UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Herramientas preliminares para la secuenciación de genomas: Genotecas con vectores de clonación

Apellidos, nombre	Vilanova Navarro, Santiago (sanvina@upv.es) Gadea Vacas, José (jgadeav@ibmcp.upv.es)
Departamento	Departamento de Biotecnología
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

La reciente evolución de las tecnologías de secuenciación ha hecho posible que los proyectos de secuenciación de genomas estén al alcance de mucha más gente. Hoy en día, incluso pequeños laboratorios con presupuestos ajustados pueden aventurarse a realizar este tipo de proyectos. El límite no es tanto el precio como la planificación previa y el análisis posterior de todos los datos que nos dan las tecnologías actuales.

Por esta razón es importante conocer el tipo de herramientas que podemos utilizar para la secuenciación (vectores, plataformas de "Next Generation Sequencing", mapas genéticos, mapas físicos, etc.), así como qué estrategias podemos seguir para llevar a cabo la misma. También es de vital importancia conocer las herramientas bioinformáticas imprescindibles para el análisis de los datos. En este punto cabe destacar que no se trata tanto de saber enumerar todos los programas que existen, sino más bien conocer qué hacen y qué tipo de programas necesitaremos para cada uno de los pasos.

Es importante tener datos del genoma a secuenciar, tanto de tamaño como de complejidad (cantidad de repetitivo, ploidía, endoduplicaciones, etc), ya que esto será crucial para elegir la estrategia a seguir. También es importante determinar cuál va ser nuestro objetivo final. Es este caso no es lo mismo querer tener un simple borrador del genoma que obtener un genoma de referencia, ya que la estrategia puede variar considerablemente. Por último, también deberemos considerar cuál será el material de partida, ya que no es lo mismo secuenciar un organismo del que es fácil obtener ADN de buena calidad que intentar reconstruir el genoma de muestras fosilizadas o de herbarios.

En este artículo se pretende revisar las primeras fases de la secuenciación de los genomas (Extracción de ADN, fragmentación y construcción de las librerías), de manera que el alumno pueda ser consciente de todas las posibilidades de las que puede disponer para que les sea más fácil seleccionar cual o cuales serán más adecuadas para un proyecto real de secuenciación.

2 Objetivos

Una vez que el alumno haya leído con detenimiento este documento, será capaz de:

1. Describir y explicar los primeros pasos a seguir para secuenciar un genoma

2. Calcular cuántos clones necesitará para poder tener una buena representación del genoma

3. Calcular el número de clones que se necesitan para poder obtener una cobertura del genoma adecuada para realizar el ensamblaje del genoma

4. Entender la terminología utilizada en la secuenciación de genomas.

3 Introducción

Aunque la manera de secuenciar genomas cambia de manera rápida y constante, existen una serie de ideas generales que pueden servir de guía para la secuenciación de los genomas. A grandes rasgos, la secuenciación de cualquier genoma sigue los siguientes pasos:

1. Extracción del ADN

2. Fragmentación del ADN
3. Construcción de las librerías Genómicas
4. Secuenciación de las librerías
5. Ensamblaje de las secuencias
6. Anotación de las secuencias obtenidas

En este artículo docente nos vamos a centrar en las primeras fases que se realizan antes de secuenciar el genoma. Las fases iniciales son de vital importancia y conocerlas facilitará el futuro éxito de la secuenciación del genoma. Estas fases iniciales se pueden dividiré en extracción de ADN, fragmentación del mismo y construcción de las librerías. En concreto en este artículo nos centraremos en la construcción de las genotecas mediante vectores de clonación.

4 Desarrollo

Para entender la construcción de las genotecas sería recomendable tener conocimientos de las técnicas generales de genética molecular (digestión, clonación, electroforesis, etc)

Como ya hemos comentado, la construcción de genotecas es fundamental para secuenciar genomas. Gracias a la construcción de diversas genotecas fue posible la secuenciación del genoma humano publicado en 2001 (Venter et al.2001; International Human Genome Consortium, 2001). En este artículo se describen los pasos a seguir haciendo hincapié en los detalles más relevantes y explicando las posibles decisiones que el alumno deberá considerar antes de empezar la realización de las genotecas. Para apoyar el proceso de aprendizaje se utilizarán enlaces a páginas web y/o vídeos donde se explican con más detalle algunos conceptos.

4.1. Extracción del ADN

El primer paso crucial para la secuenciación de todo el genoma es el aislamiento de **suficiente cantidad de ADN de calidad**. En este punto lo más importante es que tenga ausencia de contaminantes que puedan inhibir las reacciones enzimáticas. En algunos casos especiales, como cuando se trata de material antiguo proveniente de fósiles, endosimbiontes, u organismos poco abundantes o en peligro de extinción, obtener ADN de buena calidad y en suficiente cantidad resulta complicado.

La calidad del ADN se medirá tanto en espectrofotómetro como mediante gel de agarosa. Unos ratios de **absorbancia A_{260}/A_{280} y A_{260}/A_{230} cercanos a 2** son aconsejables, así como una banda clara y sin degradar en el gel de agarosa. La concentración de ADN se medirá también mediante fluorometría, ya que la absorbancia puede sobreestimar la cantidad de ADN real que tenemos.

Una vez obtenido el ADN ya podemos pasar a la construcción de las librerías genómicas.

4.2. Fragmentación del ADN

El primer paso para la construcción de librerías genómicas va a ser habitualmente la fragmentación del ADN genómico en trozos más pequeños. El tamaño dependerá del tipo de vector que se quiera utilizar, pudiendo variar desde 200 kb hasta 1000 kb.

Para la realización de genotecas con vectores de clonación usaremos siempre la digestión mediante el método enzimático.

4.2.1 Fragmentación mediante el método enzimático

La fragmentación mediante métodos enzimáticos se realiza principalmente mediante enzimas de restricción tipo II (con una diana conocida). Hay que destacar que **siempre se realizará una digestión parcial** de ADN genómico, de manera que los **fragmentos resultantes sean solapantes**, ya que de otra manera será imposible reconstruir la secuencia original del genoma. Para ello deberemos jugar con las condiciones de la digestión en cuanto a tiempo, concentración del enzima y temperatura. De esta manera obtendremos las condiciones óptimas para obtener fragmentos del tamaño adecuado para la construcción de las genotecas.

El problema principal con esta técnica, sobre todo cuando utilizamos enzimas de restricción tipo II, es que las dianas de corte no están repartidas equitativamente por todo el genoma, lo que da como resultado la pérdida de parte del genoma al final del proceso. Esto se suele solucionar realizando diferentes genotecas utilizando diferentes enzimas de restricción.

4.3. Construcción de las genotecas mediante Clonación

Antes de todo cabe definir las genotecas ("genomic libraries" en inglés) como la colección de fragmentos de ADN de un organismo que en conjunto representan idealmente (aunque no necesariamente) la totalidad del genoma. Es decir, una genoteca genómica ideal contendría todo el genoma en trozos más o menos pequeños, y, si somos capaces de ordenarlos, seremos capaces de obtener la secuencia completa del genoma.

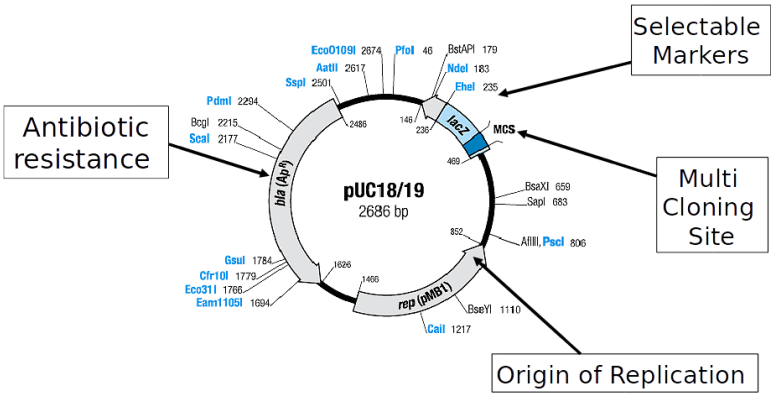
Las primeras genotecas utilizadas para la secuenciación de genomas se basaban en el uso de vectores de clonación. Aunque cada vez menos utilizadas, todavía siguen siendo clave para la obtención de genomas de referencia, sobre todo cuando se trata de genomas "complejos" con gran cantidad de repetitivo. El método más habitual de fragmentar el ADN en este tipo de genotecas es el enzimático, seguido por una selección de tamaño en gel de agarosa.

Los vectores de clonación utilizados se clasifican según el tamaño de ADN exógeno que pueden contener. De mayor a menor encontramos los plásmidos, los vectores basados en el fago lambda, los cósmidos y los BACs. Veamos en detalle las características de cada uno de ellos.

4.3.1 Los plásmidos

Los plásmidos son fragmentos de ADN circular que se encuentran de manera natural en las bacterias. Los plásmidos comerciales existentes en el mercado solo han modificado ligeramente los plásmidos naturales, de manera que nos puedan servir de herramienta. Así, se ha modificado el origen de replicación de manera que podamos obtener 500 o más copias del plásmido dentro de cada bacteria. También se han añadido genes de resistencia a antibióticos.

Una modificación fundamental es la adición de un sitio de clonación múltiple ('Multi Cloning Site" o MCS), que contiene sitios de restricción únicos y facilitan la inserción del ADN exógeno dentro del vector (Figura 1).



http://www.fermentas.com/techinfo/nucleicacids/mappuc1819.htm

***Figura 1**. Características comunes de los plásmidos. Se destacan los genes de resistencia a antibióticos, el origen de replicación modificado y el sitio de clonación múltiple (MCS).*

También poseen secuencias flanqueantes conocidas que se utilizan para poder secuenciar el fragmento incorporado cuando desconocemos su secuencia. Habitualmente este MCS se encuentra dentro de un gen de selección como el LAC-Z, que os va a permitir detectar que bacterias han incorporado el vector con el inserto (Figura 1).

El **tamaño de inserto** que habitualmente pueden contener los plásmidos varía entre **1 a 10 kb**, llegando en ocasiones a 15 kb. Normalmente estos vectores se introducen en las bacterias mediante transformación, utilizando métodos de choque térmico o electroporación. Posteriormente se crecen las bacterias y se seleccionan las adecuadas, normalmente por color, donde los clones positivos serán los que tengan color blanco. Para ver en más detalle el proceso podéis visualizar el siguiente video (https://youtu.be/IzBDO_YFNW4)

4.3.2 vectores basados en el fago lambda

Como su propio nombre indica, estos vectores se basan en la modificación del genoma del bacteriófago lambda. El genoma de este bacteriófago es de 48 kb y presenta en sus extremos una secuencia de 12 pb desemparejadas pero complementaria a la secuencia del extremo opuesto. A estos extremos se les conoce como **extremos cohesivos o extremos COS**.
Estos extremos cohesivos permiten al ADN del fago recircularizarse una vez este se introduce dentro de la bacteria. La maquinaria de este fago solo empaquetará en la cápsida ADN de un tamaño de aproximadamente 48 kb. Si es más grande o más pequeño no será empaquetado.

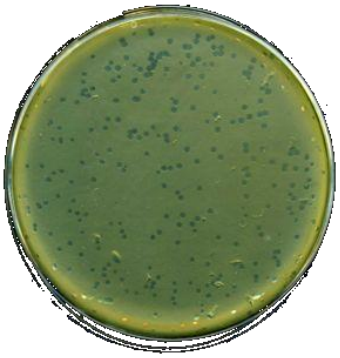


***Figura 2**. Placas de lisado de fagos lambda (círculos trasparentes) sobre césped bacteriano de E. coli.*

Los vectores comerciales eliminan todos los genes no esenciales para entrar en el ciclo lítico. Existen dos tipos de vectores de fagos λ, los de inserción y los de reemplazo

Los **vectores de inserción** pueden acomodar ADN con un tamaño de **5-11 kb**. En los **vectores de reemplazo**, puede insertarse un ADN de tamaño mayor de **8-24 kb**.
La introducción de los vectores en la bacteria se hace por infección, y se visualiza mediante la aparición de calvas en el cultivo bacteriano de *E.coli* (Figura 2).

4.3.3 Cósmidos

El siguiente vector de clonación es una mezcla entre plásmidos y fago λ. Los cósmidos son plásmidos que incorporan un segmento de ADN de bacteriófago λ que tiene el sitio final cohesivo (COS) que contiene elementos requeridos para empaquetar ADN en partículas λ. Normalmente se utiliza para clonar grandes fragmentos de ADN entre **28 y 45 kb**. Así, el vector se introduce en las bacterias mediante infección, pero una vez recircularizado dentro de la bacteria, se comporta como un plásmido (Figura 3).

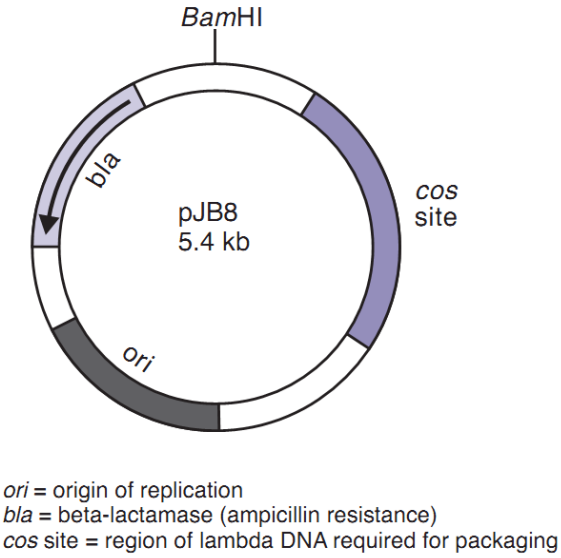


***Figura 3**. Características del cósmido. origen de replicación (ori), sitio cos (cos site), gen de resistencia a ampicilina (bla)*

4.3.4 BACs

Los vectores BACs ("Bacterial Artificial Chromosome") no son en realidad cromosomas bacterianos. Estos vectores basados en un plásmido de fertilidad funcional (o plásmido-F) de *E. coli* y son muy similares a los vectores plasmídicos. Sin embargo, el número de copias dentro de la bacteria es mucho más reducido, siendo de 1 a 2 copias por bacteria. La principal ventaja es su gran capacidad de carga de ADN exógeno, pudiendo acomodar tamaños de inserto de **150 a 300 kb**. Su manejo a efectos prácticos de laboratorio es el mismo que los de los plásmidos (Figura 4).

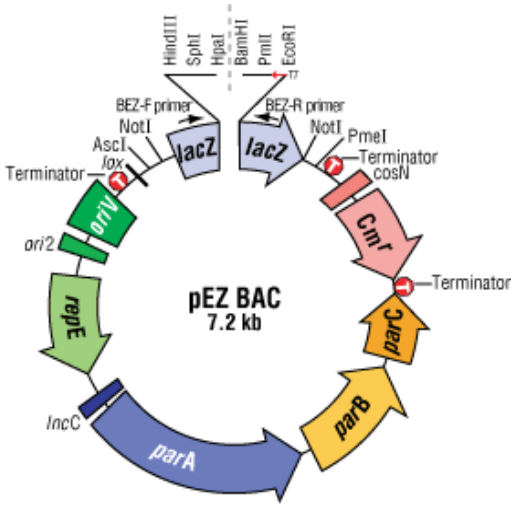


***Figura 4**. Características de un BAC. Se pueden apreciar los genes par (par A, parB y parC) que controlan el número de copias, el gen de resistencia a kanamicina (Cmr) y el "sitio de clonación múltiple" que interrumpe el gen LacZ.*

4.3.5 Elección del vector y del número de clones

La pregunta antes de iniciar a cualquier proyecto de secuenciación es: ¿Que vector debo usar para construir mi genoteca?. Aquí hay tres parámetros principales a tener en cuenta:

1. El tamaño total tiene mi genoma
2. La capacidad de carga del vector
3. El tamaño de la genoteca a construir

Estos tres parámetros se relacionan entre ellos. Por ejemplo, si utilizamos vectores plasmídicos para realizar una genoteca de un organismo con un tamaño de genoma muy grande, necesitaremos un tamaño de genoteca que va a resultar muy difícil de manejar.

Hay una fórmula que unifica todos estos parámetros, teniendo en cuenta que, por azar, no todos los clones van a ser diferentes, y que los fragmentos que contiene deben solapar. Por tanto, debemos aplicar probabilidades mediante la siguiente fórmula:

$$N = \frac{\ln(1 - P)}{\ln(1 - f)}$$

Donde:

N = Número de clones que necesito para construir mi genoteca

P= La probabilidad de encontrar un fragmento de copia única en la genoteca (habitualmente entre 95-99%)

f = Es el resultado de dividir el tamaño medio de inserto de los clones por el tamaño total del genoma.

En la tabla 1 podemos apreciar un ejemplo del tamaño de la genoteca (en número de clones) según el vector y el tamaño de genoma.

ORGANISMO	TAMAÑO GENOMA (pb)	VECTOR	TAMAÑO MEDIO DE INSERTO(kb)	P	Nº DE CLONES DE LA GENOTECA
E. coli	4 x 10^6	Plásmido	4	0.99	4600
		Lambda	18	0.99	1000
		Cósmido	40	0.99	458
		BAC	150	0.99	120
Humanos	3 x 10^9	Plásmido	4	0.99	3500000
		Lambda	18	0.99	770000
		Cosmido	40	0.99	350000
		BAC	150	0.99	92100

***Tabla 1**. Ejemplo de los distintos tamaños de genotecas dependiendo del tamaño de genoma y del tipo de vector utilizado para la construcción de la misma.*

Por ejemplo, si queremos tener una genoteca que contenga todo el genoma humano (tamaño de 3 Gb) representado dentro de plásmidos, vamos a necesitar seleccionar 3.500.000 de clones, lo que no es fácil de manejar ni de almacenar. De la misma manera, si queremos tener una genoteca que contenga todo el genoma de la bacteria *E. coli* (4 Mb) representada dentro de BACs, solo necesitaré 120 clones. Esto no sería de mucha utilidad, ya que, en vez de tener el genoma completo, lo tengo troceado en muy pocos clones.

Para poder estimar el **tamaño medio de inserto** de nuestros clones hay que seleccionar varios y digerirlos utilizando enzimas de restricción del MCS que liberen el inserto. Sumaremos los tamaños individuales de los insertos y los dividiremos por el número de clones testados.

Otro parámetro que conviene calcular es **la cobertura** que tendrá teóricamente nuestra genoteca una vez secuenciada. Por cobertura entendemos cuántas veces está representado cualquier fragmento de ADN en la genoteca. Imaginemos que buscáramos el gen cualquiera dentro de la genoteca. Si la cobertura es de 5X, entonces debería haber 5 clones que contengan el gen. Para calcularla utilizaremos la siguiente fórmula:

$$W=(N \times I)/G$$

Donde:

W = Cobertura de la genoteca
N = Número total de clones de la genoteca
I = tamaño medio de inserto de los clones
G = Tamaño total del genoma

4.3.6 Almacenamiento de la genoteca

Una vez determinado el número de clones que formarán nuestra genoteca, lo único que queda por hacer es almacenar de manera independiente estos clones. A cada uno de estos clones se los denomina clones independientes, ya que cada uno contendrá un fragmento del

ADN genómico diferente (aunque puede que haya algunos pocos con el mismo). Si estamos trabajando con bacterias, seleccionaremos el número de bacterias calculado y las creceremos en el medio adecuado en placas de 96 o 384 pocillos. De esta manera sabremos la posición de cada uno de los clones, aunque todavía no sabemos el orden de los mismos dentro del genoma. Finalmente, las placas con los clones se almacenan en un congelador a -80 ºC. En este punto nuestra genoteca está acabada y lista para ser usada (Figura 5).

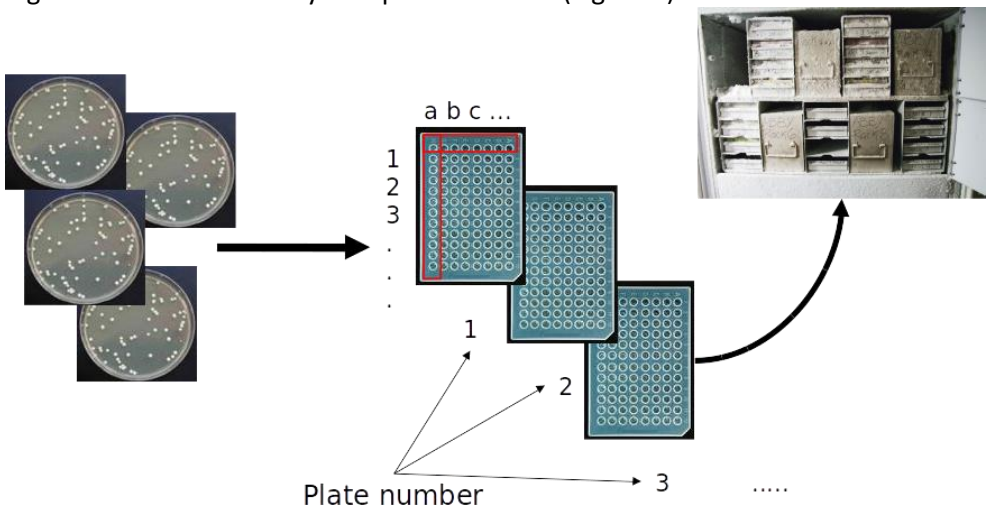


***Figura 5**. Esquema de la finalización de la genoteca. Los clones se depositan independientemente en los pocillos de las placas y finalmente se almacenan a -80ºC.*

Otro paso habitual es hacer una copia de la genoteca en filtro, lo que permite buscar clones de la genoteca que contiene fragmentos de ADN mediante hibridación. Para ello, y mediante robots especializados como el QBOT, se "imprimen" las bacterias en un macro-array de nylon. Cada uno de estos filtros puede contener más de 27.000 clones ordenados de manera que los clones positivos sean fácilmente identificables después de la hibridación (Figura 6).

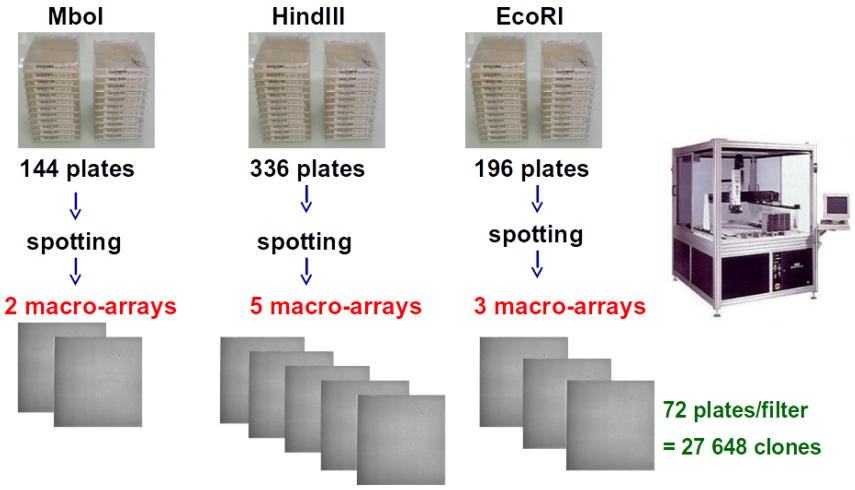


***Figura 6**. Esquema de la realización de la copia de la genoteca en filtros de nylon (macro-arrays).*

5 Cierre

A lo largo de este objeto de aprendizaje hemos visto cómo podemos realizar uno de los primeros pasos para la secuenciación de los genomas que es la construcción de genotecas utilizando vectores de clonación. En el siguiente esquema se resume el proceso a seguir (gráfico 1):

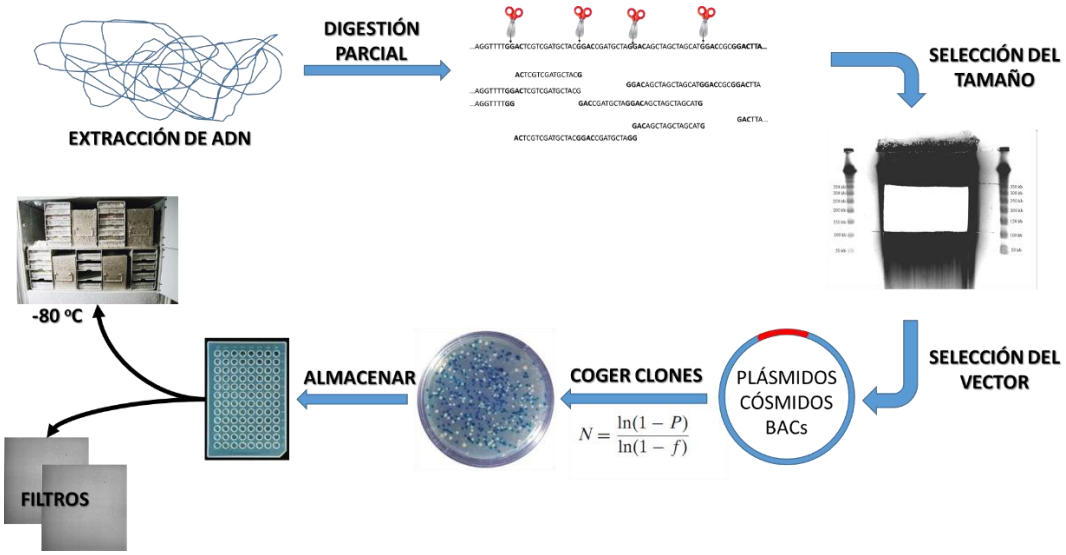


***Gráfico 1**. Estructura de la construcción de una genoteca*

Para comprobar qué realmente has aprendido intenta describir como harías una genoteca en diferentes especies con diferentes tamaños de genoma.

6 Bibliografía

6.1 Libros:

Brown, T. A. &Terence A. Brown.(2006) Gemomes 3. Tema 3 “Studying ADN” pág. 31-59.

Primrose, S., & Twyman, R. (2003). Principles of Genome Analysis and Genomics. Tema 3 “Subdividing the genome.” pág. 34-46.

6.2 Artículos de revisión

International Human Genome Sequencing Consortium, et al. “Initial sequencing and analysis of the human genome”. Nature, (2001), vol. 409, no 6822, p. 860.

Scheibye-Alsing, Karsten, et al. Sequence assembly. Computational biology and chemistry, 2009, vol. 33, no 2, p. 121-136.

Venter, J. Craig, et al. “The sequence of the human genome”. Science, 2001, vol. 291, no 5507, p. 1304-1351.