Análisis de variabilidad de secuencia *in vivo* de RNAs circulares con ribozimas

Autor: Marc Bañuls Tornero
Máster universitario en Bioinformática y bioestadística UOC-UB
Área del trabajo final: Bioinformática clínica

Nombre Consultor/a: Dr. Joan Maynou Fernández
Nombre Profesor/a responsable de la asignatura: Marc Maceira Buch, Javier Luis Cánovas Izquierdo
Nombre del Tutor externo: Dr. Marcos De la Peña
Fecha de entrega: 24/06/2020

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de variabilidad de secuencia in vivo de RNAs circulares con ribozimas*
Nombre del autor:	*Marc Bañuls Tornero*
Nombre del consultor/a:	*Dr. Joan Maynou Fernández*
Nombre del PRA:	*Marc Maceira Buch, Javier Luis Cánovas Izquierdo*
Fecha de entrega:	06/2020
Titulación:	*Máster universitario en Bioinformática y bioestadística UOC-UB*
Área del Trabajo Final:	*AreaTFM: 1(Bioinformática Clínica)*
Idioma del trabajo:	*Castellano*
Palabras clave:	*"circRNA", "retrozyme", "LTR retrotransposons"*

Resumen del Trabajo (máximo 250 palabras):

Recientemente se han descubierto retrotransposones LTR no autónomos que se expresan como RNAs circulares con ribozimas HHR, bautizados como retrozimas. En este estudio se ha realizado un análisis detallado de la variabilidad natural de un retrozima de *Fragaria ananassa* bajo diversas condiciones.

Concretamente se estudió el grado de variabilidad *in vivo* en el retrozima elegido, la cantidad de variantes de secuencia que aparecen en el tiempo, si el hecho de que el RNA del retrozima circularice o no afecta a la variabilidad del retrozima, y qué efectos tiene la variabilidad observada en su estructura secundaria. Para ello se diseñaron distintas construcciones utilizadas para la transformación

genética de dos modelos de plantas, para posteriormente purificar y extraer el RNA y realizar una secuenciación mediante Sanger.

Las secuencias obtenidas fueron ordenadas y utilizadas para realizar un análisis de variantes mediante scripts inhouse en que se utilizan Biopython, bwa-mem, BCFtools y SAMtools, mientras que las estructuras secundarias se obtuvieron mediante el servicio web RNAalifold.

En los experimentos realizados se confirmó la aparición de un mismo grupo de variantes estables que indicarían algún tipo de ventaja adaptativa o edición específica de los circRNAs que las contienen. Además, se ha demostrado que la variabilidad en retrozimas mutados que dan lugar a RNA lineales es mínima comparada con la observada en los retrozimas con circRNAs. La posible función específica de estas variantes estables queda aún por determinar en futuros experimentos, ya que ni siquiera parecieron influir en motivos conservados ni en las estructuras secundarias obtenidas.

Abstract (in English, 250 words or less):

Non-autonomous LTR retrotransposons that are expressed as circular RNAs with HHR ribozymes, named as retrozymes, have recently been discovered. In this study we carried out a detailed analysis of the natural variability of a *Fragaria ananassa* retrozyme under different conditions.

Specifically, we studied the degree of *in vivo* variability in the chosen retrozyme, the number of variants that appear over time, whether the circularization of the RNA of the retrozyme affects or not the variability of the retrozyme, and the effect of the variability in its secondary structure. Different constructs were designed for the genetic transformation of two plant models to purify and extract the RNA and perform Sanger sequencing.

The sequences obtained were ordered and used to perform an analysis of variants using inhouse scripts with Biopython, bwa-mem, BCFtools and SAMtools, while the secondary structures were obtained using the RNAalifold web service.

The experiments carried out confirmed the presence of the same group of stable variants, which would indicate some type of adaptive advantage or specific

edition of the circRNAs. Furthermore, it has been shown that the variability in mutated retrozymes that encode linear RNAs is minimal compared to the variability of circRNAs. The specific function of these stable variants remains to be determined in future experiments, since it does not seem to influence in conserved motifs or its secondary structure.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del trabajo

1.1.1 RNAs con actividad catalítica

Desde hace casi un siglo se ha venido estudiando los distintos catalizadores biológicos, siendo en 1926 cuando James Sumner purificó la primera enzima, la "ureasa". Gracias a esta purificación se encontró que la enzima estaba compuesta por un polipéptido o proteína, confirmando con la purificación de centenares de enzimas en los siguientes años que todos los catalizadores biológicos eran de naturaleza proteica.

Sin embargo, este dogma se rompió cuando en 1982 se encontraron en el protozoo *Tetrahymena thermophila* moléculas de RNA que realizaban las acciones de corte y ligación de un intrón del RNA ribosomal de manera autocatalítica [1]. Posteriormente aparecieron más estudios que indicaban la existencia de moléculas de RNA con actividad catalítica como la RNAsa P [2]**.** Desde ese momento, estos RNAs catalíticos empezaron a referenciarse como ribozimas.

Más recientemente, se ha terminado de asentar la importancia que tienen las ribozimas en biología, con la demostración de que tanto el propio ribosoma que cataliza el enlace peptídico [3]**,** como el espliceosoma que procesa el mRNA eucariótico [4] son ribozimas**.**

La mayoría de las distintas clases de ribozimas estudiadas, tanto naturales como artificiales (obtenidas en el laboratorio), catalizan el corte y/o ligación del enlace fosfodiéster del RNA. El mecanismo de corte y ligación de las ribozimas es muy parecido al mecanismo de corte de RNA catalizado por proteínas ribonucleasas, donde el oxígeno 2' de la ribosa ataca al fosfato 3' en una reacción de tipo S_N-2*,* induciendo así la formación de un fosforano bipiramidal, próximo al estado de transición. Una rotura simultánea del enlace al oxígeno 5' produce un fosfato cíclico 2'-3' y un extremo 5'-OH. En la reacción inversa (ligación), el nucleófilo 5'-OH ataca al fosfato cíclico para formar el enlace fosfodiéster [5] (Figura 1). En las ribozimas, este mecanismo es posible gracias a la complementariedad de

bases y la estructura terciaria de la secuencia, la cual permite la formación de un núcleo activo capaz de catalizar el corte en zonas específicas, como se ha observado previamente en los mecanismos de catálisis de las enzimas proteicas [6].

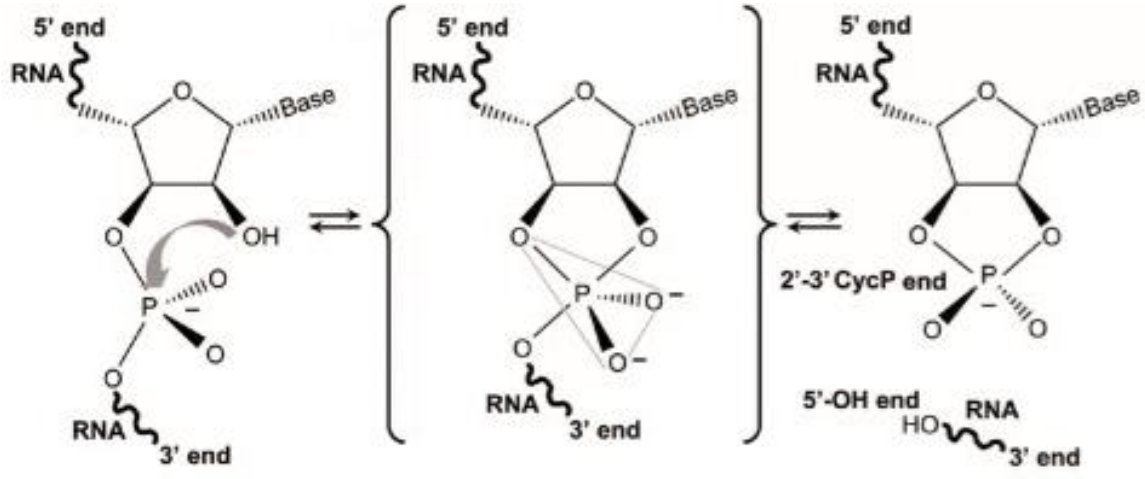

Figura 1. Reacción de transesterificación del RNA catalizado por ribozimas de autocorte. Modificado de [16].

La mayoría de las ribozimas naturales que catalizan el corte del RNA pertenecen a la familia de las ribozimas pequeñas de autocorte [7]. Éstas tienen un tamaño relativamente pequeño, rondando entre los 50 y 150 nucleótidos, y su origen evolutivo por el momento es desconocido. Aun así, las propiedades características de éstas han sido de gran importancia a la hora de ahondar en los conocimientos que se tenían del RNA, tales como su versatilidad biológica, estructura y bioquímica [8]. El aumento en los conocimientos que se tienen del RNA ha permitido asentar la hipótesis del mundo prebiótico basado en el RNA, donde las moléculas precursoras de la vida en la tierra serían capaces de contener tanto una información genética básica como la posibilidad de desempeñar por sí mismas funciones catalíticas necesarias para la vida, como es el caso de la replicación [9].

1.1.2 Ribozimas pequeñas de autocorte

Como indica su propio nombre, estas ribozimas catalizan su propio autocorte entre dos nucleótidos específicos de su secuencia. Hasta la fecha se han

descrito hasta 9 clases con este tipo de actividad catalítica: ribozimas de cabeza de martillo (Hammerhead ribozymes o HHR), ribozimas de horquilla (hairpin o HPR), ribozimas del agente delta del virus de la hepatitis humana (HDV), ribozimas de satélite varkud (VS), ribozimas de la glucosamina-6-fosfato sintasa (glmS), ribozimas twister (Tw), ribozimas twister sister (TwS), ribozimas hatchet y ribozimas pistol.

Esta familia de ribozimas se ha considerado generalmente como rarezas biológicas que fundamentalmente se encontraban en agentes subvirales con genomas mínimos de RNA circular de plantas (ribozimas HHR y HPR) o humanos (ribozima del HDV o Hepatitis Delta Virus). Recientemente se encontró la presencia de multitud de estos motivos ribozimáticos a lo largo de toda la escala biológica, desde bacteriófagos a humanos [10], lo que además ha permitido el descubrimiento de hasta 5 nuevas clases de pequeñas ribozimas, y destacando especialmente la ubiquidad de las ribozimas HHR.

1.1.2.1 Ribozima de cabeza de martillo (HHR)

Hace más de 30 años, la ribozima HHR fue la primera de las ribozimas pequeñas de autocorte que se identificó. Esta ribozima se descubrió originalmente en los genomas de RNA circular de satélites virales [7] y viroides [11] de plantas, donde participa en el procesado de los transcritos multiméricos procedentes de la replicación por círculo rodante de estos agentes infecciosos. Posteriormente se encontraron HHRs en el DNA repetitivo de genomas de algunas especies de animales, como tritones [12], tremátodos [13] y grillos de cueva [14]. No fue hasta el año 2010 que se describió la presencia de estos motivos HHR a lo largo de toda la escala biológica, pasando por procariotas y eucariotas, e incluso en el genoma del ser humano [15], lo que convirtió a la ribozima HHR en probablemente la ribozima de autocorte más ampliamente distribuida por la biosfera [16].

La estructura secundaria de las ribozimas HHR destaca por su parecido con la cabeza de un martillo (Figura 2a), la cual está formada a partir de tres hélices, dos de las cuales están cerradas mediante bucles terminales [17] (Figura 2a). Estas hélices rodean un centro catalítico compuesto por 15 nucleótidos esenciales para realizar la actividad catalítica. Según la disposición de las hélices

respecto a los extremos 3’ y 5’ del RNA, la ribozima puede presentar distintas topologías, existiendo tres tipos de HHR posibles llamados de tipo I, tipo II y tipo III (Figura2b)**.**

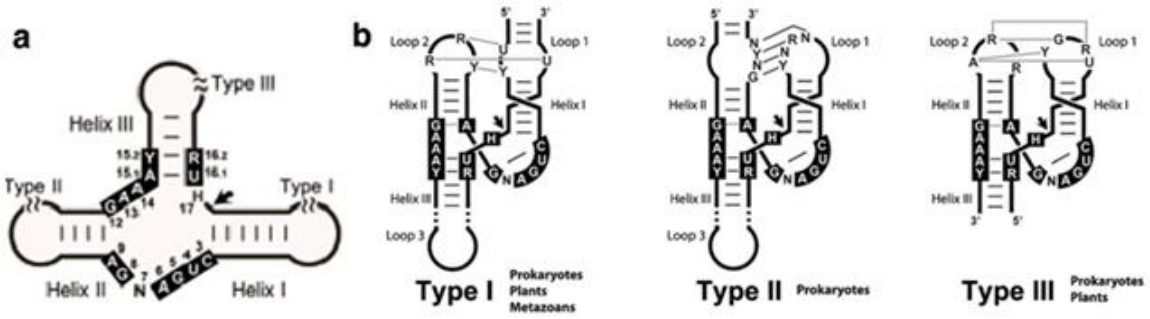

Figura 2. a: Representación de la estructura secundaria de la ribozima HHR. b: Representación de la estructura secundaria y terciaria de los tipos de HHR. Modificado de [16, 34].

Las primeras cristalizaciones de una ribozima HHR [18, 19] se realizaron a partir de ribozimas artificiales mínimas (éstas no contenían los bucles 1 y 2), donde se pudo observar el centro activo en un estado “pre-catalítico”. Este estado necesitaba una reordenación importante de los nucleótidos para que el centro activo fuera capaz de catalizar el autocorte en el residuo diana. Más adelante, en 2003, se describió la existencia conservada de interacciones terciarias entre las hélices I y II de las HHRs naturales que estabilizaban el sitio activo, solucionando así el problema encontrado en el centro activo en estado pre-catalítico. Esta estabilización del centro catalítico permitía mejorar significativamente el autocorte del RNA en un medio con bajas concentraciones de magnesio [20, 21] .Finalmente el papel de estas interacciones terciarias en el reordenamiento del centro catalítico se confirmó con la obtención de la estructura cristalográfica de una ribozima natural completa [22]. Dicho estudio estructural permitió finalmente observar que el centro catalítico seguía un mecanismo de catálisis tipo ácido-base.

Hasta la fecha, no se conoce con detalle la función concreta de la mayoría de las ribozimas HHR detectadas, a pesar de encontrarse en los genomas de multitud de especies (tanto procariotas como eucariotas). Se han observado HHRs en distintos transposones eucarióticos, ya sean autónomos (como los elementos Penélope o PLEs) [23] o no autónomos (como en los retrozimas de plantas y

animales), [24, 25]. Además, se han encontrado HHRs altamente conservadas en regiones no codificantes de genes de vertebrados superiores (amniotas) tanto dentro de intrones [15]**,** como en la región 3' UTR (UnTranslated Region) de diversos genes [26]**.** La presencia de las HHRs en estas localizaciones genómicas podría indicar que tienen una función reguladora que se ha adquirido a partir de la domesticación de ribozimas de retroelementos procedentes de metazoos inferiores [16, 17].

1.1.3 Retrotransposones

Los elementos transponibles (TE) son secuencias de DNA que poseen la capacidad intrínseca de cambiar de posición dentro del genoma en el que se encuentre. Para ello utilizan distintos mecanismos de duplicación, escisión e inserción.

Los TE se encuentran en la práctica totalidad de los organismos conocidos, desde organismos procariotas hasta el hombre. Además, un elevado porcentaje de DNA en genomas eucariotas están compuestos de TEs, viéndose este porcentaje incrementado cuando se trata de especies vegetales, con casos extremos como el genoma del maíz compuesto en un 80% por TEs [27, 28]**.** Además existen distintas clases de TEs, dependiendo principalmente de si se trasponen utilizando una molécula de RNA intermediaria mediante retrotransposición e integración como cDNA, (Retrotransposones o TEs de clase I) o si se transponen directamente a partir de una molécula de DNA (Transposones o TEs de clase II). Los retrotransposones contienen regiones de polipurinas (PPT) y motivos de unión al cebador (PBS), y además suelen utilizar una retrotranscriptasa (RT) y una endonucleasa/integrasa (EN/IN) que han sido codificadas por el mismo retrotransposón en el caso de retrotransposones autónomos, o por otro elemento transponible en el caso de los retrotransposones no autónomos.

Los TE autónomos tienen en su secuencia interna genes que codifican para proteínas enzimáticas (como sería el gen pol) y también estructurales (como los genes gag y env). Los retrotransposones autónomos están clasificados en cuatro grupos principales: retrotransposones LTR (Long Terminal Repeats),

retrotransposones no-LTR, DIRS (Dictyostelium Intermediate Repeat Sequence) y PLE (Penelope-Like Elements) [29].

Los retrotransposones PLE se encuentran principalmente en amebas, hongos y animales, excluyendo mamíferos. Estos retrotransposones contienen una pauta de lectura abierta (Open Reading Frame o ORF) que codifica para RT y EN, y producen zonas diana de duplicación (Target Site Duplication o TSD) de tamaño variable. Algunos PLE además pueden tener secuencias similares a LTRs (PLTR) pero orientados de manera directa o inversa en la secuencia [30]. Recientemente se ha descrito la presencia conservada de motivos HHR mínimos en los LTRs de PLEs con una función aún desconocida [23].

Los retrotransposones DIRS se encuentran en la mayoría de los organismos incluyendo plantas [31], y engloban a todos los retrotransposones que contienen tirosina recombinasa (YR) en vez de IN. Además, contiene Proteinasa aspártica (AP) y RNAsa H (RH), mientras que sus extremos no producen TSD y tienen estructura de repeticiones invertidas (ITR).

Dentro de los retrotransposones no-LTR destacan los LINE y SINE (Long y Short INterspersed Element respectivamente). Los retrotransposones LINE contienen entre una y dos ORF, y un promotor de la RNA polimerasa II para promover la transcripción del retrotransposón, teniendo además su RT y EN en el mismo ORF [32]. En cambio los retrotransposones SINE son elementos no autónomos compuestos por diversos RNAs, como tRNAs o rRNAs y un promotor de la RNA polimerasa III [33].

Los retrotransposones LTR son los más abundantes, especialmente en los genomas de las plantas, y se llaman así principalmente debido a que estos retrotransposones presentan en sus extremos 3’ y 5’ unas regiones denominadas como repeticiones largas terminales (Long-Terminal-Repeats o LTR). Si los retrotransposones LTR son autónomos, pueden codificar un antígeno específico de grupo, una proteasa, una integrasa, una transcriptasa inversa y una ribonucleasa H (Figura 3).

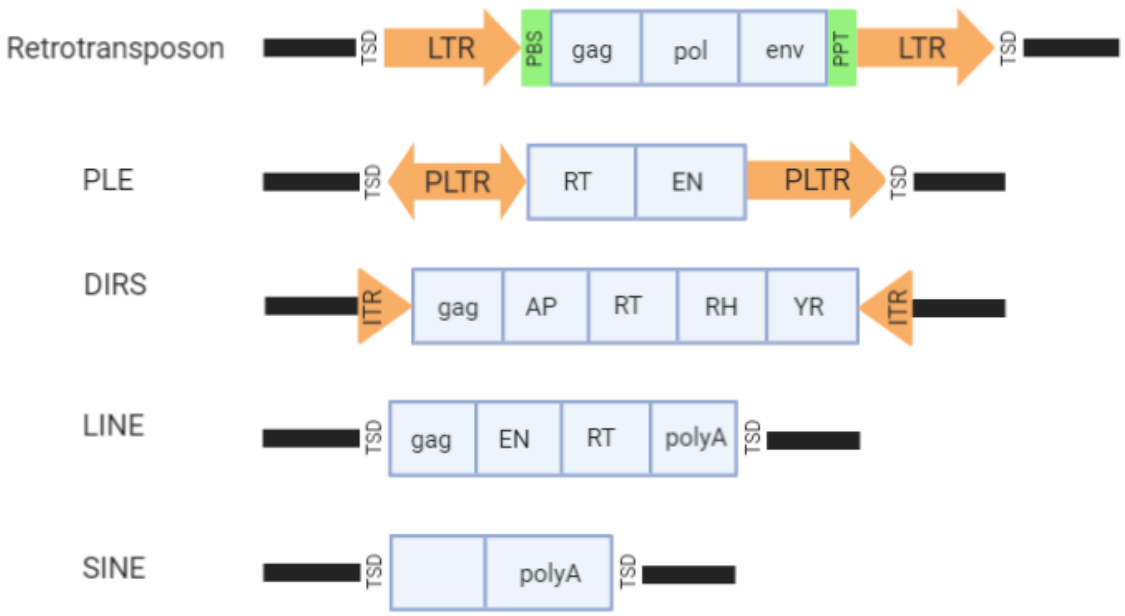

Figura 3. Estructura general de los distintos retrotransposones comentados. Las regiones codificantes se muestran en azul y los extremos LTR en naranja. Se muestra en las flechas la orientación de la región. Modificado de [33].

1.1.4 Retrozimas

Recientemente se han encontrado en genomas de distintas plantas una nueva familia de retrotransposones tipo LTR no autónomos que contienen ribozimas HHR [24], siendo llamados por ello *retrozimas* (*retro*transposones con *ribozimas*) Cada ribozima HHR se encuentra en una de las dos regiones LTR, por lo que se disponen en los extremos 3' y 5' del retrotransposón. Las HHRs se encuentran separadas por secuencias no codificantes de un tamaño entre 600 y 1000 pares de bases con una alta variabilidad (incluso entre especies no muy alejadas filogenéticamente).

Los retrozimas contienen regiones PPT y PBS características de los retrotransposones LTR (Figura 3), y en particular, éstas son muy similares a los elementos de la familia Ty3-gypsy. Esto puede indicar que el mecanismo de retrotransposición de estos retrozimas sea similar a los retrotransposones LTR Ty3-gypsy, de los que muy probablemente utiliza su maquinaria para completar el proceso de transposición. Sorprendentemente, se ha encontrado que los retrozimas se transcriben activamente en la mayoría de los tejidos de la planta, acumulándose en las células en forma de RNAs circulares (circRNAs) y lineales

del tamaño comprendido por las dos HHR del retrozima (~700 nt) [24]. Muy probablemente, estos circRNAs son los intermediarios de transposición que utiliza el retrozima para movilizarse a nuevas regiones del genoma de la planta [34] (figura 4).

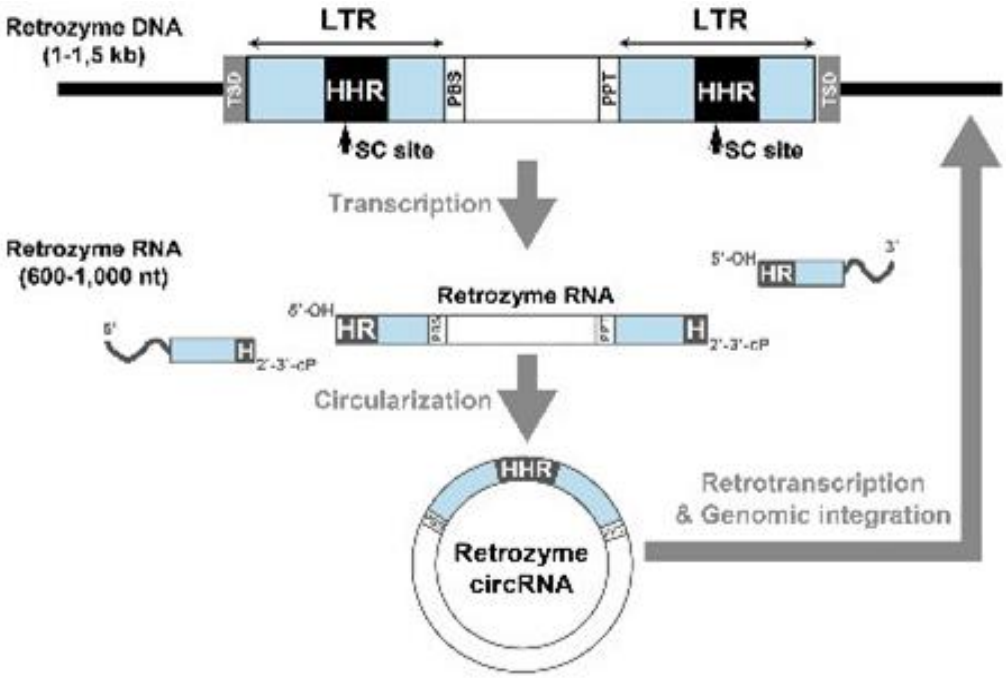

Figura 4. Representación de un retrozima genómico y su ciclo de transcripción, circularización y retrotranscripción. Los TSDs se muestran en gris y los LTR en azul. Los puntos de auto-corte (SC) se indican con flechas. Obtenido de [34].

1.1.5 Justificación del trabajo

En un estudio del grupo de investigación en el que se han realizado las prácticas y este TFM, se han descrito hasta 90 ribozimas HHR en el genoma de *Fragaria ananassa* (comúnmente conocido como fresón), donde se ha observado que los transcritos de los retrozimas se acumulan significativamente en los distintos tejidos de la planta [24]. Además, estos transcritos parecen tener una alta variabilidad en la secuencia, hipotetizándose así que los circRNAs codificados por los retrozimas podrían estar siguiendo una intensa edición a nivel de RNA, o incluso replicándose mediante un mecanismo de círculo rodante de RNA a RNA, lo que implica la introducción habitual de errores. Para profundizar en el conocimiento de estos fenómenos, se llevó a cabo la secuenciación por Sanger

de circRNAs de retrozimas de *Fragaria ananassa* obtenidos de distintos experimentos. Básicamente, se expresó de manera transitoria o estable una copia de un retrozima genómico de *F. ananassa* en dos plantas modelo como son *Nicotiana benthamiana* y *Arabidopsis thaliana*.

El objetivo original de este trabajo era llevar a cabo el estudio de variabilidad del retrozima de *Fragaria ananassa* por secuenciación masiva NGS (Illumina). Desgraciadamente, la irrupción en marzo de 2020 de la pandemia provocada por el virus SARS-CoV-2 dejó esta aproximación experimental detenida en el servicio de Secuenciación de la Universidad de Valencia. Se continuó en cualquier caso este trabajo a partir de las secuencias obtenidas por Sanger, sobre las que se realizó un análisis bioinformático mediante el estudio de los SNPs (Single Nucleotide Polymorphism) y su estructura secundaria, que permitirá la obtención de un mayor conocimiento del comportamiento de estos retrozimas en condiciones *in vivo*.

1.2 Objetivos del trabajo

1.2.1 Analizar la variabilidad *in vivo* de circRNAs respecto a una secuencia de referencia a partir de poblaciones de RNAs circulares obtenidas por transgénesis en dos modelos experimentales de plantas.

Uno de los objetivos principales de este trabajo consiste en analizar la variabilidad de las secuencias de retrozimas en su forma de circRNAs obtenidos a partir de distintas muestras biológicas. Para ello se empleó la transformación genética transitoria o estable de diversas construcciones en dos organismos vegetales distintos, de los que posteriormente se amplificaron los RNAs de interés por RT-PCR y secuenciados mediante Sanger.

Se debe tener en cuenta que para realizar el análisis de la variabilidad de las secuencias inicialmente se analizarán las variantes naturales de los retrozimas de *F. ananassa* que se encuentran depositadas en la base de datos de NCBI [35]. Estas secuencias permiten comparar las variantes naturales de las obtenidas en los distintos experimentos realizados en este trabajo. Por ello, se encuentra como objetivo la obtención de una secuencia consenso con las variantes naturales indicadas (para ello se recomienda el uso de la simbología

IUPAC). Con esta secuencia se pretenden observar las posiciones en las que aparecen las variantes naturales y de qué tipo son, para así compararlas con las variantes obtenidas a partir de la secuencia parental introducida en los distintos experimentos.
Cuando se complete la secuencia consenso de referencia, el siguiente objetivo consiste en realizar un análisis de detección de variantes en cada experimento.

1.2.2 Comparar la variabilidad de los retrozimas respecto al tiempo de transformación.

Se ha de llevar a cabo un análisis similar al del objetivo anterior, pero el objetivo en este caso varía levemente esperándose encontrar la variabilidad entre las secuencias en cada tiempo de la transformación con la secuencia de referencia.

1.2.3 Comparar la variabilidad de secuencia entre un retrozima circular y un retrozima lineal expresados *in vivo*

Se quiere observar el efecto que existe cuando se expresa en un organismo dado un retrozima que no tiene funcionales sus motivos HHR. Para cumplir este objetivo se transformará por un lado el retrozima parental de referencia, y por otro lado, un retrozima mutado en sus dominios ribozimáticos que impedirá el correcto procesado del RNA lineal, y consecuentemente, su circularización. La construcción de retrozima con HHRs mutadas se llamará lineal.
Posteriormente se pretende obtener los amplicones secuenciados de cada experimento y realizar un análisis de la variabilidad para cada uno de ellos. De esta manera se obtendrán los SNPs que demuestran la variabilidad de las secuencias. El último objetivo consiste en comparar la variabilidad entre los dos retrozimas a partir de las variables detectadas en cada experimento.

1.2.4 Observar los efectos de la variabilidad de secuencia en la estructura secundaria de los retrozimas estudiados

Se pretende obtener la estructura secundaria de los circRNA secuenciados en cada experimento. Posteriormente, se quieren estudiar los efectos de las variaciones encontradas en estas estructuras secundarias.

1.3 Metodología

Para conseguir llevar a cabo los objetivos comentados anteriormente, existen varias estrategias posibles. Por ello, se describirán a continuación algunas de dichas estrategias, y cuál se ha escogido finalmente para este trabajo.

Existen varios métodos de obtención de una secuencia consenso en formato IUPAC. Principalmente se ha pensado en realizar un alineamiento de secuencias múltiple (MSA) de los distintos amplicones de referencia (las secuencias con variantes naturales procedentes del NCBI), y para ello existen varios algoritmos que pueden servir a este propósito.

Los algoritmos más populares actualmente para el alineamiento de secuencias múltiples son el MUSCLE 3.8.31 [36], ClustalOmega 1.2.1 [37] y MAFFT 7.450 [38] entre otros. Estos algoritmos tienen cada uno ventajas e inconvenientes respecto a los otros, pero se ha observado que generalmente el algoritmo MAFFT es el que tiene una precisión y velocidad de procesamiento medio más alto de entre estos algoritmos [39]. Aun así, no parece que las diferencias entre los algoritmos respecto a precisión sean elevadas, por lo que en lo que concierne a este estudio, teniendo en cuenta la baja cantidad de muestras con un tamaño pequeño-mediano (~700 pb), da lugar a un alineamiento de las secuencias similar. Para comprobarlo se ha realizado el alineamiento utilizando los tres algoritmos mencionados y obteniendo el mismo resultado, por lo que el algoritmo escogido no afecta a la precisión del alineamiento. Por ello se escoge para este trabajo el algoritmo ClustalOmega (abreviado a ClustalO), principalmente por ser computacionalmente más rápido [39].

Para realizar el análisis de variantes se ha pensado en utilizar una de las dos herramientas más populares actualmente: SAMtools/BCFtools [40, 41] o GATK [42].

SAMtools 1.7-1 contiene un conjunto de herramientas para procesar archivos SAM (Sequence Alignment Map) y BAM (Binary Alignment Map) obtenidos previamente a partir del paquete bwa-mem 0.7.17 [43] y BCFtools 1.7-2 permite detectar y manipular variantes en formato VCF (Variant Call Format) y BCF (Binary variant Call Format) a partir de estos archivos SAM/BAM procesados. Ambas herramientas se usan en combinación para poder realizar un pipeline efectivo para detectar variantes a partir de una secuencia o genoma de referencia y las secuencias preparadas para alinear.

Otro paquete que se puede utilizar es un set de herramientas de análisis de genómas o GATK 4.1.7.0 procedente del BroadInstitute [44]. Este paquete combina herramientas propias y de otras fuentes para su uso más sencillo y compacto en la terminal. Los protocolos que distribuye esta institución se encuentran actualizados [45], aunque principalmente están enfocados a análisis de secuencias contra un genoma de referencia, especializándose concretamente en el genoma humano. Por ello, las herramientas están pensadas para ser utilizadas en secuenciaciones obtenidas mediante NGS, por lo que se espera un genoma de referencia y millones de fragmentos de secuencias en formato fastq (formato fasta con puntuaciones de calidad de la secuencia o Quality Score [46]) para alinear estos fragmentos y mapearlos en el genoma de referencia.

Dicho esto, se ha encontrado que al comparar la precisión a la hora de detectar variantes con parámetros estándar la detección de variantes mediante las herramientas de GATK tiene una precisión media mayor que cuando se usan las herramientas de SAMtools/BCFtools [47]. Aun así, aunque el uso de una u otra herramienta puede dar lugar a diferencias en el número de variantes detectadas, la modificación de los parámetros que estos proporcionan son lo que más pueden afectar a la hora de detectar eficientemente las variantes. Debido a que para la detección de variantes se están utilizando datos procedentes de secuenciación mediante Sanger cuya alta sensibilidad y especificidad de las secuencias es más elevada respecto a las obtenidas mediante NGS [48], se encontrará un ratio similar de detección de variantes en ambas herramientas. Además, cabe tener en cuenta que el número de secuencias obtenido mediante

Sanger es bajo comparado con la secuenciación NGS, por lo que esto junto a la alta especificidad y sensibilidad de la secuenciación mediante Sanger, se espera un muy bajo ratio de detecciones falsas positivas.

Como las secuencias obtenidas en los experimentos de este trabajo se encuentran en formato fasta, se han encontrado problemas de formato a la hora de intentar utilizar las herramientas de GATK (debido a los Quality Scores mencionados anteriormente). Como se ha comentado que en este trabajo la capacidad de detección de variantes será similar independientemente de si se usa SAMtools/BCFtools o GATK. Por ello se encuentra más conveniente el uso de las herramientas contenidas en SAMtools/BCFtools a la hora de realizar el pipeline de detección de variantes debido a la mayor facilidad a la hora de tratar los archivos.

Para la obtención de las estructuras secundarias se utiliza el servidor ViennaRNA [49], concretamente el servicio web RNAalifold desde la página web del Instituto de Química Teórica de la Universidad de Viena [49, 50]. En esta página web se pueden predecir automáticamente las estructuras secundarias consenso a partir de la secuencia que se le introduzca. Además permite interpretar la secuencia como circular, por lo que la estructura secundaria final obtenida tiene aún mayor fiabilidad.

1.4 Planificación del trabajo

El primer punto a realizar en el trabajo es la adquisición de conocimientos sobre los distintos métodos disponibles para poder realizar los posteriores análisis. Para ello, se ha investigado inicialmente qué herramientas son las óptimas para las tareas a realizar en este trabajo. Concretamente, las herramientas principales que se van a utilizar para los análisis son paquetes de herramientas que se utilizan desde la terminal de Linux: bwa-mem 0.7.17, SAMtools 1.7-1 y BCFtools 1.7-2. Para la visualización de los resultados se han utilizado UGENE [51]**,** y IGV [52].

1.4.1 Analizar la variabilidad de circRNAs respecto a una secuencia de referencia en los experimentos realizados

Primeramente se realiza un alineamiento utilizando el algoritmo ClustalOmega en la aplicación de escritorio UGENE con las secuencias depositadas en la página web del NCBI que contienen las variantes naturales (Figura 5). Posteriormente se obtiene en el mismo software la secuencia consenso con simbología IUPAC.

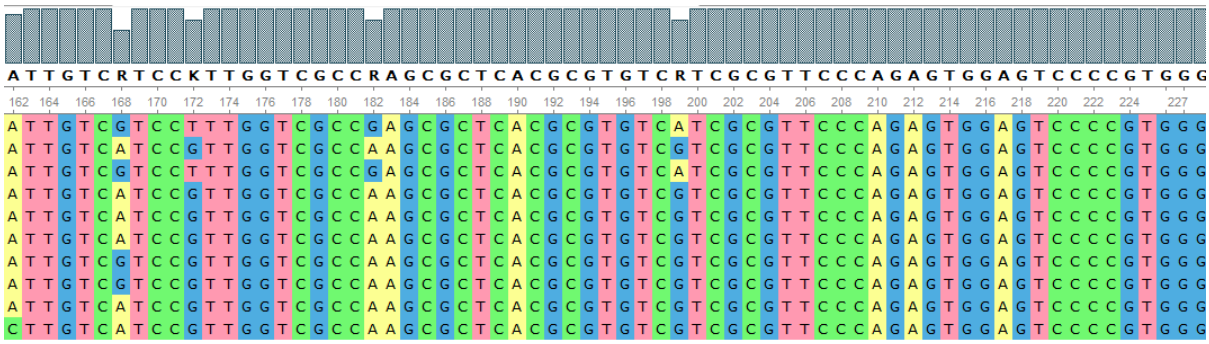

Figura 5. Visualización de las secuencias de variantes naturales con UGENE, junto a la secuencia consenso en la parte superior de las secuencias.

Las muestras secuenciadas mediante Sanger se ordenan con la posición 1 en su punto de corte de la HHR utilizando un script en Python personalizado (el código se encuentra en el anexo A).
Se crea un script bash (código en el anexo B) para que de forma automática se realice:

- Alineamiento de los amplicones de cada experimento respecto a la secuencia parental utilizando la herramienta bwa-mem en la terminal.

- Tratamiento de alineamientos y creación de los archivos necesarios para el seguimiento del análisis y comprobaciones del correcto funcionamiento del protocolo, utilizando para ello el paquete SAMtools.

- Identificación de las variantes más frecuentes mediante el uso del paquete BCFools.

1.4.2 Comparar la variabilidad entre un retrozima circular y un retrozima lineal

La primera tarea de este objetivo consiste en el filtraje manual de las secuencias y su posterior separado según la secuencia de referencia (lineal o circular) utilizada en el experimento. Para ello se debe observar manualmente el identificador en cada secuencia y separar manualmente los dos tipos de secuencias (las que se ha utilizado un retrozima circular o lineal). Para realizar este filtraje se utiliza UGENE.

Seguidamente se ordenan las secuencias y los retrozimas de referencia de cada experimento (tanto el lineal como el circular) por el punto de corte como su posición 1. En este caso también se utiliza el script de Python mencionado anteriormente (el punto de corte de las secuencias no cambia).

El alineamiento y su tratamiento junto con la detección de variantes se realiza de la misma manera que en el objetivo anterior, pero con las secuencias del experimento correspondiente.

Las secuencias de cada experimento se comparan entre ellas y sus secuencias de referencia. Además se realiza una comparación entre las variantes obtenidas en los experimentos con las variantes naturales depositadas en el NCBI. Para ello se pueden alinear las secuencias de referencia con la secuencia consenso en formato IUPAC (que contiene las variantes depositadas en el NCBI) o se puede observar la posición de cada variante en cada experimento respecto a la secuencia consenso. De una manera u otra se observa si las variantes obtenidas en cada experimento concuerdan o no con las variantes naturales.

1.4.3 Observar los efectos de la variabilidad de secuencia en la estructura secundaria de los retrozimas estudiados

Para observar los efectos de la variabilidad encontrada en cada experimento de las secuencias en su estructura secundaria, se usa el servicio web RNAalifold para predecir las estructuras secundarias consenso a partir de las alineaciones de cada experimento realizado anteriormente. Estas alineaciones se obtienen a partir del uso del algoritmo ClustalO desde la herramienta UGENE. Aunque también se podría obtener el mismo resultado desde la página web de EBI [53], se ha utilizado UGENE porque permite visualizar más fácilmente el resultado de

la alineación. De esta manera se pueden realizar modificaciones de los parámetros de alineación o se pueden visualizar posibles errores ajenos a ésta (secuencias a alinear no ordenadas correctamente por ejemplo).

2. Resto de capítulos

2.1 Experimentos realizados para la obtención de secuencias

El grupo de investigación ha realizado varios experimentos para posteriormente realizar un análisis de la variabilidad obtenida a partir de los amplicones completos secuenciados mediante Sanger en cada experimento.

En principio este trabajo iba a realizarse a partir de secuencias obtenidas mediante secuenciación NGS, pero se han sufrido retrasos en las secuenciaciones debido al SARS-CoV-2. Por ello se ha pensado en realizar los análisis en muestras secuenciadas mediante Sanger de otros experimentos realizados con anterioridad.

En todos los experimentos se obtienen construcciones utilizando un retrozima de la planta de *Fragaria ananassa* llamado a partir de ahora FaRtz.

En el primer experimento llamado 1.Fa_35S_Nb se ha realizado una construcción que contiene un retrozima de la planta de *Fragaria ananassa* (llamado a partir de ahora FaRtz como ya se ha dicho anteriormente) bajo el control de expresión del promotor fuerte 35S en un vector adecuado para agroinfiltración. Esta construcción se utiliza para transformar de forma transitoria plantas de *Nicotiana benthamiana*.

Al cabo de 6 días post-agroinfiltración de la construcción en las plantas, se extraen, clonan y purifican los circRNA producidos por el retrozima. Posteriormente se secuencian los circRNA mediante secuenciación Sanger (Figura 6a).

En el segundo experimento llamado 2.Fa_Nos_Nb se ha obtenido una construcción con el promotor débil Nos que contiene el retrozima FaRtz. Esta construcción también se utiliza para transformar de forma transitoria la planta *Nicotiana benthamiana*. En este experimento se han extraído clonado y

purificado los circRNA producidos por el retrozima en distintos días post-agroinfiltración, para ser secuenciados posteriormente. Concretamente, las extracciones de circRNA se han llevado a cabo a los 2, 4, 6 y 8 post-agroinfiltración (Figura 6b).

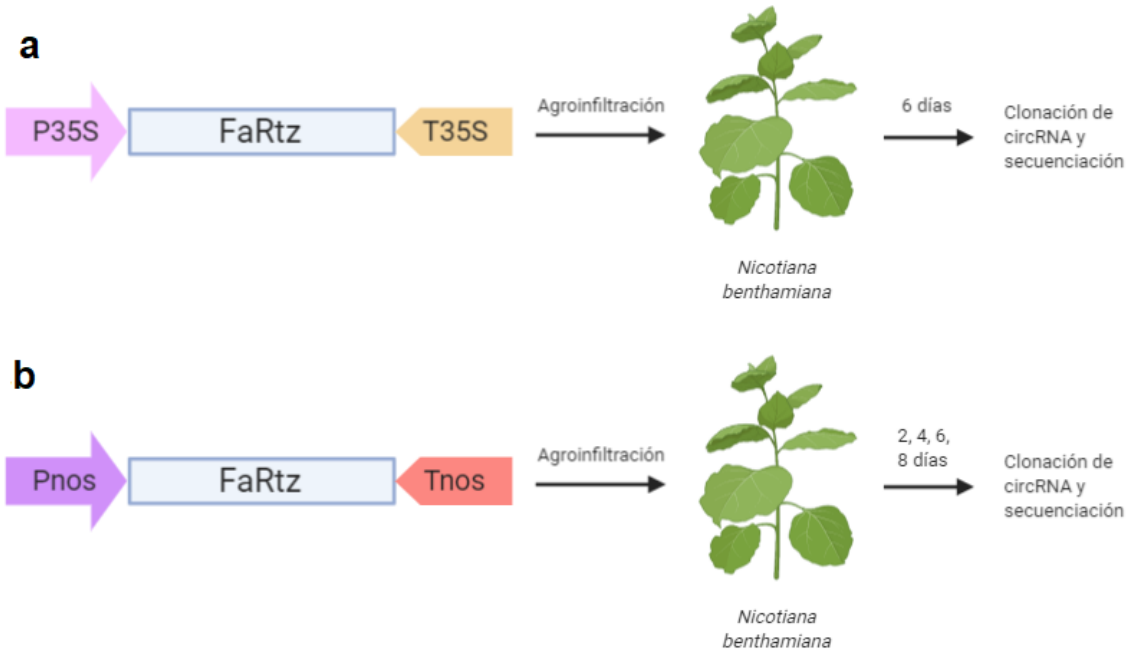

Figura 6. Representación esquemática del proceso realizado en el experimento 1.Fa_35S_Nb (a) y en el experimento 2.Fa_Nos_Nb (b).

El tercer experimento llamado 3.Fa_Nos_At es significativamente diferente a los dos anteriores, ya que se realizan transformaciones estables de la planta *Arabidopsis thaliana* utilizando dos construcciones distintas. Una construcción es la misma que la ya utilizada anteriormente en el experimento 2.Fa_Nos_Nb, en el que se utiliza Nos como promotor y se espera que el FaRtz exprese de manera estable circRNAs (Figura 7a). La otra construcción en cambio tiene como promotor también Nos pero el FaRtz utilizado ha sido mutado para que sus regiones HHR no sean funcionales (llamado FaRtzNull). De esta manera se espera que la planta transformada con FaRtzNull exprese de manera estable RNAs lineales (éstos no circularizan debido a la inactividad de las regiones HHR) (Figura 7b).

Finalmente se extraen, clonan y purifican los circRNAs y RNAs (expresados por FaRtz y FaRtzNull respectivamente) de las plantas transgénicas estables para su posterior secuenciación.

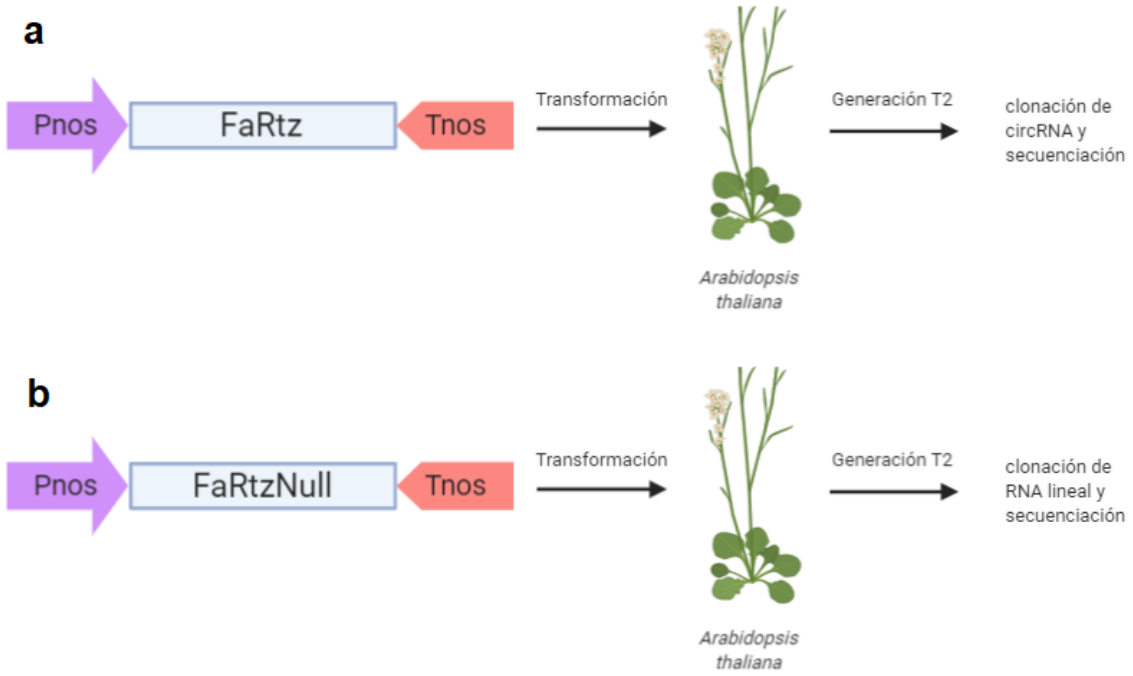

Figura 7. Representación esquemática del proceso realizado en el experimento 3.Fa_Nos_At utilizando el retrozima que expresa circRNAs (a) y el que expresa RNAs lineales (b).

2.2 Preparación de las secuencias

Según el experimento en el que se ha realizado el análisis de variabilidad de secuencias, los primeros pasos difieren entre ellos, ya que cada experimento debe tratarse de una manera determinada para obtener unos resultados finales satisfactorios.

Como se ha comentado anteriormente, en cada experimento se han realizado las construcciones utilizando unos promotores concretos. Por lo tanto, la posición de inicio y fin de la secuencia variará según los promotores utilizados. Como estas secuencias provienen de circRNAs, a la hora de realizar el alineamiento en su forma lineal, los puntos de inicio y fin de secuencia pueden dar lugar a alineamientos erróneos. Por ello, se ha pensado en realizar una ordenación de todas las secuencias utilizando el punto de corte en el HHR como primera posición de la secuencia. Para ordenar las secuencias se ha diseñado un script inhouse con Python (ver anexo A).

Para el diseño de ese script se ha utilizado el paquete BioPython 1.77 [54] para facilitar la lectura y tratamiento de las secuencias, utilizando sus distintas funciones para detectar más fácilmente las secuencias fasta.
Para poder utilizar el script, se debe ejecutar desde la terminal introduciendo:
python3 rearrange_sequences.py

Al ejecutar el script, se pide al usuario que introduzca secuencialmente (el script va pidiendo que se introduzcan los distintos valores) el directorio donde se encuentra el archivo fasta que hay que reordenar, el nombre del nuevo archivo, y el patrón por el que se quiere empezar la secuencia.
Con los datos principales introducidos, el script crea un nuevo archivo con el nombre asignado donde se van a introducir las secuencias conforme se reordenan. Con la función “SeqIO.parse” se identifican las secuencias como fasta y de manera automática detecta y separa cada secuencia para realizar la acción que se desee. Posteriormente se utiliza la función “seq.find” con el patrón dado por el usuario para que se utilice esta secuencia como patrón en las siguientes líneas de script.
Finalmente se escriben en el nuevo archivo las secuencias ordenadas en formato fasta. Para ello se introduce para cada nueva secuencia el símbolo “>” (para marcar el inicio de una nueva secuencia como se indica en el formato fasta), la propia secuencia desde el patrón de ésta hasta su última posición, y posteriormente la propia secuencia desde la primera posición hasta el patrón, acabando con un cambio de línea (para que la siguiente secuencia empiece en una nueva línea).
Este código en resumen recorta la secuencia por el patrón indicado y ubica desde la parte del patrón hasta el final como inicio de secuencia, y seguidamente se introduce la parte de la secuencia que se encuentra antes del patrón. De esta manera se obtiene la secuencia empezando por el patrón deseado reordenando satisfactoriamente la secuencia.
El script realizado se utiliza en cada grupo de secuencias de cada experimento utilizando como patrón la posición del punto de corte. Esta posición tiene el patrón “GGGGTCT”, el cual es único en las secuencias. De esta manera se han obtenido los archivos fasta preparados para los siguientes pasos.

Para obtener la secuencia parental con las variantes naturales en formato IUPAC se utiliza la aplicación UGENE para leer las secuencias en formato fasta. Cuando las secuencias están cargadas en la aplicación, se selecciona la opción “Align” y “ClustalO”**.** Este alineamiento se realiza con los parámetros estándar debido a las pocas mutaciones existentes en las secuencias, por lo que es improbable que uno u otro parámetro modifiquen significativamente el alineamiento.

Después de realizar el alineamiento se comprueba que las secuencias están alineadas correctamente. Si es así, se indica en la aplicación que se guarde la secuencia consenso con las variantes naturales en formato IUPAC. Finalmente se copia esta secuencia, para así poder utilizarla en los posteriores análisis.

En el experimento 1.Fa_35S_Nb debido a que las secuencias estaban dando problemas, se tuvieron que secuenciar las dos hebras en la mayoría de los casos. Por ello un primer paso consiste en el filtraje de las hebras que no estén completas o tengan errores graves en su secuencia. Para hacer esto se ha pensado en cargar el archivo con las secuencias fasta en el UGENE y comparar cada pareja de secuencias (se han secuenciado la secuencia directa y la secuencia reversa, que se ha convertido a la secuencia complementaria para igualar ambas secuencias). La secuencia de cada “pareja” que se encuentre incompleta o tenga más errores se descarta manualmente. Debido a los problemas de lectura que ha causado guardar las secuencias resultantes en formato fasta, la mejor opción encontrada ha sido hacer una copia del archivo fasta original y eliminar manualmente las secuencias que se han visualizado previamente en UGENE.

En el experimento 2.Fa_Nos_Nb se realiza una separación previa al análisis según el día de extracción de los amplicones secuenciados. Esto se realiza manualmente, ya que ya están todos correctamente nombrados. Entonces se ha creado un archivo para cada día de extracción distinto, para así poder facilitar su estudio por separado.

En el experimento 3.Fa_Nos_At, previamente a los análisis, se han localizado qué secuencias provienen del uso de cada construcción, para ver si éstas podían circularizar o no. Este proceso se ha realizado manualmente, ya que cada secuencia tiene en el nombre de qué construcción proviene. Por lo tanto, se han organizado en distintos archivos fasta las secuencias procedentes de la secuencia FaRtz y FaRtzNull por separado.

Las secuencias obtenidas se ordenan desde el punto de corte como posición inicial utilizando el script de Python indicado previamente. Posteriormente se ha empezado a realizar el análisis de la variabilidad de las secuencias.

2.3 Obtención de variantes

Para ejecutar todas las funciones para obtener las variantes de manera automatizada, se ha creado un script bash (ver anexo B). Este script se ejecuta desde la terminal con el comando:
bash variants_analysis.sh

Al ejecutar el script se pide al usuario indicar el directorio donde se encuentran los archivos, la secuencia o genoma de referencia y las secuencias con las que se realiza el análisis de variantes. El desglose de los comandos ejecutados por el script es:

Para empezar el protocolo de análisis de variabilidad primero se debe crear un archivo con el índice de la secuencia de referencia. Para ello se utiliza la herramienta bwa:
bwa index fartz_parental_corte.fasta

Ahora que la secuencia de referencia tiene un índice, se alinean las secuencias a analizar utilizando la secuencia de referencia y el algoritmo bwa-mem. A partir de esta operación se obtiene un archivo SAM, el cual contiene el alineamiento con las secuencias donde concuerdan (o no) todas las posiciones marcadas con un asterisco. Debido a la poca cantidad de secuencias en cada experimento y la alta fiabilidad de las secuenciaciones mediante Sanger no debería ser necesario aplicar parámetros adicionales a los que ya están aplicados por la herramienta.
bwa mem fartz_parental_corte.fasta exp20_r.fasta > exp20_r.sam

Como paso intermedio, se requiere la conversión del archivo SAM a formato binario, llamado BAM. Para ello se utiliza la función view de la herramienta SAMtools:
samtools view -b exp20_r.sam > exp20_r.bam

Otro paso intermedio consiste en ordenar las secuencias por su posición respecto a la secuencia de referencia. En el caso de que todas las secuencias empiecen en la misma o una posición similar (como es el caso de este trabajo) igualmente se recomienda su uso rutinario para poder observar posibles errores en el alineamiento:
samtools sort -o exp20_r_s.bam exp20_r.bam

Con los datos del experimento en formato BAM, ya se puede empezar la detección de variantes respecto a la secuencia parental. Para ello se utiliza el paquete BCFtools, concretamente la herramienta mpileup.
bcftools mpileup -O b -o exp20_raw.bcf -f fartz_parental_corte.fasta exp20_r_s.bam

La herramienta mpileup analiza la información del archivo BAM y realiza las suposiciones pertinentes sobre la probabilidad de que cada posición tenga o no variantes. Para que se realice el cálculo correctamente, se requiere introducir la secuencia de referencia (de nuevo). De esta manera, compara los valores obtenidos en el alineamiento previo con las posiciones y bases de la secuencia de referencia.
Además cabe tener en cuenta que el formato BCF que se obtiene en este paso es un formato binario, por lo que no se puede obtener información de las variantes detectadas por el momento.
Para la detección de los SNPs se utiliza el paquete BCFtools, concretamente la herramienta call:
bcftools call -mv -o exp20_var.vcf exp20_raw.bcf

Concretamente se indica que tan solo se muestren las variantes detectadas con el parámetro -v.
Teóricamente se debe realizar un filtraje de las variantes detectadas mediante el paquete BCFtools con la herramienta filter, pero se ha observado que las variantes detectadas en todos los experimentos no cambian al pasar por este filtro. De igual manera, al ser una operación rutinaria, se realiza:
bcftools filter exp20_var.vcf > exp20_final.vcf

En este punto ya se han obtenido las variantes del experimento en el que se haya realizado el procedimiento. Para poder visualizar mediante la aplicación IGV los SNPs respecto a la secuencia de referencia, se requiere de un archivo intermedio del experimento en formato BAI. Este archivo se utiliza para indexar y poder visualizar su archivo BAM correspondiente. Para crear el respectivo archivo BAI se utiliza el paquete SAMtools y la herramienta index:
samtools index exp20_r.bam

Con el archivo BAI, la secuencia de referencia y el respectivo archivo VCF, se puede observar en IGV las variantes detectadas en este protocolo. Además señalando cada variante se indican varios parámetros y más información sobre el SNP.
Como en este estudio la cantidad de variantes obtenidas no es elevada, se puede observar directamente el archivo VCF para encontrar la posición en las que se encuentran las variantes, el nucleótido de la secuencia de referencia y el nucleótido en que ha variado.

2.4 Obtención de las estructuras secundarias

Para obtener las estructuras secundarias de las secuencias analizadas, se utiliza el servicio RNAalifold. Para poder utilizar este servicio se requiere de alineamientos en formato Clustal o fasta. A partir de un alineamiento generado en UGENE utilizando el algoritmo ClustalO para cada experimento, se obtienen los alineamientos en el formato necesario para poder utilizar el servicio RNAalifold. Introduciendo estos alineamientos y seleccionando los parámetros estándar indicando que se quiere una estructura secundaria circular, la estructura secundaria se genera automáticamente.

2.5 Análisis de los resultados

2.5.1 Comparación de variantes encontradas

En el experimento 1.Fa_35S_Nb se han encontrado 5 variantes respecto a la secuencia de referencia (Tabla 1). De estas 5 mutaciones, 4 de ellas son transiciones y 1 es una transversión.

Tabla 1. Variantes encontradas en las secuencias obtenidas en el experimento 1.Fa_35S_Nb.

1.Fa_35S_Nb

Posición	Referencia	Variante	Tipo
438	A	G	Transición
596	A	G	Transición
611	G	A	Transición
630	A	G	Transición
676	U	G	Transversión

En el experimento 2.Fa_Nos_Nb se han encontrado seis variantes (tabla 2), 4 transiciones y 2 transversiones. Como se ha dicho anteriormente, en este experimento se han obtenido secuencias a los 2, 4, 6 y 8 días, pero en todas las secuenciaciones se han obtenido exactamente las mismas variantes en las mismas posiciones, por lo que solo se muestran aquí las variantes de uno de los días.

Tabla 2. Variantes encontradas en las secuencias obtenidas en el experimento 2.Fa_Nos_Nb.

2.Fa_Nos_Nb (Días 2 a 8)

Posición	Referencia	Variante	Tipo
12	C	U	Transición
102	C	A	Transversión
596	A	G	Transición
611	G	A	Transición
630	A	G	Transición
676	U	G	Transversión

En el experimento 3.Fa_Nos_At como se ha dicho anteriormente se ha utilizado una construcción con la secuencia mutada para que no circularice y otra construcción donde la secuencia sí circulariza (la secuencia de referencia utilizada en los otros experimentos). En la secuencia no circularizada se ha encontrado tan solo una variante (tabla 3) que es una transversión, mientras que

en la secuencia circularizada se han encontrado 7 variantes (tabla 3) de las cuales 5 son transiciones y 2 son transversiones. Comparando las variantes de ambas secuencias se observa que la secuencia circularizada tiene un total de 6 variantes más que la secuencia que no circulariza. Esto refuerza la hipótesis de que la circularización de la secuencia es necesaria para la replicación y/o edición del RNA del retrozima, provocando bien durante la replicación (posiblemente del tipo de círculo rodante) o mediante factores de edición del RNA, la aparición de mutaciones en la secuencia.

Tabla 3. Variantes encontradas en las secuencias obtenidas en el experimento 3.Fa_Nos_At, tanto del retrozima que expresa circRNA como del que expresa RNA lineal.

3.Fa_Nos_At (circRNA)				3.Fa_Nos_At (RNA lineal)			
Posición	Referencia	Variante	Tipo	Posición	Referencia	Variante	Tipo
12	C	U	Transición	663	C	U	Transición
102	C	A	Transversión				
438	A	G	Transición				
596	A	G	Transición				
611	G	A	Transición				
630	A	G	Transición				
676	U	G	Transversión				

Ahora se puede realizar una comparación de todas las variantes encontradas en los distintos experimentos. Observando las posiciones y cada tipo de variante (transición o transversión) se pueden detectar variantes similares entre los distintos experimentos realizados.

Además se va a comparar la secuencia consenso con las variantes naturales obtenidas del NCBI con la secuencia de referencia utilizada en cada experimento. De esta manera se puede observar si las mutaciones obtenidas en cada experimento ya han sido encontradas como variantes obtenidas en el NCBI o se han detectado variantes causadas durante el experimento.

En el experimento 1.Fa_35S_Nb se encuentran exactamente las mismas 5 variantes que aparecen en las secuenciaciones de circRNA del experimento 3.Fa_Nos_At. En cambio las dos primeras variantes encontradas en las secuenciaciones de circRNA del experimento 3.Fa_Nos_At (las posiciones 12 y 102) no aparecen en el experimento 1.Fa_35S_Nb.

De igual manera, realizando una comparación entre las variantes encontradas en el experimento 2.Fa_Nos_Nb y las secuenciaciones de circRNA del experimento 3.Fa_Nos_At, excepto por una variante en la posición 438 de las secuenciaciones de circRNA, todas las demás variantes son iguales, tanto en la posición en la que se encuentran como con en el tipo de variante. En cambio, en las secuenciaciones del RNA lineal del experimento 3.Fa_Nos_At tan solo aparece una variante, la cual además no aparece reflejada en ningún otro experimento.

Finalmente comparando las variantes encontradas entre el experimento 1.Fa_35S_Nb y el experimento 2.Fa_Nos_Nb, se encuentran similitudes en 4 variantes (que se encuentran en las posiciones 596, 611, 630 y 676) de las 5 que tiene 1.Fa_35S_Nb. Esta coincidencia en las variantes encontradas entre los dos experimentos indicaría que estas posiciones del retrozima bien tienen una alta presión de selección o bien son puntos diana de los factores de edición del RNA. De todas las variantes encontradas en los experimentos, las que también se han encontrado como variantes naturales ya depositadas en el NCBI son las que se encuentran en las posiciones 438, 596, 630 y 676. Cabe destacar además que las variantes naturales son el mismo nucleótido que las encontradas en los distintos experimentos, lo que nos podría estar indicando algún tipo de reversión de la secuencia parental genómica hacia secuencias de RNA tal y como se dan *in vivo*.

Con las observaciones realizadas, se encuentra que una elevada parte de las variantes se encuentra entre la posición 400 y 700, indicando que la mayor tasa de variación aparece en estas regiones. Como observación más destacada, se ha encontrado entre las comparaciones de los tres experimentos cuatro variantes que aparecen en todos ellos en las posiciones 596, 611, 630 y 676, siendo las tres primeras variantes transiciones y la cuarta variante una transversión. De las variantes naturales depositadas en el NCBI mencionadas anteriormente, tres de estas variantes concuerdan con las variantes que aparecen en todos los experimentos, siendo concretamente las de las posiciones 596, 630 i 676. Esto implica que estas variantes no parecen haber sido causa de los experimentos que se han realizado, sino que de manera natural el retrozima revierte a estas variantes.

Para saber qué puede causar estas mutaciones conservadas, se observa la posición en la que se encuentran las mutaciones en el retrozima. Para ello se indican todas las variantes encontradas en los distintos experimentos (excepto el del RNA lineal por ser una variante que solo aparece en estas secuencias) en el retrozima (Figura 9). En el retrozima circRNA se puede observar que todas las variantes menos una se encuentran en las región de la LTR, incluída la HHR. La única variante que no se encuentra en esta región, aparte de ser una transición, se encuentra en un extremo de la región intermedia del retrozima, y muy cerca del motivo PPT.

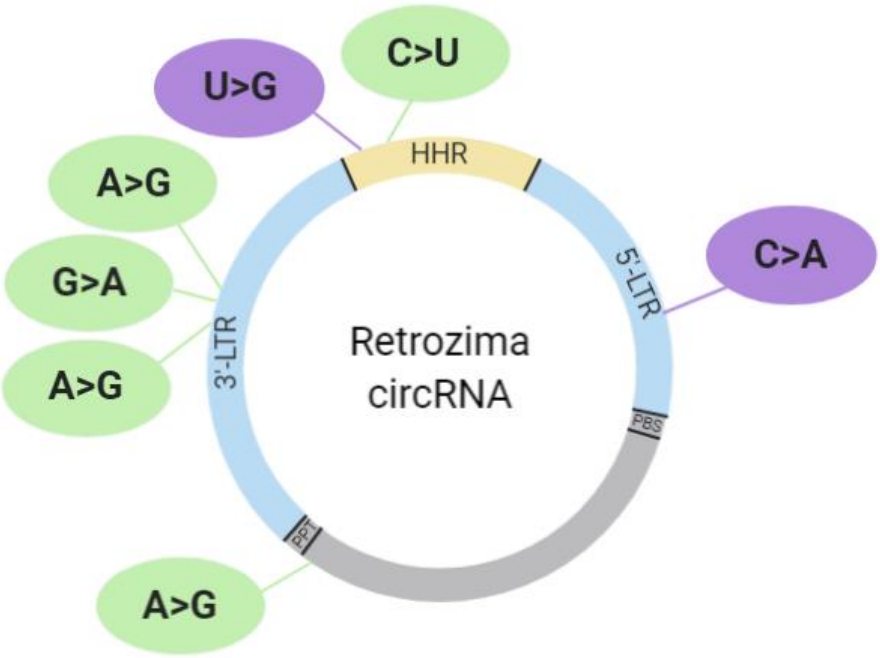

Figura 8. Representación del circRNA de un retrozima con las variantes encontradas marcadas en sus posiciones. Las transiciones se muestran en verde y las transversiones en morado.

2.5.2 Comparación de estructuras secundarias

Se han obtenido 4 estructuras secundarias coloreadas según la probabilidad de emparejamiento en los pares de bases predichos. Se ha obtenido una estructura para el experimento 1.Fa_35S_Nb, otra para el experimento 2.Fa_Nos_Nb (tan solo se obtiene la estructura secundaria del día 2 ya que las variantes no han cambiado entre los distintos días, por lo que no habrá cambios en sus respectivas estructuras secundarias) y dos para el experimento 3.Fa_Nos_At,

siendo una para las variantes detectadas a partir de circRNA y otra para las variantes detectadas a partir de RNA lineal.

Observando la estructura secundaria del experimento 1.Fa_35S_Nb y la del experimento 2.Fa_Nos_Nb (Figura 9a y 9b respectivamente), se encuentran diferencias visuales fácilmente discernibles a pesar de ser muy similares en su secuencia y variantes detectadas. Las partes donde mayor similitud se encuentra es en la región de la ribozima HHR, señalada en ambas estructuras en la parte superior de las estructuras. Aparte de esto, parece que existe una similitud en la estructura general, variando principalmente en las terminaciones o ramificaciones procedentes de la estructura principal. El experimento 2.Fa_Nos_Nb tiene una estructura predicha con unas probabilidades de pares de bases más elevada en casi toda su estructura, indicando que es más probable que esta estructura esté correctamente predicha en comparación a la obtenida en el experimento 1.Fa_35S_Nb.

Las dos estructuras secundarias obtenidas en 3.Fa_Nos_At también contienen claras diferencias entre ellas (Figura 10a y 10b)., aunque aquí la diferencia en la cantidad de variantes en cada secuencia es suficiente motivo para dichos cambios. En la estructura secundaria obtenida a partir de circRNAs las probabilidades de cada par de bases son muy bajas, indicando que la predicción de esta estructura es menos probable, siendo el caso contrario en la estructura secundaria obtenida a partir de RNA lineales. En este caso la mayoría de la estructura secundaria tiene las probabilidades de pares de bases muy elevada, indicando que es probable que la predicción de la estructura sea correcta. Aun así, la presencia de la HHR es difícil de distinguir debido a la elevada cantidad de hélices predichas en ambas estructuras. En principio el hecho de que la HHR sea difícil de distinguir en la estructura secundaria procedente de secuencias de RNA lineal puede ser debido a la inactivación de ésta mencionada en el experimento, dificultando la formación del centro catalítico esperado. En cambio, no se esperaba la estructura del HHR formada en la estructura secundaria de secuencias de circRNA.

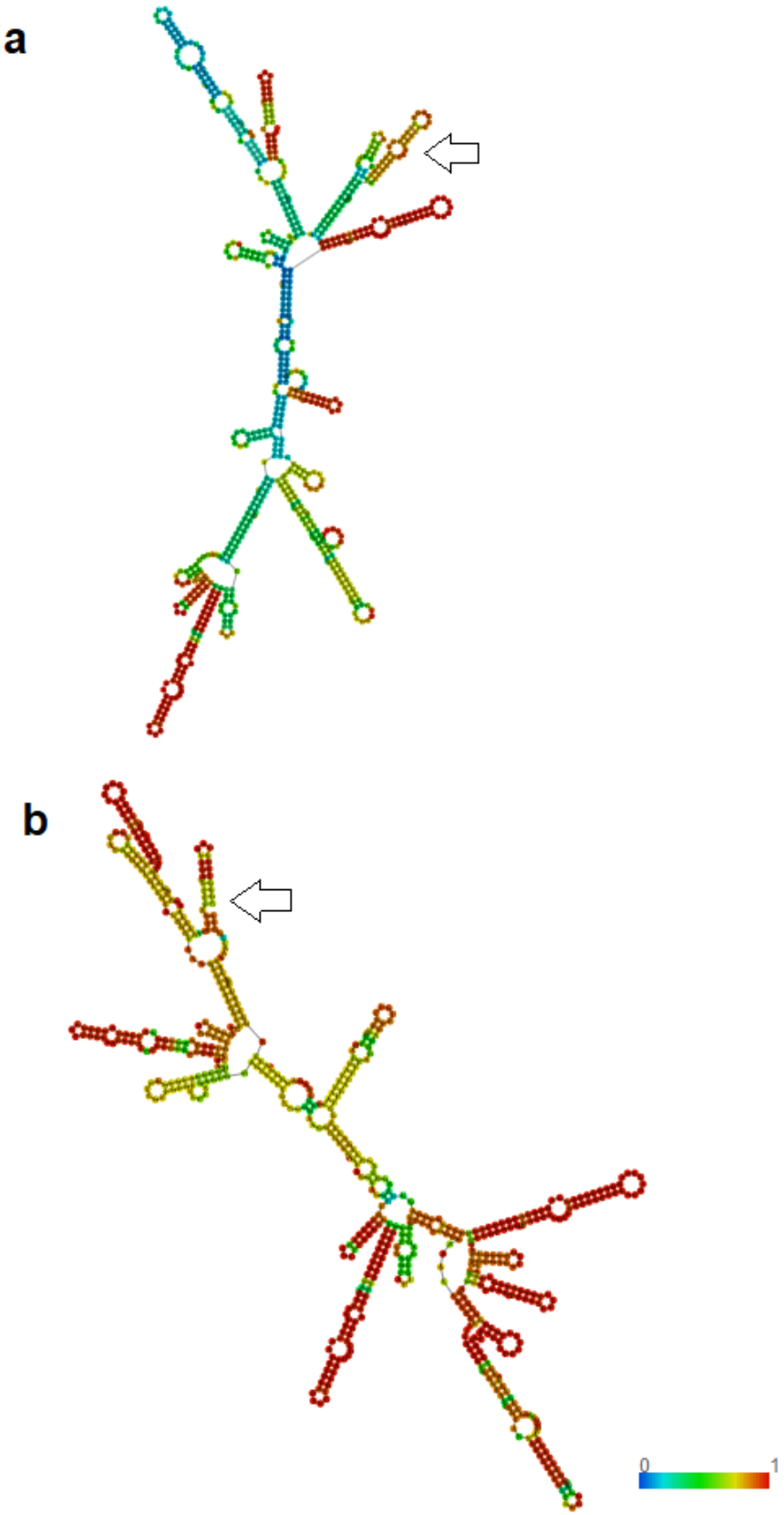

Figura 9. Estructuras secundarias obtenidas de los experimentos 1.Fa_35S_Nb (a) y 2.Fa_Nos_Nb (b). Los colores indican las probabilidades de los pares de base. La flecha indica la región HHR.

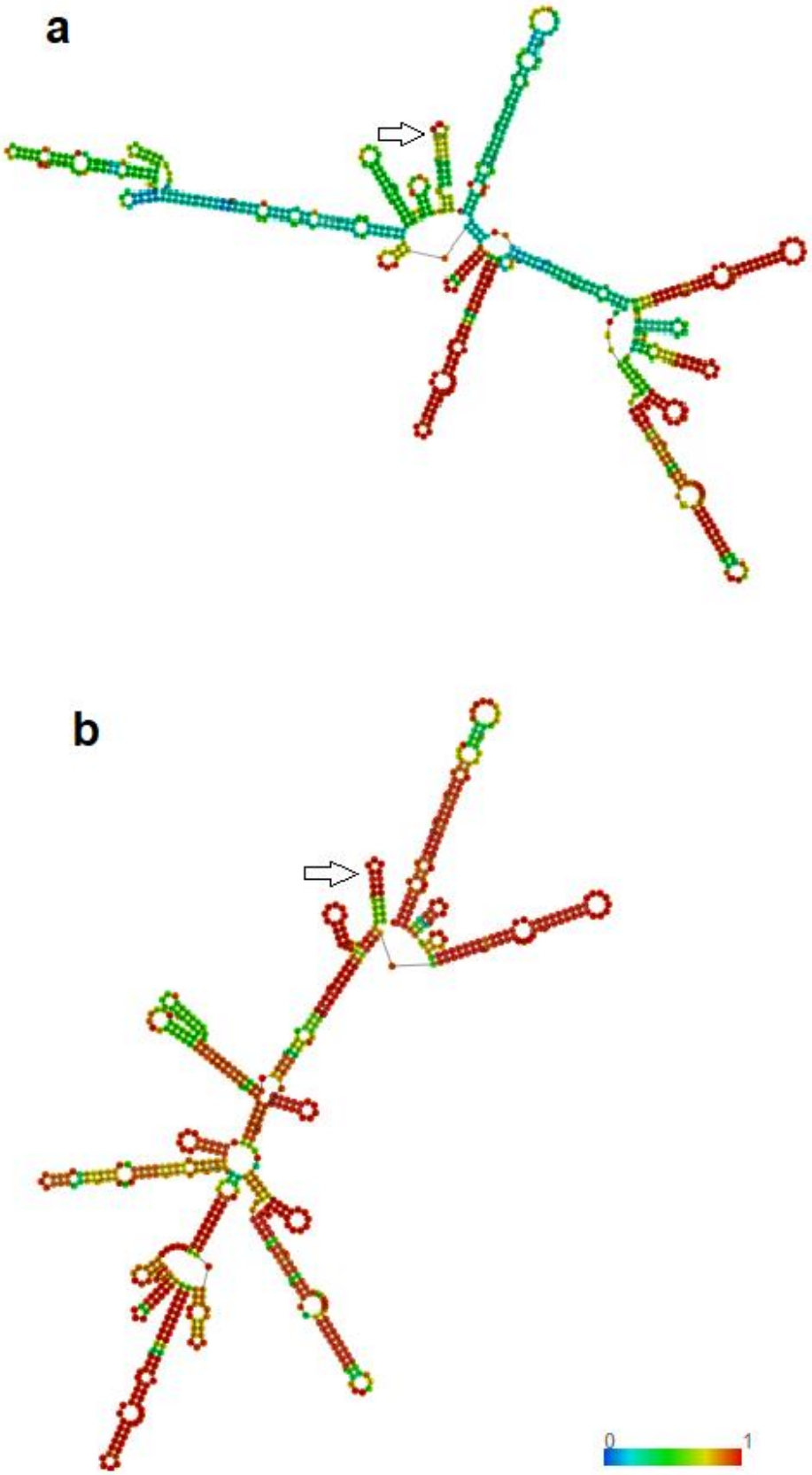

Figura 10. Estructuras secundarias obtenidas del experimentos 3.Fa_Nos_At procedentes de circRNA (a) o de RNA lineal (b). Los colores indican las probabilidades de los pares de base. La flecha indica la región HHR.

Como última comparación se pueden observar las estructuras secundarias entre el experimento 2.Fa_Nos_Nb y la estructura secundaria obtenida a partir de circRNAs del experimento 3.Fa_Nos_At. En este caso, al haber tan solo una variante que difiere entre las dos estructuras, éstas deberían ser similares en la mayoría de la secuencia. Sin embargo, excepto una región que se encuentra igual en ambas estructuras (la región inferior derecha de las imágenes), el resto es diferente en varios puntos de sus estructuras. Donde más se destacan estas diferencias es en la estructura de las HHR, la cual contiene un mayor número de hélices próximas en el experimento 3.Fa_Nos_At haciéndola difícil de distinguir, a diferencia de la HHR de 2.Fa_Nos_Nb.

Debido a la elevada cantidad de diferencias entre todas las estructuras obtenidas y sus dispares probabilidades en sus pares de bases, no se han podido obtener conclusiones claras respecto a cómo pueden afectar las variantes a la funcionalidad del retrozima.

3. Conclusiones

En este trabajo se ha conseguido la obtención de las variantes o SNPs en tres experimentos. A partir de las variantes de cada experimento se han extraído varias conclusiones.

En el experimento 1.Fa_35S_Nb se ha comprobado que al cabo de 6 días han aparecido SNPs o variantes respecto a la secuencia de referencia con la que se había obtenido la construcción correspondiente. Esto se valida con la hipótesis inicial de que debido a la posible replicación de este retrozima por círculo rodante o por edición del RNA, aumenta la variabilidad de la secuencia.

En el experimento 2.Fa_Nos_Nb los SNPs o variantes detectadas al segundo día post-agroinfiltración son exactamente los mismos que en los encontrados en las posteriores extracciones (a los 4, 6 y 8 días post-agroinfiltración). En este experimento se podría esperar un aumento de variantes detectadas con el paso del tiempo si estas fueran debido a la replicación del círculo rodante mencionada en el anterior experimento. En cambio, con los resultados obtenidos se lleva a pensar que la variabilidad del retrozima no se ve especialmente incrementada

con el paso del tiempo, siendo los primeros días donde mayor variabilidad se da en el retrozima. Esto podría ocurrir debido a que los niveles de expresión del retrozima disminuyen al cabo del tiempo conforme se va acumulando el circRNA en los tejidos [24], disminuyendo entonces su posible replicación y por tanto disminuyendo la variabilidad. También es factible que el proceso de edición se complete al principio de la expresión y solo observemos las moléculas ya editadas. Estas teorías por el momento no están probadas, y se investigarán con detalle en estudios futuros.

Con el análisis de variantes del experimento 3.Fa_Nos_At se ha confirmado que el hecho de que el transcrito del retrozima pueda o no circularizar afecta significativamente a la variabilidad en su secuencia. Concretamente, el hecho de que el transcrito del retrozima no circularice tiene como consecuencia una disminución significativa en la aparición de variantes. Con esto se puede confirmar que la circularización del RNA procedente del retrozima es condición necesaria para que el retrozima sufra una alta variabilidad de secuencia.

Como conclusión del análisis de varianza cabe comentar que la mayoría de las variantes detectadas se encuentran en las regiones LTR, por lo que estas variaciones no deberían tener un impacto en la actividad del retrozima. Sin embargo, las variantes detectadas en la región HHR y la región intermedia (sobre todo las transversiones) pueden dar lugar a un cambio en la funcionalidad de estas regiones.

En las estructuras secundarias obtenidas en este estudio se han observado grandes diferencias entre los distintos experimentos, incluso en estructuras que tenían pocas variantes para afectar a su estructura. Esto es debido principalmente a que algunas predicciones de estructuras tienen una baja probabilidad en sus pares de bases, indicando que la probabilidad de que aparezcan estas estructuras no es elevada. Respecto al efecto de las variantes en las estructuras secundaria, no se han podido obtener resultados relevantes en este estudio.

Los retrozimas requieren de un extenso estudio para poder entender sus mecanismos, y en este trabajo se ha realizado un análisis superficial de su variabilidad. Esto implica que hay varias líneas de trabajo que pueden llevarse a cabo para obtener más conocimientos sobre los retrozimas. Por ejemplo, para probar la hipótesis de que la expresión del retrozima disminuye en el tiempo, se

podrían analizar los niveles de expresión del retrozima en el tiempo mediante un experimento similar. También para determinar los posibles efectos de las variantes en la estructura secundaria se podría realizar un análisis de una mayor cantidad de experimentos similares para obtener una mayor cantidad de secuencias y consecuentemente estructuras secundarias. De esta manera se podrían establecer patrones y cambios entre las distintas estructuras secundarias e investigar así los cambios funcionales producidos por éstos.

4. Glosario

At: *Arabidopsis thaliana*

circRNA: RNA circular

DIRS: Secuencia Repetida Intermediaria de Dictyostelium

EN: Endonucleasa

Fa: *Fragaria ananassa*

GATK: Herramientas de análisis de genomas

glmS: Ribozimas de la glucosamina-6-fosfato sintasa

HDV: Ribozima Hepatitis Delta Virus

HHR: Ribozimas de cabeza de martillo

HPR: Ribozimas de horquilla

IN: Integrasa

ITR: Repetición terminal invertida

LINE: Elementos largos intercalados

LTR: Repeticiones largas terminales

MSA: Alineamiento de secuencias múltiple

NGS: Secuenciación de la siguiente generación (Next Generation Sequencing)

ORF: Pauta de lectura abierta

PBS: Sitio de Unión al Primer

PLE: Elementos similares a los elementos Penelope

PLTR: Secuencias similares a LTR en PLE

PPT: Tracto Polipurina

RH: RNasaH

RT: Retrotransciptasa

SINE: Elementos cortos intercalados

TE: Elementos transponibles

TSD: Zona diana de duplicación

Tw: Ribozimas Twister

Tws: Ribozimas Twister sister

UTR: Regiones no traducidas

VS: Ribozimas de satélite Varkud

YR: Recombinasa

5. Bibliografía

[1] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech, “Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena,” *Cell*, vol. 31, no. 1, pp. 147–157, 1982.

[2] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman, “The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme,” *Cell*, vol. 35, no. 3 PART 2, pp. 849–857, 1983.

[3] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz, “The structural basis of ribosome activity in peptide bond synthesis,” *Science*, vol. 289, no. 5481, pp. 920–930, 2000.

[4] S. Valadkhan and J. L. Manley, “Splicing-related catalysis by protein-free snRNAS,” *Nature*, vol. 413, no. 6857, pp. 701–707, 2001.

[5] T. J. Wilson and D. M. J. Lilley, “Do the hairpin and VS ribozymes share a common catalytic mechanism based on general acid-base catalysis? A critical assessment of available experimental data,” *RNA*, vol. 17, no. 2. RNA, pp. 213–221, 2011.

[6] E. A. Doherty and J. A. Doudna, “Ribozyme structures and mechanisms,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 30. Annu Rev Biophys Biomol Struct, pp. 457–475, 2001.

[7] G. A. Prody, J. T. Bakos, J. M. Buzayan, I. R. Schneider, and G. Bruening, “Autolytic processing of dimeric plant virus satellite RNA,” *Science (80-.).*, vol. 231, no. 4745, pp. 1577–1580, 1986.

[8] A. R. Ferré-D’Amaré and W. G. Scott, “Small self-cleaving ribozymes.,” *Cold Spring Harbor perspectives in biology*, vol. 2, no. 10. Cold Spring Harb Perspect Biol, 2010.

[9] W. Gilbert, “Origin of life: The RNA world,” *Nature*, vol. 319, no. 6055, p. 618, 1986.

[10] M. De La Peña and I. García-Robles, “Ubiquitous presence of the hammerhead ribozyme motif along the tree of life,” *RNA*, vol. 16, no. 10, pp. 1943–1950, 2010.

[11] C. J. Hutchins1, P. D. Rathjen2, A. C. Forster, and R. H. Symons, “Self-

cleavage of plus and minus RNA transcripts of avocado sunblotch viroid," Nucleic Acids Res, vol. 14, no. 9, pp. 3627-3640, 1986.

[12] L. M. Epstein and J. G. Gall, "Self-cleaving transcripts of satellite DNA from the newt," *Cell*, vol. 48, no. 3, pp. 535–543, 1987.

[13] G. Ferbeyre, J. M. Smith, and R. Cedergren, "Schistosome Satellite DNA Encodes Active Hammerhead Ribozymes," *Mol. Cell. Biol.*, vol. 18, no. 7, pp. 3880–3888, 1998.

[14] A. A. Rojas *et al.*, "Hammerhead-mediated processing of satellite pDo500 family transcripts from Dolichopoda cave crickets," vol. 28, no. 20, pp. 4037-4043, 2000.

[15] M. De La Peña and I. García-Robles, "Intronic hammerhead ribozymes are ultraconserved in the human genome," *EMBO Rep.*, vol. 11, no. 9, pp. 711–716, 2010.

[16] C. Hammann, A. Luptak, J. Perreault, and M. De La Peña, "The ubiquitous hammerhead ribozyme," *RNA*, vol. 18, no. 5. Cold Spring Harbor Laboratory Press, pp. 871–885, 2012.

[17] M. De La Peña, I. García-Robles, and A. Cervera, "The hammerhead Ribozyme: A long history for a short RNA," *Molecules*, vol. 22, no. 1. MDPI AG, 2017.

[18] H. W. Pley, K. M. Flaherty, and D. B. McKay, "Three-dimensional structure of a hammerhead ribozyme," *Nature*, vol. 372, no. 6501, pp. 68–74, 1994.

[19] W. G. Scott, J. T. Finch, and A. Klug, "The crystal structure of an AII-RNAhammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage," *Cell*, vol. 81, no. 7, pp. 991–1002, 1995.

[20] A. Khvorova, A. Lescoute, E. Westhof, and S. D. Jayasena, "Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity," *Nat. Struct. Biol.*, vol. 10, no. 9, pp. 708–712, 2003.

[21] M. De La Peña, S. Gago, and R. Flores, "Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity," The EMBO Journal Vol. 22, no. 20, pp. 5561-5570, 2003

[22] M. Martick and W. G. Scott, "Tertiary Contacts Distant from the Active Site Prime a Ribozyme for Catalysis," *Cell*, vol. 126, no. 2, pp. 309–320, 2006.

[23] A. Cervera and M. De La Peña, “Eukaryotic Penelope-Like Retroelements Encode Hammerhead Ribozyme Motifs,” Mol Biol Evol, vol. 31, no. 11, pp. 2941-2947, 2014.

[24] A. Cervera, D. Urbina, and M. de la Peña, “Retrozymes are a unique family of non-autonomous retrotransposons with hammerhead ribozymes that propagate in plants through circular RNAs,” *Genome Biol.*, vol. 17, no. 1, p. 135, 2016.

[25] A. Cervera, M. De La Peña, “Small circRNAs with self-cleaving ribozymes are highly expressed in diverse metazoan transcriptomes,” *Nucleic Acids Res.*, vol. 48, no. 9, pp. 5054–5064, 2020.

[26] M. Martick, L. H. Horan, H. F. Noller, and W. G. Scott, “A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA,” *Nature*, vol. 454, no. 7206, pp. 899–902, 2008.

[27] P. S. Schnable *et al.*, “The B73 maize genome: Complexity, diversity, and dynamics,” *Science (80-.).*, vol. 326, no. 5956, pp. 1112–1115, 2009.

[28] M. Morgante, “Plant genome organisation and diversity: the year of the junk!,” *Current Opinion in Biotechnology*, vol. 17, no. 2. Elsevier Current Trends, pp. 168–173, 2006.

[29] T. Wicker *et al.*, “A unified classification system for eukaryotic transposable elements,” *Nature Reviews Genetics*, vol. 8, no. 12. Nature Publishing Group, pp. 973–982, 2007.

[30] W. Makałowski, A. Pande, V. Gotea, and I. Makałowska, “Transposable elements and their identification,” *Methods Mol. Biol.*, vol. 855, pp. 337–359, 2012.

[31] W. Makałowski, V. Gotea, A. Pande, and I. Makałowska, “Transposable elements: Classification, identification, and their use as a tool for comparative genomics,” in *Methods in Molecular Biology*, vol. 1910, Humana Press Inc., pp. 177–207, 2019.

[32] H. S. Malik, W. D. Burke, and T. H. Eickbush, “The age and evolution of non-LTR retrotransposable elements,” *Mol. Biol. Evol.*, vol. 16, no. 6, pp. 793–805, 1999.

[33] S. Orozco-Arias, G. Isaza, and R. Guyot, “Retrotransposons in plant genomes: Structure, identification, and classification through bioinformatics and machine learning,” *International Journal of Molecular*

Sciences, vol. 20, no. 15. MDPI AG, 2019.

[34] M. de la Peña and A. Cervera, “Circular RNAs with hammerhead ribozymes encoded in eukaryotic genomes: The enemy at home,” *RNA Biol.*, vol. 14, no. 8, pp. 985–991, 2017.

[35] “National Center for Biotechnology Information.” [Online]. Available: https://www.ncbi.nlm.nih.gov/. [Accessed: 16-Jun-2020].

[36] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” Nucleic Acids Res., vol. 32, no. 5, pp. 1792-1797, 2004.

[37] F. Sievers and D. G. Higgins, “Clustal Omega for making accurate alignments of many protein sequences,” *Protein Sci.*, vol. 27, no. 1, pp. 135–145, 2018.

[38] K. Katoh and D. M. Standley, “Article Fast Track MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability,” Mol Biol Evol., vol. 30, no. 4, pp. 772-780, 2013.

[39] “Performance Evaluation of Leading Protein Multiple Sequence Alignment Methods,” Evol Bioinform Online., vol. 10, pp. 205-217, 2014.

[40] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[41] “bcftools.” [Online]. Available: http://samtools.github.io/bcftools/bcftools.html. [Accessed: 09-April-2020].

[42] A. McKenna *et al.*, “The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, 2010.

[43] “bwa.1.” [Online]. Available: http://bio-bwa.sourceforge.net/bwa.shtml. [Accessed: 09-April-2020].

[44] “Broad Institute.” [Online]. Available: https://www.broadinstitute.org/. [Accessed: 16-March-2020].

[45] “Best Practices Workflows – GATK.” [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows. [Accessed: 16-March-2020].

[46] M. Frampton and R. Houlston, “Generation of Artificial FASTQ Files to Evaluate the Performance of Next-Generation Sequencing Pipelines,” *PLoS One*, vol. 7, no. 11, p. e49110, 2012.

[47] H. M. Schilbert, A. Rempel, and B. Pucker, “Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data,” *Plants*, vol. 9, no. 4, p. 439, 2020.

[48] W. Mu, H. M. Lu, J. Chen, S. Li, and A. M. Elliott, “Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing,” *J. Mol. Diagnostics*, vol. 18, no. 6, pp. 923–932, 2016.

[49] R. Lorenz *et al.*, “ViennaRNA Package 2.0,” *Algorithms Mol. Biol.*, vol. 6, no. 1, 2011.

[50] “ViennaRNA Web Services.” [Online]. Available: http://rna.tbi.univie.ac.at/. [Accessed: 18-May-2020].

[51] K. Okonechnikov, O. Golosova, and M. Fursov, “Genome analysis Unipro UGENE: a unified bioinformatics toolkit,” *Bioinforma. Appl. NOTE*, vol. 28, no. 8, pp. 1166–1167, 2012.

[52] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nature Biotechnology*, vol. 29, no. 1. Nature Publishing Group, pp. 24–26, 01-2011.

[53] “Clustal Omega < Multiple Sequence Alignment < EMBL-EBI.” [Online]. Available: https://www.ebi.ac.uk/Tools/msa/clustalo/. [Accessed: 17-April-2020].

[54] P. J. A. Cock *et al.*, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

6. Anexos

ANEXO A. Python script para reordenar las secuencias a partir de un patrón

```
import Bio
from Bio import SeqIO
from Bio.SeqRecord import SeqRecord
from Bio.Seq import Seq
input_file = input("Write the path to your fasta sequence: ")
output_file = input("Name the fasta sequence reordered: ")
restriction = input("Write the place where it has to start reordering the sequence: ")

with open(output_file, "w") as f:
        for seq_record in SeqIO.parse(input_file, "fasta"):
                cut = seq_record.seq.find(restriction)
                f.write(">" + str(seq_record.id) + "\n" +
                str(seq_record.seq[cut:]) + str(seq_record.seq[:cut]) + "\n")
print("The reordering should be done, restart the script to reorder another fasta sequence")
```

ANEXO B. Bash script para detectar las variantes de las secuencias en formato VCF

```
#!/bin/bash

#!/bin/bash

# Input of names from files that contain the reference sequence/genome
# and the target sequences

read -p 'Input the path: ' path
if [[ -d $path ]]; then
        echo 'Path loaded succesfully'
else
        until [[ -d $path ]]; do
                read -p 'Path not found, try again: ' path
        done
        echo 'Path loaded succesfully'
fi
cd $path


read -p 'Input the reference genome/sequence: ' ref
if [[ -f $ref ]]; then
        echo 'Reference genome/sequence loaded'
else
        until [[ -d $ref ]]; do
                read -p 'Path not found, try again: ' ref
        done
        echo 'Reference genome/sequence loaded'
fi


read -p 'Input file with sequences to align with reference: ' seq
if [[ -f $seq ]]; then
```

```
        echo 'Sequences loaded'
else
        until [[ -d $seq ]]; do
                read -p 'Path not found, try again: ' seq
        done
        echo 'Sequences loaded'
fi

# Sequence alignment with reference genome/sequence

## Index the reference genome/sequence
bwa index $ref

## Sequence alignment with the reference sequence

bwa mem $ref $seq > ${seq%%.*}.sam

## sam to bam conversion
samtools view -S -b ${seq%%.*}.sam > ${seq%%.*}.bam

## Ordering of bam coordinates:
samtools sort -o ${seq%%.*}_s.bam ${seq%%.*}.bam


# Variant detection

## Calculate read coverage position in the reference sequence
bcftools mpileup -O b -o ${seq%%.*}_raw.bcf -f $ref ${seq%%.*}_s.bam

## Variant detection:
bcftools call -mv -o ${seq%%.*}_var.vcf ${seq%%.*}_raw.bcf

## Filter and report of SNP variants in vcf format
vcfutils.pl varFilter ${seq%%.*}_var.vcf > ${seq%%.*}_final.vcf
```

echo 'End of script, to generate more vcf files restart again.'