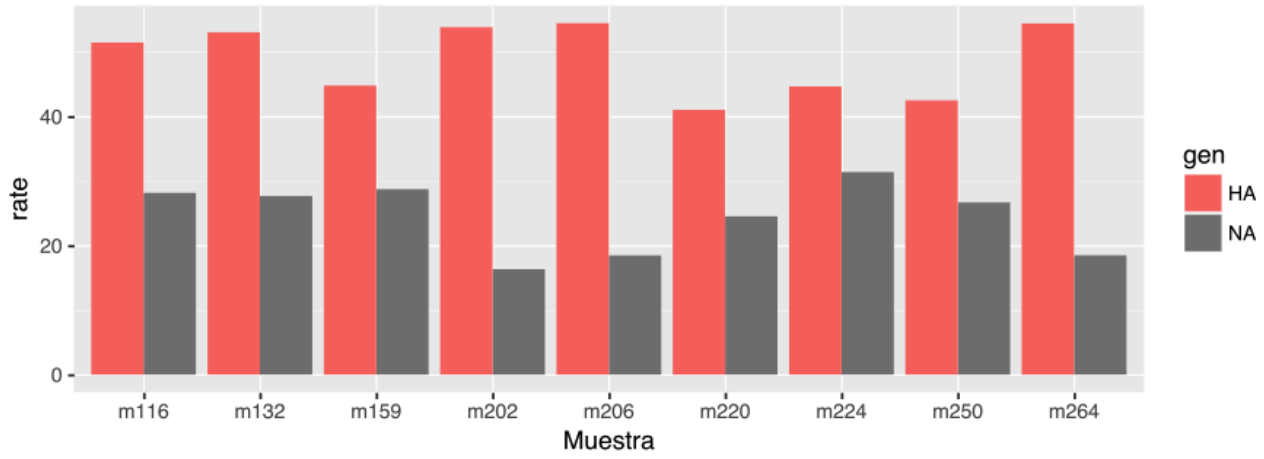


Conteo:

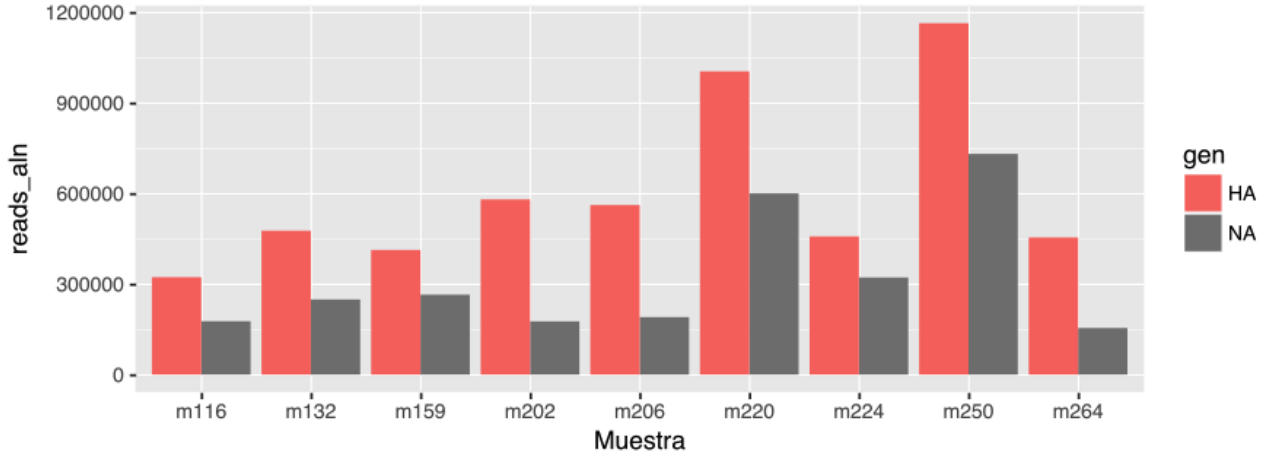


Alineamiento con Bowtie

Procentaje:

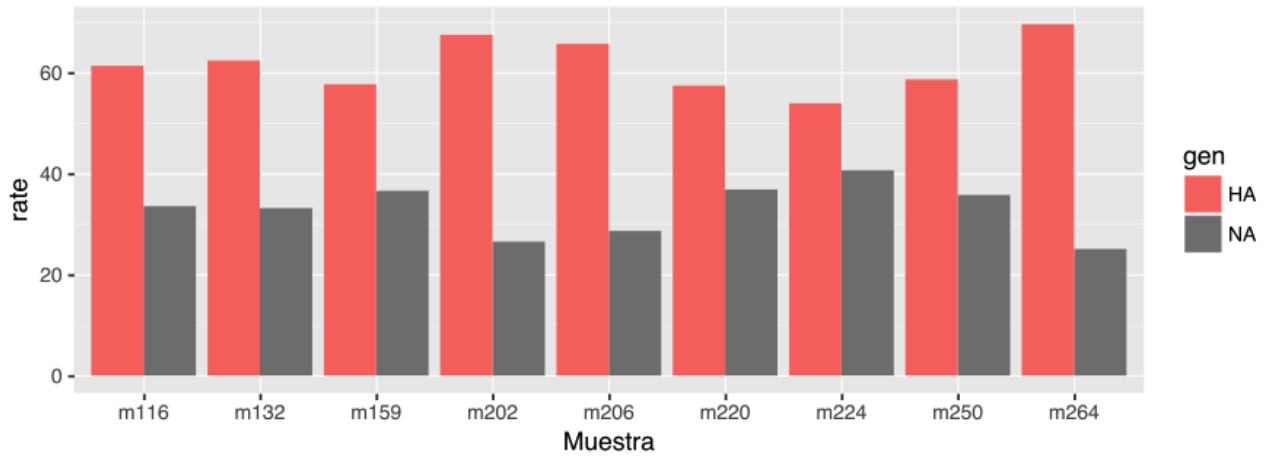


Conteo:

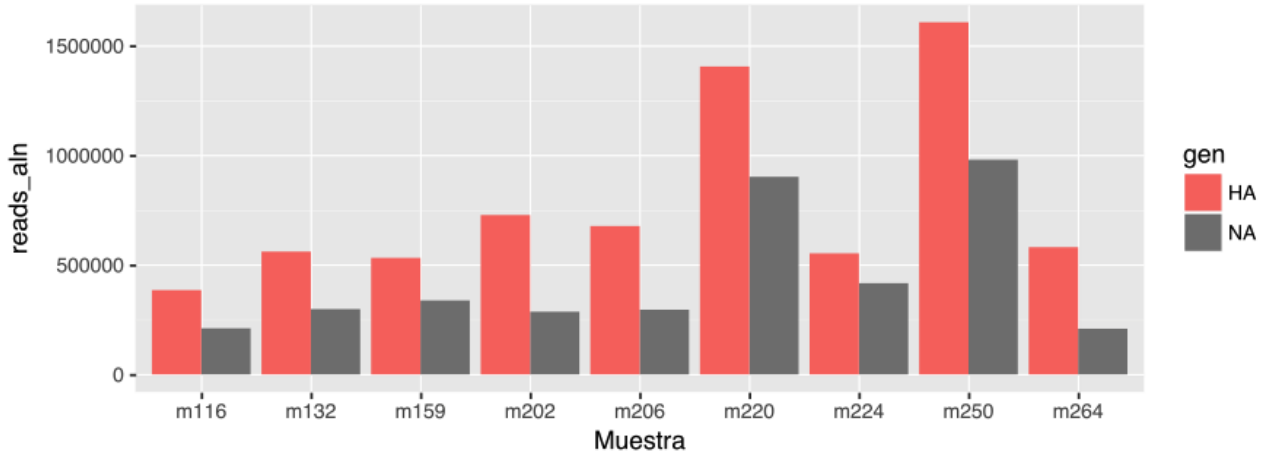


Alineamiento con Bowtie2

Porcentaje:



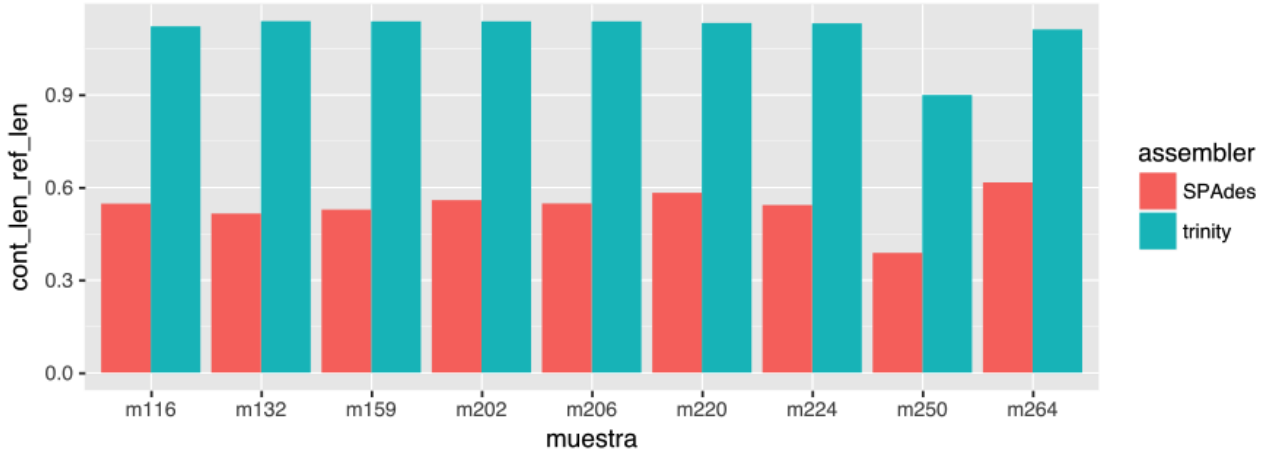
Conteo:



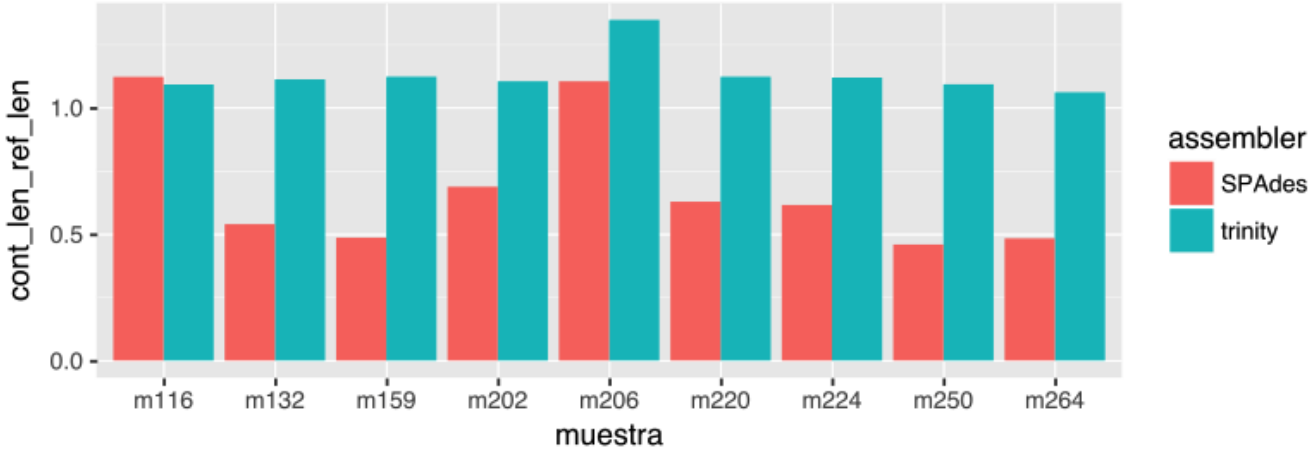
Ensambladores:

Figura 5. Resultados globales sobre los ensambladores. Relación largo de contig vs largo de gen.

HA)



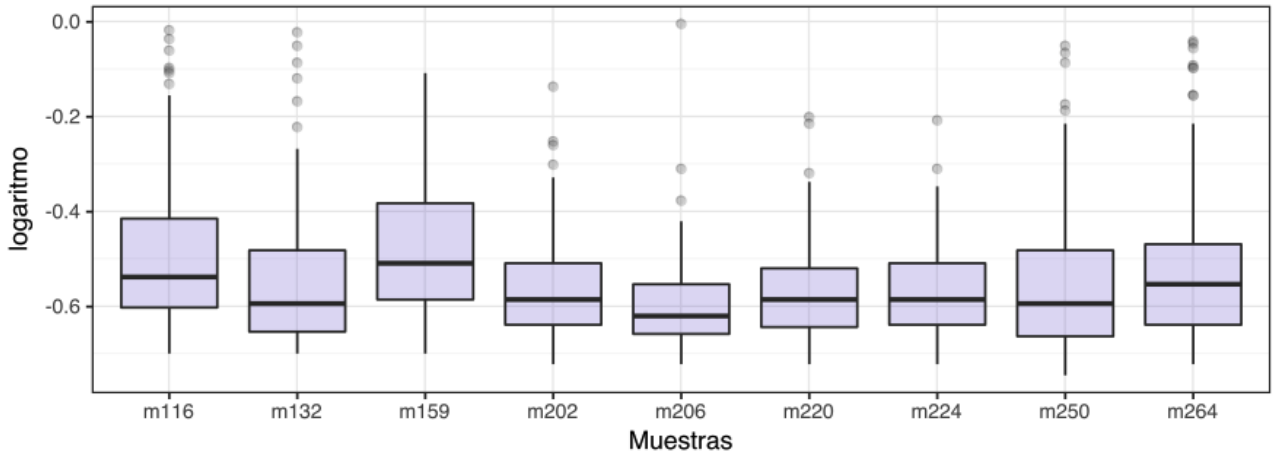
NA)

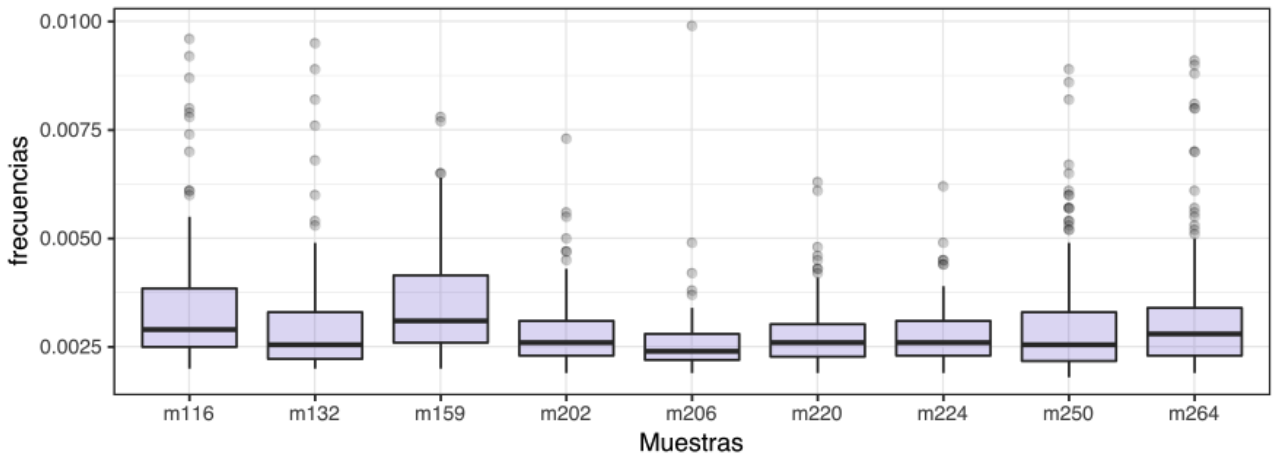


SNPs:

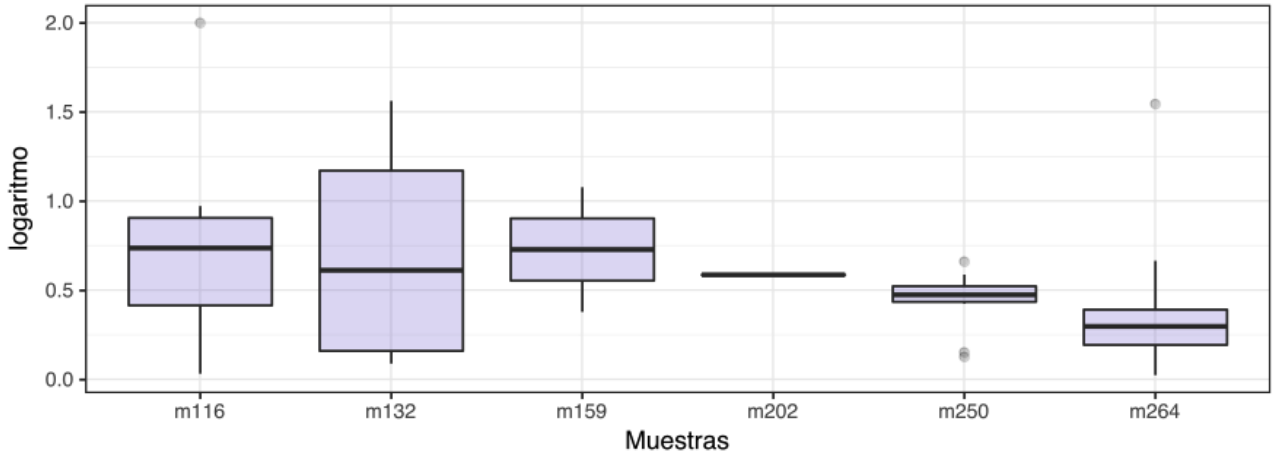
HA)

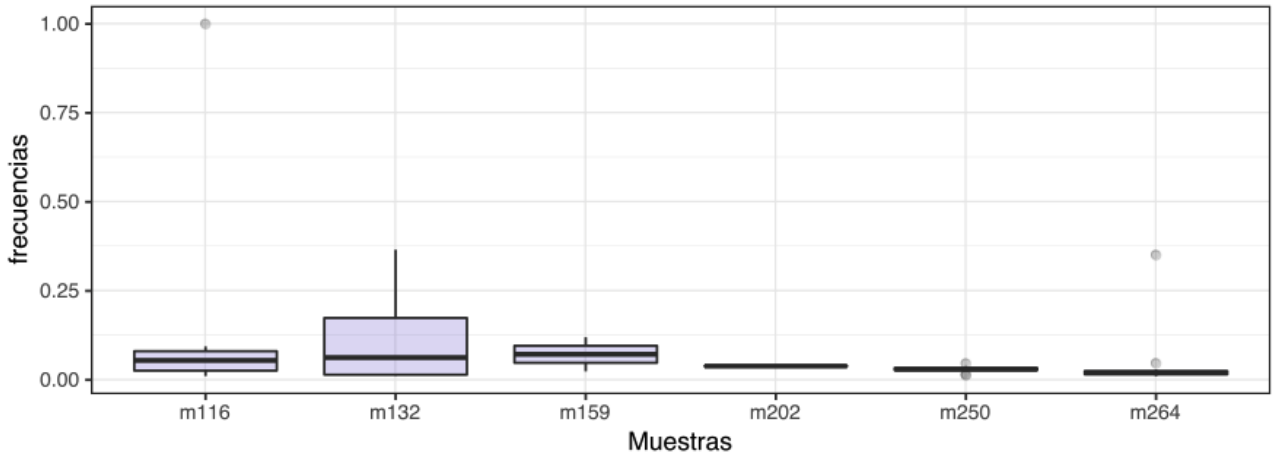
Minoritarias:





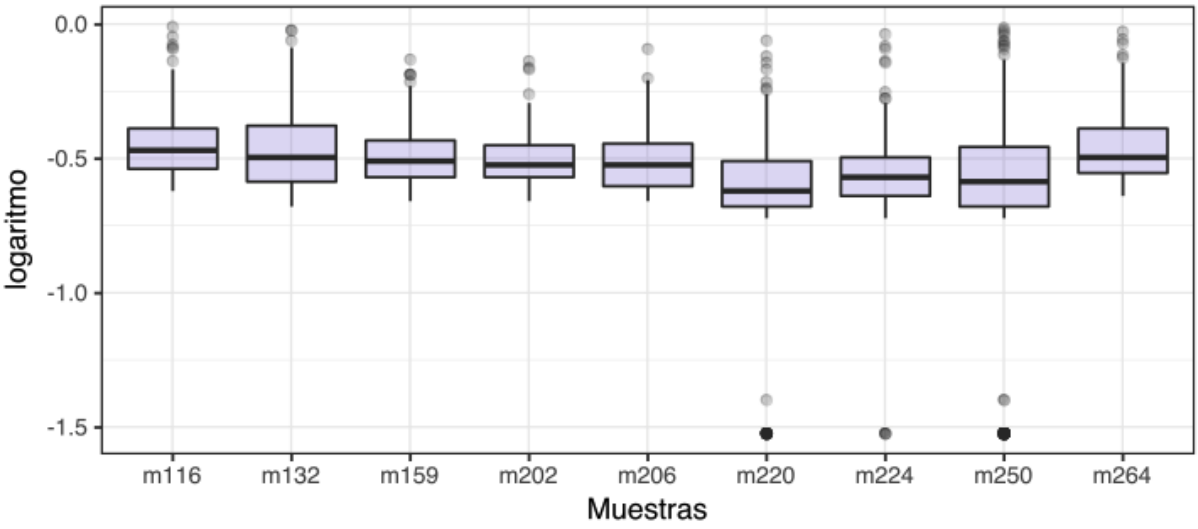
Mayoritarias:

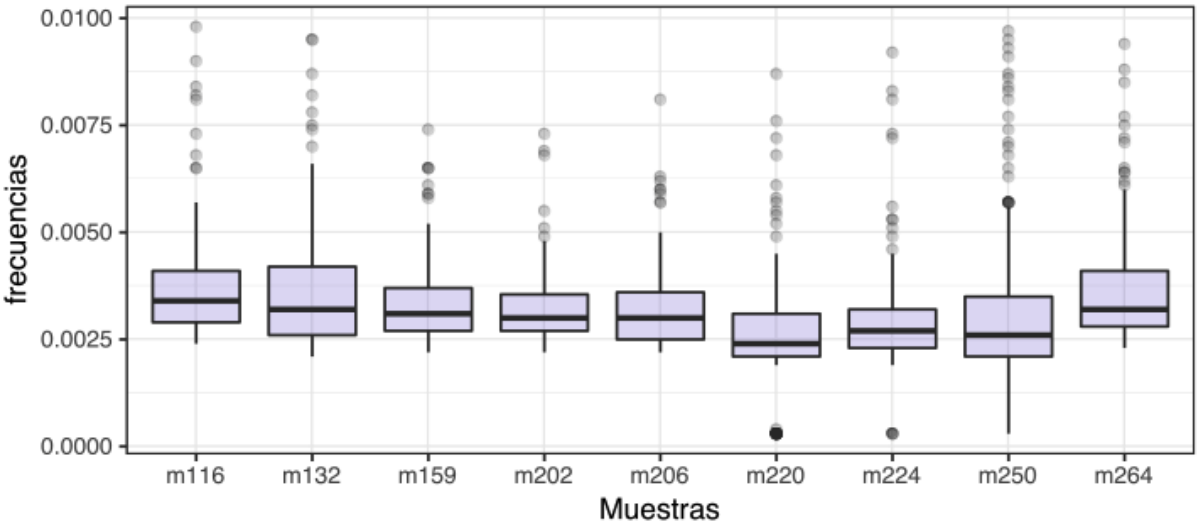




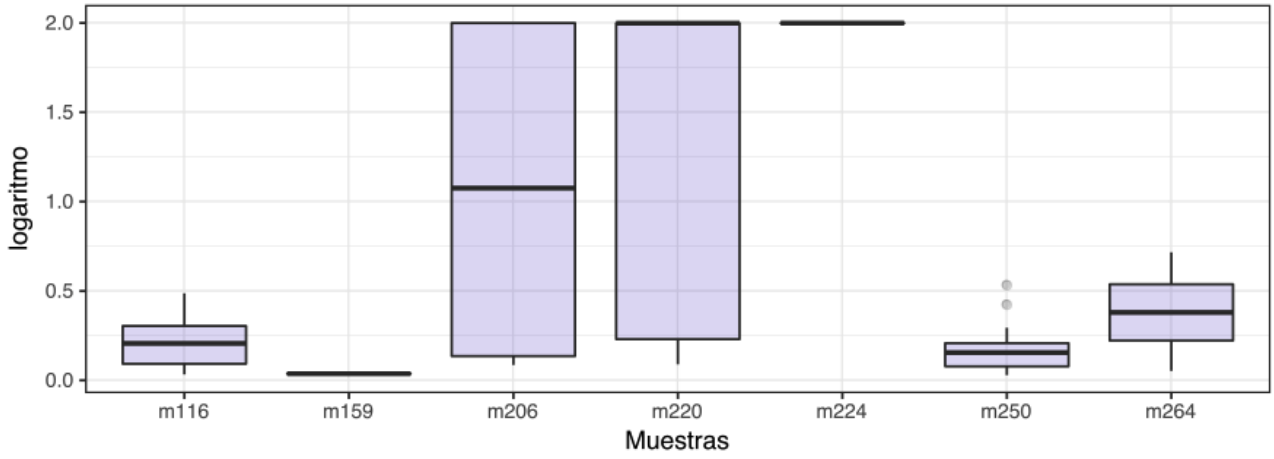
NA)

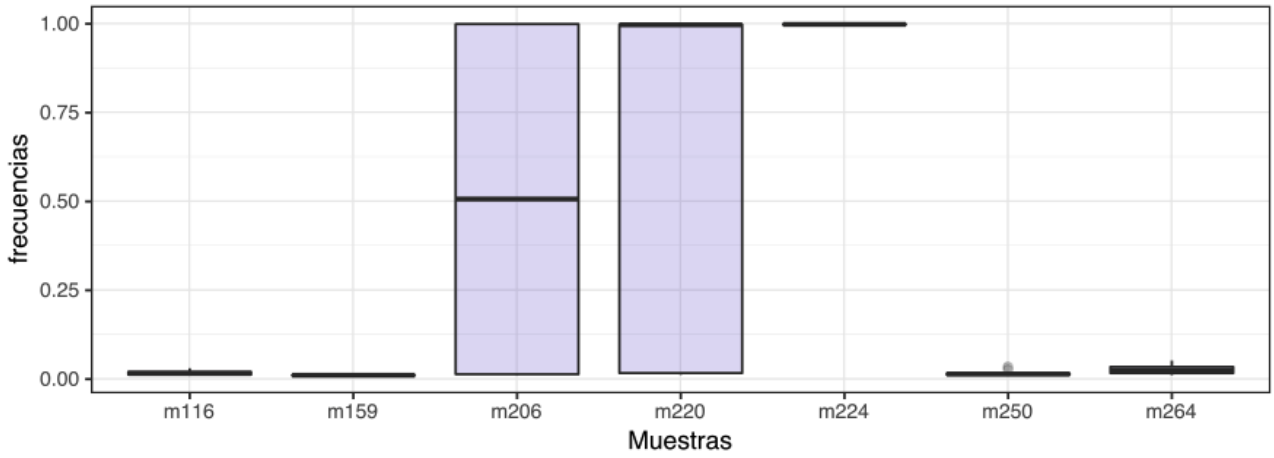
Minoritarias:





Mayoritarias:





Publicación derivada de esta tesis

Arch Virol
DOI 10.1007/s00705-015-2434-y

ORIGINAL ARTICLE

An evolutionary insight into Newcastle disease viruses isolated in Antarctica

Martin Soñora[1] · Pilar Moreno[1] · Natalia Echeverría[1] · Sabrina Fischer[1] · Victoria Comas[1] · Alvaro Fajardo[1] · Juan Cristina[1]

Received: 9 December 2014 / Accepted: 17 April 2015

Abstract The disease caused by Newcastle disease virus (NDV) is a severe threat to the poultry industry worldwide. Recently, NDV has been isolated in the Antarctic region. Detailed studies on the mode of evolution of NDV strains isolated worldwide are relevant for our understanding of the evolutionary history of NDV. For this reason, we have performed Bayesian coalescent analysis of NDV strains isolated in Antarctica to study evolutionary rates, population dynamics, and patterns of evolution. Analysis of F protein cleavage-site sequences of NDV isolates from Antarctica suggested that these strains are lentogenic. Strains isolated in Antarctica and genotype I reference strain Ulster/67 diverged from ancestors that existed around 1958. The time of the most recent common ancestor (MRCA) was established to be around 1883 for all class II viruses. A mean rate of evolution of 1.78×10^{-3} substitutions per site per year (s/s/y) was obtained for the F gene sequences of NDV strains examined in this study. A Bayesian skyline plot indicated a decline in NDV population size in the last 25 years. The results are discussed in terms of the possible role of Antarctica in emerging or re-emerging viruses and the evolution of NDV populations worldwide.

Electronic supplementary material The online version of this article (doi:10.1007/s00705-015-2434-y) contains supplementary material, which is available to authorized users.

✉ Juan Cristina
cristina@cin.edu.uy

[1] Laboratorio de Virologia Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la Republica, Igua 4225, 11400 Montevideo, Uruguay

Introduction

The disease caused by Newcastle disease virus (NDV) is one the most important diseases of poultry, affecting the poultry industry worldwide [1]. NDV belongs to the genus *Avulavirus* of the family *Paramyxoviridae*, and its genome is a non-segmented, single-stranded, negative-sense RNA molecule of approximately 15,186 nucleotides (nt) in length [2].

NDV isolates have been grouped by virulence phenotype, with lentogenic, mesogenic, and velogenic strains, in order of increasing virulence [3]. Lentogenic viruses typically cause subclinical infections or mild respiratory disease. Mesogens are of intermediate virulence, usually resulting in moderate respiratory disease with occasional nervous signs. Velogens are the most virulent viruses and may cause extensive hemorrhagic lesions, particularly in the gastrointestinal tract (viscerotropic), and/or a predominance of nervous signs (neurotropic) [4].

NDV infection is initiated by the action of two envelope glycoproteins. One of these mediates attachment of the virus to a host-cell receptor and is designated HN (hemagglutinin-neuraminidase). The other glycoprotein, designated as the fusion (F) protein, is responsible for virus penetration into the host cell and syncytium formation [5]. The F protein plays a key role in viral virulence and is a major target for the immune response [6]. The NDV F protein is a trimeric type I integral membrane protein that is synthesized as an inactive precursor, F0 (66 kDa), which is posttranslationally cleaved by host-cell proteases into two disulfide-linked subunits, the N-terminal F2 (12.5 kDa) and the C-terminal F1 (55 kDa) [7, 8]. The sequence of the F protein cleavage site is a major determinant of NDV pathogenicity. The cleavage sites of virulent NDV strains usually contain multiple basic residues, whereas avirulent strains have fewer basic residues [9].

Published online: 27 May 2015

The consensus sequence of the F protein cleavage site of velogenic and mesogenic strains is ^{112}(R/K)RQ(R/K)RF^{117}, while the consensus sequence of the lentogenic F cleavage site is ^{112}(G/E)(K/R)Q(G/E)RL^{117} [10]. Most of the recent virulent NDV strains bear the virulence motif ^{112}RRQKRF^{117} at the cleavage site of their F0 protein [11, 12]. Seven neutralizing epitopes have been mapped on the F protein of NDV [5, 13, 14]. Critical amino acids involved in neutralization sites are sites 72, 74, 75, 78, 79 and 343, as well as a stretch of amino acids from residues 157 to 171 [14, 15].

NDV strains are divided into two classes based on genetic analysis: class I strains, which are mainly isolated from wild birds and are generally avirulent, and class II strains, which are isolated from wild and domestic birds can be either virulent or avirulent [16]. Class I viruses comprise a single genotype, while class II viruses are divided into 18 or possibly 19 genotypes (I–XIX) [17–19]. Strains of genotypes V, VI and VII of class II are currently circulating in chickens throughout the world [20].

Since NDV was first reported in poultry in 1926, vaccination has been widely used for prevention and control of the disease caused by NDV [21]. The most commonly used live vaccines are LaSota and Clone-30, which belong to genotype II [22]. Characterization of NDV strains is important to evaluate field changes, anticipate new outbreaks, and develop adequate control measures [23]. Large gaps in our current knowledge in the areas of epidemiology and evolution limit the possibilities for controlling the disease [24, 25].

Three main panzootics have occurred in the last century. The first one (1926 to 1960) was caused by viruses belonging to genotypes II, III and IV, while the second (1960 to 1973) and third (1970–1980) were caused by viruses of genotypes V–VI [14]. Severe outbreaks in Western and Southern Europe [26, 27], South Africa [28] and Taiwan [29] in the 1990s were caused by genotype VII, the currently circulating genotype in Asia, Africa and Europe [14]. A recent outbreak of NDV in South America (Venezuela) has also been attributed to a genotype VII virus, suggesting that viruses of this genotype are spreading worldwide [30, 31].

In 2010, infection by virulent NDV was confirmed in 80 countries, including infections of wild birds in Canada, Germany, Israel, Italy, Kenya, Mongolia and the USA, and infections in domestic poultry in countries of North and South America, Europe, Africa, and Asia [32]. Moreover, recent studies revealed the isolation of NDV in penguins from King George Island in the Antarctic region [33]. Detailed studies on the mode of evolution of these new NDV strains are relevant for inferring the evolutionary history of NDV. In order to gain insight into these matters, Bayesian coalescent studies were performed to investigate the evolutionary rates, population dynamics and patterns of evolution of NDV.

Materials and methods

Sequences

Nucleotide sequences from NDV strains were obtained using ARSA from the DDBJ database (available at: http://arsa.ddbj.nig.ac.jp/). Strain names and accession numbers can be found in Supplementary Material Table 1.

Sequence alignment and *in silico* translation of nucleotide sequence

Sequences were aligned using the MUSCLE program [34]. Nucleotide sequences were translated to amino acids *in silico* using software from the MEGA 5 program [35].

Bayesian coalescent Markov chain Monte Carlo (MCMC) analysis

In order to gain insight into the evolutionary rate and mode of evolution of NDV strains, we used a Bayesian Markov MCMC approach as implemented in the BEAST package v.1.7.5 [36]. For strains included in these analyses, see Supplementary Material Table 1. First, software from the Datamonkey server [37] was used to identify the optimal evolutionary model that best fitted our sequence dataset. Akaike information criteria and the hierarchical likelihood ratio test indicated that the HKY + Γ model was the most accurate. Using this model and 50 million steps of MCMC, different population models were tested (constant population size, exponential population growth, expansion population growth, logistic population growth and Bayesian Skyline). Statistical uncertainty in the data was reflected by the 95 % highest probability density (HPD) values. Results were examined using the TRACER v1.5 program (available from http://beast.bio.ed.ac.uk/Tracer) from the BEAST package. Convergence was assessed with ESS (effective sample size) values after a burn-in of 2 million steps. Models were compared by calculating the Bayes factor (BF) [38] from the posterior output of each of the models using the TRACER v1.5 program as explained on the BEAST website (http://beast.bio.ed.ac.uk/Model comparison). A log BF (natural log units) values greater than 2.3 indicates strong evidence against the null model. The Bayesian skyline model was the best fit to the data. Maximum clade credibility trees were generated using the Tree Annotator program from the BEAST package and the FigTree program v1.4.1 (available at: http://tree.bio.ed.ac.uk) was used for the visualization of the annotated trees.

Bayesian skyline plots (BSPs) were used to infer how the effective population size has changed over time [38, 39].

Results

Mapping of amino acid substitutions found in the fusion proteins of NDV strains isolated in Antarctica

Previous studies have identified NDV strains isolated in Antarctica as class II strains [33]. In order to gain insight into the virulence status of these strains, partial F gene sequences from NDV isolates from Antarctica (positions 4502 to 4995 relative to NDV reference strain LaSota, accession number AF077761) were aligned with the corresponding sequences of members of nine genotypes of class II strains for which complete genome sequences had been determined. For names and accession numbers of NDV strains included in this analysis, see Supplementary Material Table 1. Once aligned, they were translated *in silico* to amino acids using the MEGA 5 program [35], and the results are shown in Figure 1.

The F protein cleavage-site sequence of NDV isolated in Antarctica is ^{112}GKQGRLI^{118}, suggesting that the NDV strains isolated in that region of the world and included in these studies are lentogenic strains. Nevertheless, more studies will be needed to address this issue. Moreover, no amino acid substitutions were found at positions 72, 74, 75, 78 and 79 of the F2 protein, which were previously shown to be involved in neutralization [15]. An N-linked glycosylation acceptor site (N-X-S/T, where X corresponds to any amino acid except aspartic acid or proline) at position 85-87 of the F2 protein is also conserved [9, 40], as are the cysteine residues at positions 25 and 76 of the F2 protein [41].

Bayesian coalescent analysis of NDV strains isolated in Antarctica

In order to determine the evolutionary rate and mode of evolution of the NDV population, we used a Bayesian Markov chain Monte Carlo (MCMC) approach as implemented in the BEAST package [36]. In this case, the same F gene sequences from NDV strains isolated in Antarctica were aligned with corresponding sequences from 74 NDV strains, representing class I and genotypes I to XIX of class II strains. Names and accession numbers of NDV strains included in these analyses can be found in Supplementary Material Table 1. After performing the alignment and determining that the optimal evolutionary model is HKY + Γ, different population dynamic models were

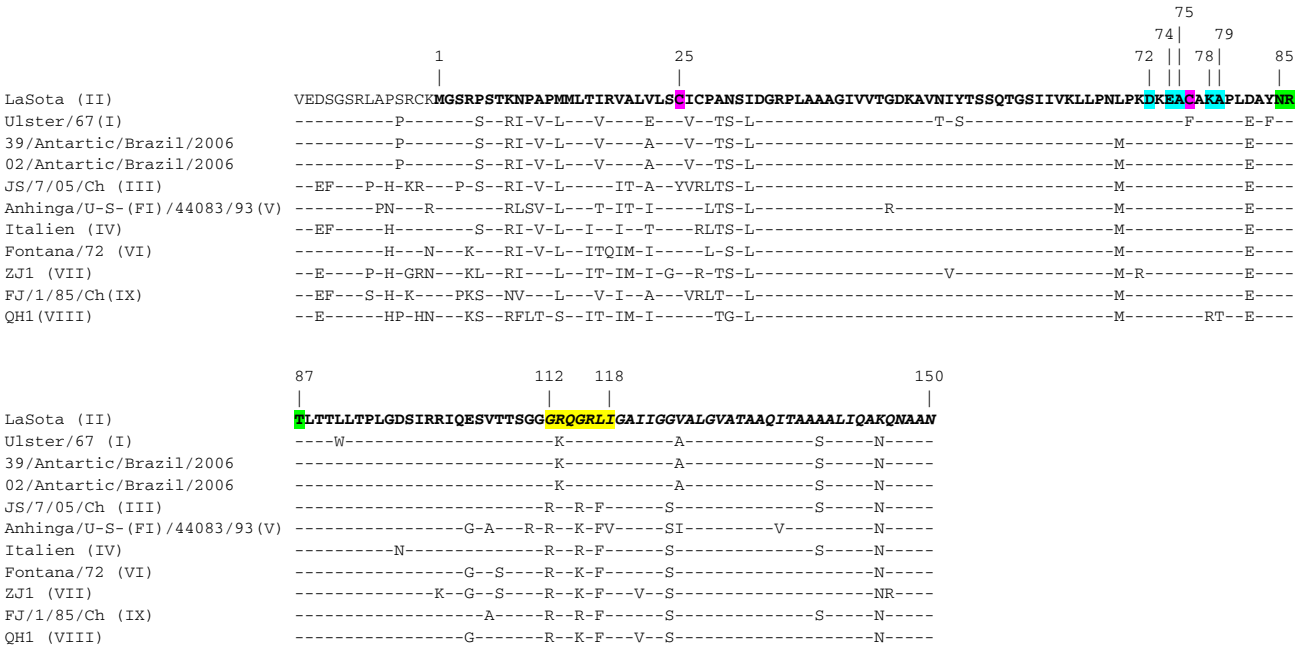


Fig. 1 Alignment of F amino acid sequences of NDV strains. Strain names are shown at the left side of the figure, and their class II genotype is indicated in parentheses. Identity to the LaSota strain (genotype II) is indicated by a dash. F2 sequences are shown in bold, and F1 sequences are shown in bold and italics. Numbers above the alignment indicate amino acid positions. The F protein cleavage site is highlighted in yellow. Amino acid substitutions detected in antigenic sites in neutralization escape mutants are indicated in turquoise [5, 6, 13]. A potential acceptor site for N-linked glycosylation at residues 85-87 is highlighted in green [8]. Cysteine residues at positions 26 and 76, which are conserved among most NDV isolates, are highlighted in fuchsia [22]

tested. The results for 50 million steps of MCMC analysis, using the HKY + Γ model, a relaxed clock and the Bayesian skyline model [42] are shown in Table 1. A mean rate of 1.78×10^{-3} substitutions per site per year (s/s/y) was obtained for the F gene sequences of NDV strains used in these studies. A maximum clade credibility tree revealed that all class II genotype strains have evolved from ancestors that existed around 1883 (130 years before the most recent isolates included in these studies, see Fig. 2). Both classes of NDV strains evolved from ancestors that existed around 1819 (Table 1). Strains isolated in Antarctica and genotype I reference strain Ulster/67 diverged from ancestors around 1958 (Fig. 2). BSPs suggested that a constant effective population size was maintained until the late 1980s (Fig. 3), where a decline in the population is observed.

Discussion

NDV strains isolated from penguins in Antarctica were assigned to genotype I of class II (Fig. 2), in agreement with previous reports [33] and with antigenic studies of NDV isolated from penguins from Antarctica that showed a reaction against a monoclonal antibody raised against NDV Ulster/67 strain (genotype I) [43]. Viruses of this genotype have been associated with outbreaks in Australia that occurred between 1998 and 2000 [44]. Genotype I viruses from these same outbreaks were found to be velogenic, and previous reports have shown that the origin of these viruses can be traced back to low-virulence NDV strains circulating in waterfowl just prior to the outbreak [45].

NDV strains circulating in one particular avian species may have the ability to cause disease in other avian species. For example, NDV strains from pigeons have been reported to be responsible for outbreaks in chickens [46–48]. Moreover, virtually all domestic and wild bird species are susceptible to infection with NDV [49]. Therefore, although the possibility of direct contact between penguins and chickens seems unlikely, other wild birds may act as carriers of different NDV strains through transmission routes that are not yet fully understood [19].

The presence of NDV strains in Antarctica, where other avian species live, indicates the importance of NDV strain characterization in all regions of the world. Genotypes V, VI, and VII of class II are currently circulating worldwide in chickens [20]. The role of Antarctica in maintaining other NDV genotypes not circulating at the moment also reinforces the relevance of in-depth NDV surveillance studies.

The F protein cleavage-site sequence has been shown to be a major determinant of NDV virulence [50]. The F protein cleavage sites of NDV strains isolated in Antarctica were found to have the consensus cleavage site of avirulent strains (Fig. 1). These cleavage sequences are insensitive to intracellular proteases and depend on extracellular secreted proteases for cleavage, limiting the replication of avirulent strains to the respiratory and enteric tracts [8–10]. More studies will be needed in order to confirm the avirulent (lentogenic) phenotype of NDV isolated from penguins in Antarctica.

Bayesian coalescent analysis revealed a rate of evolution of 1.78×10^{-3} s/s/y for NDV strains (see Table 1). This evolutionary rate is slightly higher than the rate estimated in a recent study for full-length NDV F gene sequences (1.35×10^{-3} s/s/y), although it lies within the confidence intervals of these estimations (0.71 - 1.98×10^{-3} s/s/y) [21]. This evolutionary rate is comparable to rates previously estimated for other fast-evolving RNA viruses such as human immunodeficiency virus type 1 (gp160env; 2.4×10^{-3} s/s/y) [51], human respiratory syncytial virus (G; 1.9×10^{-3} s/s/y) [52] and hepatitis C virus (E2; 3.4×10^{-3} s/s/y) [53].

The time of the most recent common ancestor (MRCA) was established to be around 1883 for all class II viruses (Fig. 2). This estimate is in agreement with previous reports that established the time of the MRCA for class II NDV strains to be around 1885 [21]. This finding is also in line with studies done by Macpherson in 1956, which suggest that a disease outbreak in domestic birds in

Table 1 Bayesian coalescent inference of Newcastle disease viruses

Group[a]	Parameter	Value[b]	HPD[c]	ESS[d]
F gene sequences	Log likelihood	-5576	-5595 to -5558	4010
	Posterior	-9101	-9055 to -9150	402
	Prior	-3525	-3573 to -3483	287
	Mean rate[e]	1.78×10^{-3}	9.22×10^{-4} to 2.56×10^{-3}	228
	Root age (years)	194	104 to 308	221
	MRCA[f]	1819	1705 to 1909	

[a] See Supplementary Material Table 1 for strains included in this analysis. [b] In all cases, mean values are shown. [c] High probability density values. [d] Effective sample size. [e] Mean rate was calculated in substitutions/site/year. [f] Year of the most common recent ancestor

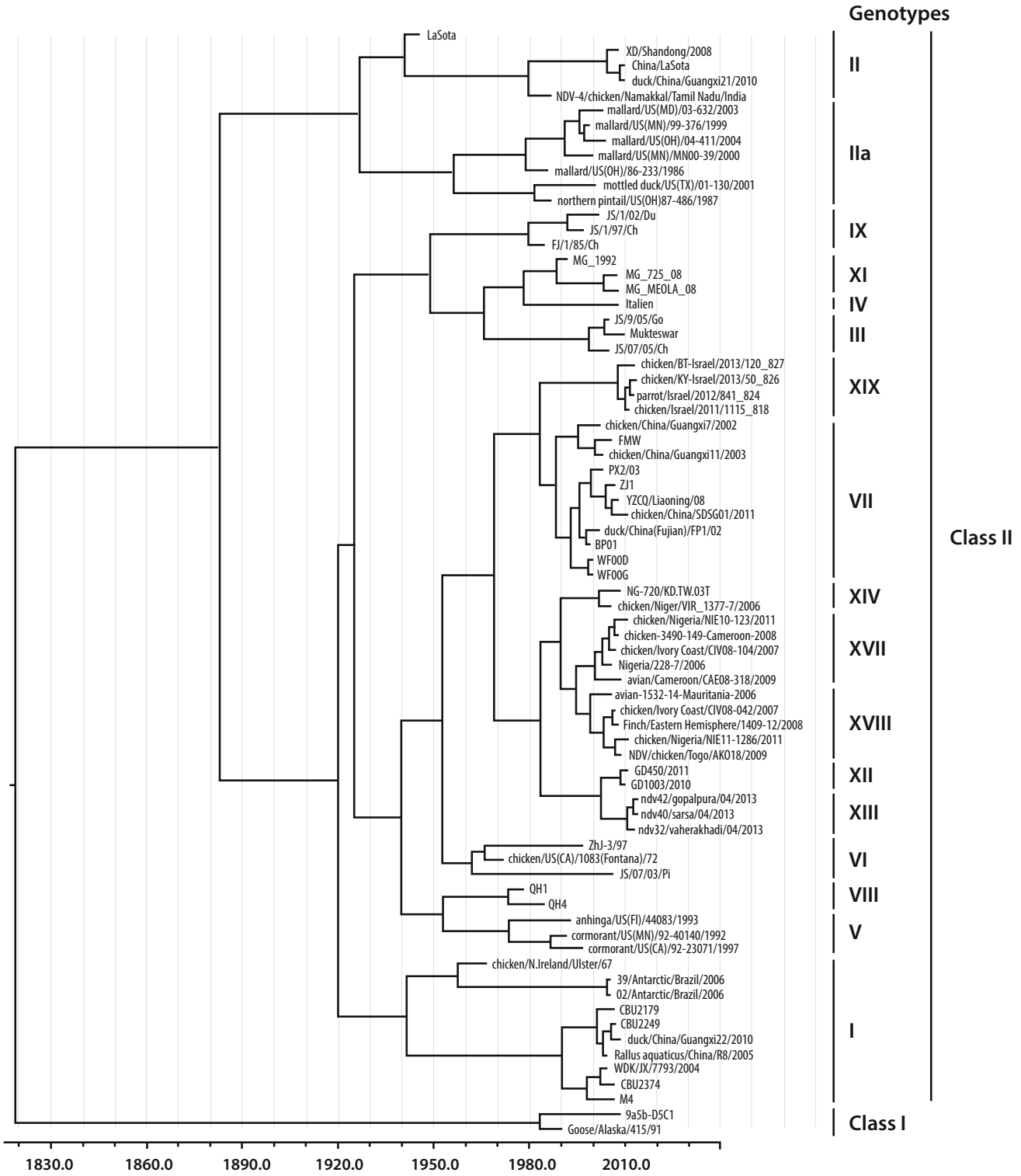


Fig. 2 Bayesian MCMC phylogenetic tree analysis of F genes of NDV strains. A maximum-credibility clade obtained using the HKY + Γ model, the Bayesian Skyline model and a relaxed clock (uncorrelated exponential) is shown. The tree is rooted to the MCRA. Years are indicated on the *x*-axis. Strains are shown by name and their genotypes are indicated on the right side of the figure. Strains isolated in Antarctica are shown by black arrows

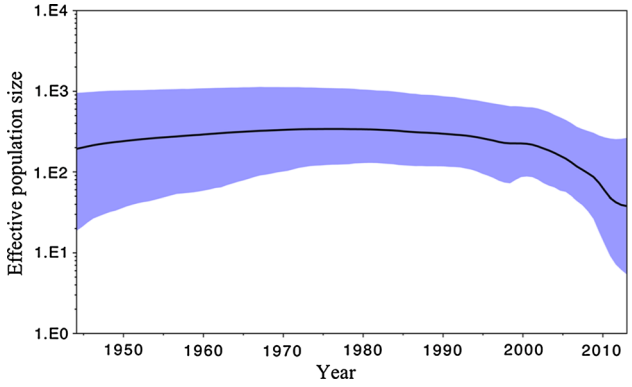


Fig. 3 Bayesian skyline plot depicting the population history of NDV strains. The *x*-axis indicates the year and the *y*-axis shows the product of effective population size and the generation length in years. The thick solid black line is the median estimate, and the blue area shows the 95 % highest probability density (HPD) values [38]

northwest Scotland between 1897 and 1898 was due to NDV [54].

In recent studies, Chong *et al.* have investigated the demographic history of NDV class II genotypes I-VII through Bayesian coalescent approaches, suggesting the maintenance of a constant effective population size until the late 1990s, when an abrupt decline with a posterior recovery (around 2000) was observed [21]. Roughly similar results were suggested by the analyses performed in the present study, which are summarized in a BSP supported by a narrow 95 % HPD (Fig. 3). Interestingly, the population dynamics observed in the last years of our analysis suggest a different behavior compared to what was reported previously, since a persistent continuous decrease in the effective population size was observed. This behavior can be explained by the larger number of class II genotypes considered in the present analysis (I-XIX), as distinct genotypes have been reported previously to exhibit different population dynamics [21]. Although the reasons for the observed decline are currently unknown, both climate change and avian influenza control measures have been suggested previously as possible factors [21]. More studies should be conducted in order to address these issues.

Considering that NDV seems to evolve rapidly towards higher virulence [55] and that several studies have reported not only increased pathogenicity but also outbreaks in vaccinated animals and increased host range [56, 57], it is becoming clear that it is important to conduct in-depth characterization of new strains isolated during the course of outbreaks worldwide to determine how these viruses are evolving. Additionally, studying viruses isolated from different wild birds and environments might contribute to our understanding of how NDV evolves and spreads around the world.

Acknowledgments We thank Instituto Antártico Uruguayo and Base Científica Antártica Artigas, Uruguay, for encouragement and support. We also thank Agencia Nacional de Investigación e Innovación (ANII) for support through project PE_ALI_2009_1_1603 and PEDECIBA, Uruguay.

References

1. Samal SK (2011) Newcastle disease and related avian paramyxoviruses. In: Samal SK (ed) The biology of paramyxoviruses. Caister Academic Press, Norfolk, pp 69–114
2. Lamb R, Parks G (2007) Paramyxoviridae: the viruses and their replication. In: Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE (eds) Fields virology. Lippicott Williams & Wilkins, Philadelphia, pp 1449–1496
3. Kim LM, King DJ, Curry PE et al (2007) Phylogenetic diversity among low-virulence Newcastle Disease Viruses from waterfowl and shorebirds and comparison of genotype distributions to those of poultry-origin isolates. J Virol 81:12641–12653
4. Alexander DJ (2003) Newcastle disease virus, other avian paramyxoviruses, and pneumovirus infections. In: Saif YM, Barnes HJ, Glisson JR, Fadly AM, McDougald LR, Swayne DE (eds) Disease of poultry, 11edn. Iowa State University Press, Ames, pp 63–87
5. Toyoda T, Gotoh B, Sakaguchi T, Kida H, Nagai Y (1988) Identification of amino acids relevant to three antigenic determinants on the fusion protein of Newcastle disease virus that are involved in fusion inhibition and neutralization. J Virol 62:4427–4430
6. Neyt C, Geliebter J, Slaoui M, Morales D, Meulemans G, Burny A (1989) Mutations located on both F1 and F2 subunits of the Newcastle Disease virus fusion protein confer resistance to neutralization with monoclonal antibodies. J Virol 63:952–954
7. Nagai Y, Hamaguchi M, Toyoda T (1989) Molecular biology of Newcastle disease virus. Prog Vet Microbiol Immunol 5:16–64
8. Samal S, Khattar S, Kumar S, Collins PL, Samal SK (2012) Coordinate deletion of N-glycans from the heptad repeats of the fusion F protein of Newcastle Disease virus yields a hyperfusogenic virus with increased replication, virulence, and immunogenicity. J Virol 86:2501–2511
9. Panda A, Huang Z, Elankumaran S, Rockemann DD, Samal SK (2004) Role of fusion protein cleavage site in the virulence of Newcastle disease virus. Microb Pathog 36:1–10
10. De Leeuw OS, Koch G, Hartog L, Ravenshorst N, Peeters BPH (2005) Virulence of Newcastle disease virus is determined by the cleavage site of the fusion protein and by both the stem region and globular head of the haemagglutinin–neuraminidase protein. J Gen Virol 86:1759–1769
11. Choi KS, Lee EK, Jeon WJ, Kwon JH (2010) Antigenic and immunogenic investigation of the virulence motif of the Newcastle disease virus fusion protein. J Vet Sci 11:205–211
12. Pedersen JC, Senne DA, Woolcock PR, Kinde H, King DJ, Wise MG, Panigrahy B, Seal BS (2004) Phylogenetic relationships among virulent Newcastle disease virus isolates from the 2002–2003 outbreak in California and other recent outbreaks in North America. J Clin Microbiol 42:2329–2334
13. Yusoff K, Nesbit M, McCartney H, Meulemans G, Alexander DJ, Collins MS, Emmerson PT, Samson AC (1989) Location of neutralizing epitopes on the fusion protein of Newcastle disease virus strain Beaudette C. J Gen Virol 70:3105–3109
14. Maminiaina OF, Gil P, Briand FX et al (2010) Newcastle Disease Virus in Madagascar: identification of an original genotype

possibly deriving from a died out ancestor of genotype IV. PLoS ONE 5:e13987
15. Mase M, Murayama K, Karino A, Inoue T (2010) Analysis of the fusion protein gene of Newcastle Disease viruses isolated in Japan. J Vet Med Sci 73:47–54
16. Czeglédi A, Ujvári D, Somogyi E, Wehmann E, Werner O, Lomniczi B (2006) Third genome size category of avian paramyxovirus serotype 1 (Newcastle disease virus) and evolutionary implications. Virus Res 120:36–48
17. Diel DG, da Silva LH, Liu H, Wang Z, Miller PJ, Afonso CL (2012) Genetic diversity of avian paramyxovirus type 1: proposal for a unified nomenclature and classification system of Newcastle disease virus genotypes. Infect Genet Evol 12:1770–1779
18. Fernandes CC, Varanib AM, Lemos EGM, de Miranda VFO, Silva KR, Fernando FS, Montassiera MFS, Montassiera HJ (2014) Molecular and phylogenetic characterization based on the complete genome of a virulent pathotype of Newcastle disease virus isolated in the 1970s in Brazil. Infect Genet Evol 26:160–167
19. Snoeck CJ, Owoade AA, Couacy-Hymann E, Alkali BR, Okwen MP, Adeyanju AT, Komoyo GF, Nakouné E, Le Faou A, Muller CP (2013) High genetic diversity of Newcastle disease virus in poultry in West and Central Africa: cocirculation of genotype and newly defined genotypes XVII and XVIII. J Clin Microbiol 51:2250–2260
20. Xiao S, Paldurai A, Nayak B, Mirande A, Collins PL, Samal SK (2013) Complete genome sequence of a highly virulent Newcastle disease virus currently circulating in Mexico. Genome Announc. doi:10.1128/genomeA.00177-12
21. Chong YL, Padhi A, Hudson PJ, Poss M (2010) The effect of vaccination on the evolution and population dynamics of avian paramyxovirus-1. PLoS Pathog 6:e1000872
22. Rui Z, Juan P, Jingliang S, Jixun Z, Xiaoting W, Shouping Z, Xiaojiao L, Guozhong Z (2010) Phylogenetic characterization of Newcastle disease virus isolated in the mainland of China during 2001–2009. Vet Microbiol 141:246–257
23. Zhang S, Wang X, Zhao C, Liu D, Hu Y, Zhao J, Zhang G (2011) Phylogenetic and pathotypical analysis of two virulent Newcastle disease viruses isolated from domestic ducks in China. PLoS ONE 6:e25000
24. Susta L, Miller PJ, Afonso CL, Brown CC (2011) Clinicopathological characterization in poultry of three strains of Newcastle disease virus isolated from recent outbreaks. Vet Pathol 48:349–360
25. Afonso CL, Miller PJ (2013) Newcastle disease: progress and gaps in the development of vaccines and diagnostic tools. Dev Biol 135:95–106
26. Lomniczi B, Wehmann E, Herczeg J et al (1998) Newcastle disease outbreaks in recent years in western Europe were caused by an old (VI) and a novel genotype (VII). Arch Virol 143:49–64
27. Herczeg J, Wehmann E, Bragg RR, Travassos-Dias PM, Hadjiev G, Werner O, Lomniczi B (1999) Two novel genetic groups (VIIb and VIII) responsible for recent Newcastle disease outbreaks in Southern Africa, one (VIIb) of which reached Southern Europe. Arch Virol 144:2087–2099
28. Abolnik CHR, Bisschop SP, Parker ME, Romito M, Viljoen GJ (2004) A phylogenetic study of South African Newcastle disease virus strains isolated between 1990 and 2002 suggests epidemiological origins in the Far East. Arch Virol 149:603–619
29. Yang CY, Shieh HK, Lin YL, Chang PC (1999) Newcastle disease virus isolated from recent outbreaks in Taiwan phylogenetically related to viruses (genotype VII) from recent outbreaks in western Europe. Avian Dis 43:125–130
30. Perozo F, Marcano R, Afonso CL (2012) Biological and phylogenetic characterization of a genotype VII Newcastle disease virus from Venezuela: efficacy of field vaccination. J Clin Microbiol 50:1204–1208
31. Diel DG, Susta L, Garcia SC, Killian ML, Brown C, Miller PJ, Afonso CL (2012) Complete genome and clinicopathological characterization of a virulent Newcastle disease virus isolate from South America. J Clin Microbiol 50:378–387
32. OIE (2011) World Animal Health Information Database (WAHID) Interface. http://web.oie.int
33. Thomazelli LM, Araujo J, Oliveira DB et al (2010) Newcastle disease virus in penguins from King George Island on the Antarctic region. Vet Microbiol 146:155–160
34. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5:113
35. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
36. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214
37. Delport W, Poon AF, Frost SD, Kosakovsky-Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457
38. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confindence. PLoS Biol 4:e88
39. Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian selection of continuous time Markov chain evolutionary models. Mol Biol Evol 18:1001–1013
40. Paldurai A, Kumar S, Nayak B, Samal SK (2010) Complete genome sequence of highly virulent neurotropic Newcastle disease virus strain Texas GB. Virus Genes 41:67–72
41. Seal BS (2004) Nucleotide and predicted amino acid sequence analysis of the fusion protein and hemagglutinin-neuraminidase protein genes among Newcastle disease virus isolates. Phylogenetic relationships among the Paramyxovirinae based on attachment glycoprotein sequences. Funct Integr Genomics 4:246–257
42. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185–1192
43. Alexander DJ, Manvell RJ, Collins MS, Brockman SJ, Westbury HA, Morgan I, Austin FJ (1989) Characterization of paramyxoviruses isolated from penguins in Antarctica and sub-Antarctica during 1976–1979. Arch Virol 109:135–143
44. Gould AR, Kattenbelt JA, Selleck P, Hansson E, Della-Porta A, Westbury HA (2001) Virulent Newcastle disease in Australia: molecular epidemiological analysis of viruses isolated prior to and during the outbreaks of 1998–2000. Virus Res 77:51–60
45. Kattenbelt JA, Stevens MP, Gould AR (2006) Sequence variation in the Newcastle disease virus genome. Virus Res 116:168–184
46. Alexander DJ (1998) Newcastle disease and other avian paramyxoviruses. A laboratory manual for the isolation and identification of avian pathogens. American Association of Avian Pathologists, Kennett Square, pp 156–163
47. Alexander DJ (1997) Newcastle disease and other avian Paramyxoviridae infections. In: Calnek BW (ed) Diseases of poultry. Mosby-Wolfe Iowa State University Press, Ames, pp 541–569
48. Werner O, Römer-Oberdörfer A, Köllner B, Manvell RJ, Alexander DJ (1999) Characterization of avian paramyxovirus type 1 strains isolated in Germany during 1992 to 1996. Avian Pathol 28:79–88
49. Alexander DJ, Senne DA (2008) Newcastle disease. In: Saif YM, Barnes HJ, Glisson JR, Fadly AM, McDougald LR, Swayne DE (eds) Diseases of poultry, 12th edn. Blackwell Publishing, Ames, pp 75–100
50. Wakamatsu N, King DJ, Seal BS, Peeters BP, Brown CC (2006) The effect on pathogenesis of Newcastle disease virus LaSota

strain from a mutation of the fusion cleavage site to a virulent sequence. Avian Dis 50:483–488
51. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BBH, Wolinsky S, Bhattacharya T (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789–1796
52. Zlateva KT, Lemey P, Moes E, Vandamme AM, Van Ranst M (2005) Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. J Virol 79:9157–9167
53. Allain JP, Dong Y, Vandamme AM, Moulton V, Salemi M (2000) Evolutionary rate and genetic drift of hepatitis C virus are not correlated with the host immune response: studies of infected donor–recipient clusters. J Virol 74:2541–2549
54. Macpherson LW (1956) Some observations on the epizootiology of New Castle Disease. Can J Comp Med Vet Sci 20:155–168
55. Miller PJ, Decanini EL, Afonso CL (2009) Newcastle disease: Evolution of genotypes and the related diagnostic challenges. Infect Genet Evol 10:26–35
56. Nakamura K, Ohtsu N, Nakamura T, Yamamoto Y, Yamada M, Mase M, Imai K (2008) Pathologic and immunohistochemical studies of Newcastle disease (ND) in broiler chickens vaccinated with ND: severe nonpurulent encephalitis and necrotizing pancreatitis. Vet Pathol 45:928–933
57. Wan H, Chen L, Wu L, Liu X (2004) Newcastle disease in geese: natural occurrence and experimental infection. Avian Pathol 33:216–221

Publicaciones realizadas durante el transcurso de esta tesis

Research Article

EVIDENCE OF INCREASING DIVERSIFICATION OF ZIKA VIRUS STRAINS ISOLATED IN THE AMERICAN CONTINENT†

Running Head: Zika virus evolution in the Americas

Fabián Aldunate[1], Fabiana Gámbaro[1], Alvaro Fajardo[1], Martín Soñora[1] and Juan Cristina[1*]

[1] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la Republica, Igua 4225, 11400 Montevideo, Uruguay.

***Correspondence:**

Juan Cristina,

Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la Republica, Igua 4225, 11400 Montevideo, Uruguay.

Phone: +598 2525 09 01,

FAX: +598 2525 08 95,

E-mail: cristina@cin.edu.uy

†This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/jmv.24910]

Additional Supporting Information may be found in the online version of this article.

Received 10 July 2017; Revised 28 July 2017; Accepted 1 August 2017

Journal of Medical Virology

DOI 10.1002/jmv.24910

Biochemical and Biophysical Research Communications xxx (2017) 1–7

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Host influence in the genomic composition of flaviviruses: A multivariate approach

Diego Simón [a], Alvaro Fajardo [b], Martín Sóñora [b], Adriana Delfraro [c], Héctor Musto [a, *]

[a] Laboratorio de Organización y Evolución del Genoma, Unidad de Genómica Evolutiva, Facultad de Ciencias (FC), Universidad de la República (UDELAR), Iguá 4225, Montevideo 11400, Uruguay
[b] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, FC, UDELAR, Uruguay
[c] Sección Virología, FC, UDELAR, Uruguay

ARTICLE INFO

Article history:
Received 16 May 2017
Received in revised form
9 June 2017
Accepted 15 June 2017
Available online xxx

Keywords:
Flavivirus
Base composition
Dinucleotides
Codon usage
Amino acids

ABSTRACT

Flaviviruses present substantial differences in their host range and transmissibility. We studied the evolution of base composition, dinucleotide biases, codon usage and amino acid frequencies in the genus *Flavivirus* within a phylogenetic framework by principal components analysis. There is a mutual interplay between the evolutionary history of flaviviruses and their respective vectors and/or hosts. Hosts associated to distinct phylogenetic groups may be driving flaviviruses at different pace and through various sequence landscapes, as can be seen for viruses associated with *Aedes* or *Culex* spp., although phylogenetic inertia cannot be ruled out. In some cases, viruses face even opposite forces. For instance, in tick-borne flaviviruses, while vertebrate hosts exert pressure to deplete their CpG, tick vectors drive them to exhibit GC-rich codons. Within a vertebrate environment, natural selection appears to be acting on the viral genome to overcome the immune system. On the other side, within an arthropod environment, mutational biases seem to be the dominant forces.

1. Introduction

The genus *Flavivirus* belongs to the family *Flaviviridae*, together with *Hepacivirus*, *Pegivirus* and *Pestivirus*. According to the International Committee of Virus Taxonomy, the genus comprises 53 species with wide global distribution, as well as an increasing number of unclassified or tentative species [1]. They are positive-sense single-stranded RNA viruses of about 11 kb, with a 5′ type I cap structure and lacking a poly(A) tail at the 3′ end. Their genome is translated in a single polyprotein which is cleaved in three structural proteins (C, prM and E) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) [2].

Despite the similarity in their genomic organization, there are substantial differences in the host range and transmissibility among them. Most known species are arboviruses, which are transmitted horizontally between hematophagous arthropods and susceptible vertebrate hosts, and are classified in mosquito-borne flaviviruses (MBFV) and tick-borne flaviviruses (TBFV). However, some species only replicate in bats or rodents with not-known vector associated to them (NKV). Furthermore, several species only infect mosquitoes, which are referred to as insect-specific flaviviruses (ISFV) [3–9].

The taxonomic relationship among flaviviruses has been extensively investigated through different approaches, originally based on antigenic cross-reactivity in neutralization, complement fixation and hemagglutination inhibition assays [10,11]. Lately, phylogenetic reconstructions based on nucleotide and amino acid sequences allowed a deeper understanding of the diversity of the genus. Several methodological approaches were followed to analyze different genes and complete coding regions [3–8,12–14]. As a result of these efforts it was found that the general pattern of the inferred phylogenetic relationships correlates with the main epidemiological aspects as host range, vectors and related diseases. Nevertheless, the comparison of different analyses evidences phylogenetic incongruences that difficult a proper definition of the taxonomic relationships.

Another approach to understand both the evolution and the phylogenetic relationships, is to analyze compositional properties of each virus, such as base composition, dinucleotide biases, codon

* Corresponding author.
E-mail addresses: dsimon@fcien.edu.uy (D. Simón), afajardo@cin.edu.uy (A. Fajardo), msonora@cin.edu.uy (M. Sóñora), adelfraro@gmail.com (A. Delfraro), hmusto@gmail.com (H. Musto).

http://dx.doi.org/10.1016/j.bbrc.2017.06.088
0006-291X/

Title: A DETAILED COMPARATIVE ANALYSIS OF CODON USAGE BIAS IN ZIKA VIRUS.

Author: Juan Cristina Alvaro Fajardo Martín Soñora Gonzalo Moratorio Héctor Musto

PII: S0168-1702(16)30118-6
DOI: http://dx.doi.org/doi:10.1016/j.virusres.2016.06.022
Reference: VIRUS 96925

To appear in: *Virus Research*

Received date: 16-2-2016
Revised date: 8-6-2016
Accepted date: 14-6-2016

Please cite this article as: Cristina, Juan, Fajardo, Alvaro, Soñora, Martín, Moratorio, Gonzalo, Musto, Héctor, A DETAILED COMPARATIVE ANALYSIS OF CODON USAGE BIAS IN ZIKA VIRUS.Virus Research http://dx.doi.org/10.1016/j.virusres.2016.06.022

A DETAILED COMPARATIVE ANALYSIS OF CODON USAGE BIAS IN ZIKA VIRUS.

Juan Cristina[a]*, Alvaro Fajardo[a], Martín Soñora[a], Gonzalo Moratorio[a,b] and Héctor Musto[c].

[a]Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Igua 4225, 11400 Montevideo, Uruguay.

[b] Viral Populations and Pathogenesis laboratory. Institut Pasteur, CNRS UMR 3569, Paris, France.

[c]Laboratorio de Organización y Evolución del Genoma, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay.

* Corresponding author: Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Iguá 4225, 11400 Montevideo, Uruguay. Phone: +5982 525 09 01, FAX: +5982 525 08 95, e-mail: cristina@cin.edu.uy

Virus Research 223 (2016) 140–146

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Naturally occurring NS3 resistance-associated variants in hepatitis C virus genotype 1: Their relevance for developing countries

Natalia Echeverría[a,1], Gabriela Betancour[a,1], Fabiana Gámbaro[a], Nelia Hernández[b], Pablo López[b], Daniela Chiodi[b], Adriana Sánchez[b], Susana Boschi[c], Alvaro Fajardo[a], Martín Sóñora[a], Gonzalo Moratorio[a], Juan Cristina[a], Pilar Moreno[a,*]

[a] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, 11400 Montevideo, Uruguay
[b] Clínica de Gastroenterología, Hospital de Clínicas, Facultad de Medicina, Universidad de la República, 11600 Montevideo, Uruguay
[c] Laboratorio de Biología Molecular, Asociación Española, Palmar 1465, Montevideo, Uruguay, Uruguay

ARTICLE INFO

Article history:
Received 6 July 2016
Received in revised form 15 July 2016
Accepted 18 July 2016
Available online 19 July 2016

Keywords:
HCV variability
DAAs
Resistance-associated variants

ABSTRACT

Hepatitis C virus (HCV) is a major cause of global morbidity and mortality, with an estimated 130–150 million infected individuals worldwide. HCV is a leading cause of chronic liver diseases including cirrhosis and hepatocellular carcinoma. Current treatment options in developing countries involve pegylated interferon-α and ribavirin as dual therapy or in combination with one or more direct-acting antiviral agents (DAA). The emergence of resistance-associated variants (RAVs) after treatment reveals the great variability of this virus leading to a great difficulty in developing effective antiviral strategies. Baseline RAVs detected in DAA treatment-naïve HCV-infected patients could be of great importance for clinical management and outcome prediction. Although the frequency of naturally occurring HCV NS3 protease inhibitor mutations has been addressed in many countries, there are only a few reports on their prevalence in South America. In this study, we investigated the presence of RAVs in the HCV NS3 serine protease region by analysing a cohort of Uruguayan patients with chronic hepatitis C who had not been treated with any DAAs and compare them with the results found for other South American countries. The results of these studies revealed that naturally occurring mutations conferring resistance to NS3 inhibitors exist in a substantial proportion of Uruguayan treatment-naïve patients infected with HCV genotype 1 enrolled in these studies. The identification of these baseline RAVs could be of great importance for patients' management and outcome prediction in developing countries.

1. Introduction

Hepatitis C virus (HCV) is a significant human pathogen affecting nearly 3% of the world's population, and is a leading cause of chronic liver diseases including cirrhosis and hepatocellular carcinoma. Infections with HCV have become a major cause of liver cancer and one of the most common indications for liver transplantation (Hoofnagle, 2002; Martin et al., 2013; Pawlotsky, 2003; Simmonds, 2004; World Health Organization, 2015).

HCV belongs to the family *Flaviviridae* and has a single stranded positive sense RNA genome that is 9.6 kb in length. This genome contains a single open-reading frame and encodes a unique polyprotein that is processed to yield ten structural (core, E1 and E2) and non-structural (p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B) proteins (Scheel and Rice, 2013).

The high error rate of the viral RNA-dependent RNA-polymerase and the pressure exerted by the host immune system, have driven the evolution of HCV towards the development of a global diversity that reveals the existence of seven genetic lineages (genotypes 1–7) and more than 67 subtypes (Smith et al., 2014). Subtypes 1a, 1b and 3a are widely distributed and account for the vast majority of infections in Western countries including the South American region (World Health Organization, 2015).

HCV NS3 protein is responsible for processing the non-structural region of the viral polyprotein. NS3 is a bifunctional protein with an amino-terminal domain exhibiting a zinc-dependent serine protease activity, and a carboxyl-terminal one with helicase activity (Bartenschlager and Lohmann, 2000).

* Corresponding author at: Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Mataojo 2055, 11400 Montevideo, Uruguay.
E-mail address: pmoreno@cin.edu.uy (P. Moreno).
[1] both authors contributed equally to this work.

http://dx.doi.org/10.1016/j.virusres.2016.07.008
0168-1702/

Journal of Medical Virology

Bayesian Coalescent Inference Reveals High Evolutionary Rates and Diversification of Zika Virus Populations

Alvaro Fajardo,[1] Martín Soñora,[1] Pilar Moreno,[1] Gonzalo Moratorio,[1,2] and Juan Cristina[1]*

[1]*Molecular Virology Laboratory, CIN, Facultad de Ciencias, Universidad de la Republica, Montevideo, Uruguay*
[2]*Viral Populations and Pathogenesis laboratory. Institut Pasteur, Paris, France*

Zika virus (ZIKV) is a member of the family *Flaviviridae*. In 2015, ZIKV triggered an epidemic in Brazil and spread across Latin America. By May of 2016, the World Health Organization warns over spread of ZIKV beyond this region. Detailed studies on the mode of evolution of ZIKV strains are extremely important for our understanding of the emergence and spread of ZIKV populations. In order to gain insight into these matters, a Bayesian coalescent Markov Chain Monte Carlo analysis of complete genome sequences of recently isolated ZIKV strains was performed. The results of these studies revealed a mean rate of evolution of 1.20×10^{-3} nucleotide substitutions per site per year (s/s/y) for ZIKV strains enrolled in this study. Several variants isolated in China are grouped together with all strains isolated in Latin America. Another genetic group composed exclusively by Chinese strains were also observed, suggesting the co-circulation of different genetic lineages in China. These findings indicate a high level of diversification of ZIKV populations. Strains isolated from microcephaly cases do not share amino acid substitutions, suggesting that other factors besides viral genetic differences may play a role for the proposed pathogenesis caused by ZIKV infection. ***J. Med. Virol.***

KEY WORDS: Zika; coalescent; bayesian; evolution; microcephaly

INTRODUCTION

Zika virus (ZIKV) is a flavivirus, whose natural transmission cycle involves mosquitoes vectors from the *Aedes* (*Ae*.) genus, while humans are occasional hosts [Hayes, 2009]. Clinical manifestations of disease range from asymptomatic cases to fever, headache, malaise, and cutaneous rash. ZIKV is transmitted primarily by *Ae. aegypti* mosquitoes [Hayes, 2009]. Other mosquitoes species, like *Ae. albopictus*, can transmit the virus. Both mosquitoes species are found throughout the Americas, where also transmit Dengue and Chikungunya viruses [Hennessey et al., 2016].

ZIKV genome consists of a single-stranded positive sense RNA molecule of 10,794 nt in length. It has two non-coding regions at the 5′ and 3′ end of the genome. This genome encode for a single long open reading frame encoding a polyprotein that is cleaved into capsid (C), precursor of membrane (prM), envelope (E), and seven non-structural proteins (NS) [Kuno and Chang, 2007].

ZIKV was isolated for the first time in 1947, from the blood of a sentinel Rhesus monkey stationed in the Zika forest, Uganda [Dick et al., 1952]. Although, ZIKV enzootic activity was reported in diverse countries of Africa and Asia, few human cases were reported until 2007, when an epidemic took place in Micronesia [Duffy et al., 2009]. A large ZIKV outbreak took place in French Polynesia during 2013–2014 and then spread to other Pacific Islands [Musso, 2015]. In early 2015, a ZIKV epidemic outbreak took place in Brazil, currently estimated at 440,000–1,300,000 cases [Campos et al., 2015]. By January 20th, 2016, ZIKV locally transmitted cases were reported from Puerto Rico and 19 other

Grant sponsor: Agencia Nacional de Investigación e Innovación (ANII); Grant sponsor: PEDECIBA; Grant sponsor: Comisión Sectorial de Investigación Científica (CSIC); Grant sponsor: Grupos I+D program

Conflict of interest: None.

*Correspondence to: Juan Cristina, Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la Republica, Igua 4225, 11400 Montevideo, Uruguay. E-mail: cristina@cin.edu.uy

Accepted 2 June 2016

DOI 10.1002/jmv.24596
Published online in Wiley Online Library (wileyonlinelibrary.com).

Virus Research 197 (2015) 1–7

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Phylogenetic analysis of the neuraminidase gene of pandemic H1N1 influenza A virus circulating in the South American region

Victoria Comas[a], Gonzalo Moratorio[a,b], Martin Soñora[a], Natalia Goñi[c], Silvana Pereyra[d], Silvana Ifran[d], Pilar Moreno[a], Juan Cristina[a,*]

[a] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Igua 4225, 11400 Montevideo, Uruguay
[b] Viral Populations and Pathogenesis laboratory. Institut Pasteur, CNRS UMR 3569, Paris, France
[c] Centro Nacional de Referencia de Influenza, Departamento de Laboratorios de Salud Pública, Ministerio de Salud Pública, Alfredo Navarro 3051 acceso norte, 11200 Montevideo, Uruguay
[d] Laboratorio de Biología Molecular, Asociación Española Primera de Socorros Mutuos, Br. Artigas 1515, 11300 Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form
30 September 2014
Accepted 8 November 2014
Available online 3 December 2014

Keywords:
Pandemic
Influenza A virus
Evolution
Neuraminidase

ABSTRACT

Molecular characterization of circulating influenza A viruses (IAV) in all regions of the world is essential to detect mutations potentially involved in increased virulence, anti-viral resistance and immune escape. In order to gain insight into these matters, a phylogenetic analysis of the neuraminidase (NA) gene of 146 pandemic H1N1 (H1N1pdm) influenza A virus strains isolated in Argentina, Brazil, Chile, Paraguay, Peru and Uruguay from 2009 to 2013 was performed. Comparison of vaccine strain A/California/7/2009 included in the influenza vaccine recommended for the Southern hemisphere from 2010 through 2013 influenza seasons and strains isolated in South America revealed several amino acid substitutions. Mapping of these substitutions revealed that most of them are located at the surface of the protein and do not interfere with the active site. 3.4% of the strains enrolled in these studies carried the H275Y substitution that confers resistance to oseltamivir. Strains isolated in South America differ from vaccine in two predicted B-cell epitope regions present at positions 102–103 and 351–352 of the NA protein. Moreover, vaccine and strains isolated in Paraguay differ also in an epitope present at position 229. These differences among strains isolated in South America and vaccine strain suggests that these epitopes may not be present in strains isolated in this region. A potential new N-linked glycosylation site was observed in the NA protein of an H1N1pdm IAV strain isolated in Brazil. The results of these studies revealed several genetic and antigenic differences in the NA of H1N1pdm IAV among vaccine and strains circulating in South America. All these findings contribute to our understanding of the course of genetic and antigenic evolution of H1N1pdm IAV populations circulating in the South American region and, consequently, contribute to the study and selection of future and more appropriate vaccines and anti-viral drugs.

1. Introduction

Influenza A virus (IAV) is a member of the family *Orthomyxoviridae* and contains eight segments of a single-stranded RNA genome with negative polarity (Neumann et al., 2004). IAV causes 300,000–500,000 deaths worldwide each year, and in pandemic years, this number can increase to 1 million (in 1957–1958) or as high as 50 million, as was seen in 1918–1919 (Nguyen-Van-Tam and Hampson, 2003). IAV exhibits a rapid evolution and complex molecular dynamics patterns due to its wide host range, high substitutions rates and rapid replication (Holmes, 2010). Hemagglutinin (HA) and neuraminidase (NA) are the two envelope glycoproteins that are responsible for attaching the virions to the host receptors, determining pathogenicity, and releasing newly produced viral particles (Li et al., 2011). Amino acid substitutions on these glycoproteins can modify virus replication and impact over the potential spread in the human population (Pizzorno et al., 2012; Abed et al., 2006). The NA is also playing an important role as a target of the single calls of available anti-influenza drugs, e.g. NA inhibitors.

The first influenza pandemic of this century was declared in April of 2009, with the emergence of a novel H1N1 IAV strain (H1N1pdm) in Mexico and the USA (CDC, 2009; WHO, 2009a,b,c). This virus rapidly spread to the South American region, where it was

* Corresponding author. Tel.: +598 2525 09 01; fax: +598 2525 08 95.
E-mail address: cristina@cin.edu.uy (J. Cristina).

http://dx.doi.org/10.1016/j.virusres.2014.11.008
0168-1702/