ANEXO I

FORMATOS DE ARCHIVO

FORMATO FASTA

Un archivo FASTA se reconoce por las extensiones ". fasta, .fna, .ffn, .faa, .frn".
Este formato se emplea para representar aquellas secuencias provenientes, por ejemplo, de procesos de secuenciación, en forma de ácidos nucleicos o péptidos. Visualmente, se pueden encontrar dos partes separadas: la primera línea o "*defline*" aparece señalada por el símbolo (">") y consta de la descripción de la secuencia en una única línea, siendo la primera palabra adyacente al símbolo (">") el identificador de la secuencia. El resto de las líneas representan la secuencia mediante códigos de una letra, códigos estándar de ácidos nucleicos y aminoácidos de la IUB/IUPAC, la Figura 9 ejemplifica el formato FASTA. No obstante, hay excepciones como la posibilidad de utilizar guiones para denotar un *gap – con longitud indeterminada*– o el uso de minúsculas.

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Figura 9. Ejemplo visual de un archivo formato FASTA (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION [NCBI], s.f.).

En el caso de los archivos denominados multi-FASTA, lo que aparece son una serie de secuencias en formato FASTA.

FORMATO FASTQ

FASTQ es un formato de archivo que integra lecturas provenientes de secuenciación, y la puntuación de la calidad, asociada a dichas lecturas, base por base.
Se representa con la extensión ".fastq", tiene dos variantes con respecto al estándar original, SANGER, que se denominan variantes Solexa e Illumnia. No obstante, la versión de SANGER de FASTQ es la que ha obtenido la hegemonía, siendo apoyada por diferentes herramientas bioinformáticas como son SSAHA2 (Ning *et al*., 2001), MAQ, Velvet, BWA (Li y Durbin, 2009) and BowTie que se emplean en procesos de ensamblaje y mapeo.
En un archivo FASTQ aparecen 4 líneas distintas por lectura. La primera línea comienza con una arroba ('@'), a la que le sigue un identificador de registro (ID). Adicionalmente pueden describirse otras características como la longitud de la secuencia, comentarios o identificaciones adicionales; siendo de esta forma un campo sin límite de caracteres y de libre escritura. La segunda, corresponde con la secuencia "problema", no existe limitación de caracteres y se suele emplear el uso de mayúsculas. Aunque, en algunos casos aparecen minúsculas, e incluso ambas. La tercera línea corresponde con el final de la línea de secuencia y el comienzo de la línea de calidad. Generalmente, esta línea contiene únicamente el símbolo "+". En los primeros archivos FASTQ la tercera línea contenía, además del símbolo antes mencionado, una repetición del ID plasmado en la primera línea que, por motivos de tamaño de archivo, ha quedado eliminada con el paso del tiempo. La última de las cuatro líneas presenta la calidad de secuenciación de cada uno de los nucleótidos que componen la secuencia. Por lo tanto, esta línea y la segunda deben tener la misma extensión. Los símbolos que marcan la calidad pertenecen al subconjunto de caracteres imprimibles ASCII. Entre ellos aparece la arroba, lo que implica que, en el caso de que una línea comience por "@", la herramienta que analice el archivo debe comprobar la longitud de dicha línea para no que no haya una confusión con la línea de cabecera (Cock *et al*., 2010).
Toda la explicación anterior se reduce visualmente en la Figura 10.

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Figura 10. Ejemplo visual de un archivo formato FASTQ (ILLUMINA, 2020a).

FORMATO YALM

YALM es un acrónimo recursivo de *YAML Ain't Markup Language* que en castellano se traduce como "YAML no es un lenguaje de marcado" (https://yaml.org/), y los archivos de este tipo se nombran mediante la extensión ".ylm". El objetivo del formato reside en facilitar la legibilidad de los archivos, al mismo tiempo que permitir su empleabilidad por la amplia mayoría de lenguajes de más usados en el mundo de la informática como Python, Java, JavaScript, Perl o Haskell. En resumen, YALM es capaz de mostrar cualquier tipo de dato mediante una combinación de datos escalares, listas y funciones resumen (función hash) (YAML AIN'T MARKUP LANGUAGE VERSION 1.2, 2009). Los archivos YAML se emplean con predominancia en archivos de configuración.

Como ejemplo se exponen los archivos YAML pertenecientes al pipeline presentado en la tesis Figura 11 y 12.

```yaml
name: nf-mtg_diamond-0.9.22
channels:
  - default
  - conda-forge
  - bioconda
dependencies:
  - python=2.7
  - bioconda::diamond=0.9.22
```

Figura 11. Archivo "diamond.ylm" que contiene las instrucciones para crear el ambiente conda que alberga el paquete de *software* DIAMOND (0.9.22).

```yaml
name: prokka1.14.6
channels:
  - conda-forge
  - bioconda
  - defaults
dependencies:
  - prokka=1.14.6
```

Figura 12. Archivo "prokka.ylm" que contiene las instrucciones para crear el ambiente conda que alberga el paquete de *software* PROKKA (1.14.6).

FORMATO SAM/BAM

SAM o *Sequence Alignment/Map format* es un tipo de archivo que se encuentra delimitado por tabulaciones (*TAB-delimited text*) y se denota por la extensión ".sam". El esquema de este formato se divide en las líneas de cabecera que son opcionales y están precedidas por el símbolo "@", y las líneas de alineamiento. Estas últimas están compuestas por once campos obligatorios que se detallan en la Tabla 5 y se representan en la Figura 13.

Además, existen otros campos opcionales que pueden aparecen en un archivo SAM que pueden aparecer en cualquier orden.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Tabla 5. Descripción de los campos obligatorios que aparecen en un archivo con formato SAM.
[11]Los nombres de las secuencias de referencia pueden contener cualquier carácter ASCII imprimible, a excepción de algunos caracteres de puntuación, y no pueden empezar por "*" o "=" (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021).

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Figura 13. Ejemplo visual de un archivo formato SAM (THE SAM/BAM FORMAT SPECIFICATION WORKING GROUP, 2021).

Por otro lado, el formato BAM se considera la versión binaria comprimida del formato SAM. Como ventaja, tiene la propiedad de que está diseñado para ser fácilmente comprimible, pero es más difícil de procesar y menos legible que los archivos SAM. Los archivos BAM también poseen una cabecera con información como el nombre de la muestra o la longitud, y una parte de alineamientos que contiene seis campos entre los que se encuentran el número de lecturas de cada muestra (RG), la calidad de alineamiento single-end y paired-end (SM/AS) o el nombre de la etiqueta del amplicón (XN) (ILLUMINA, s.f. a).

FORMATO GTF

Los archivos con extensión ".gtf", en inglés *General Transfer Format* , son el output típico de anotación que presenta la siguiente estructura (Figura 14) de nueve campos o columnas separadas por tabulador:

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pse
udogene";
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gene_sourc e "havana"; gene_bi
otype "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-002"; transcript_source "havana";
```

Figura 14. Ejemplo visual de un archivo formato GTF (ENSEMBL, s.f.).

La primera columna corresponde al cromosoma donde se encuentra el gen en cuestión. La segunda muestra la fuente de la que proviene la anotación (en este caso Ensembl); la tercera caracteriza la secuencia anotada, pudiendo ser etiquetada como: "start_codon", "CDS" o "stop_codon". Las dos columnas adyacentes marcan el inicio y el final de la secuencia problema, respectivamente. La siguiente columna, se denomina de score y se representa con un punto que "indica la posible presencia de señal del point floating value" (CRG.EU, sf).
La séptima columna marca la hebra de DNA que se transcribe para codificar dicha secuencia. Es decir, las opciones son "-" o "+". La penúltima define el marco de lectura (ORF) por el que se

comienza a transcribir – 1, 2 o 3 – y la última, proporciona información sobre los identificadores de los genes y sus características (separadas por ";"), puede incluir su "E-value".

FORMATO GFF

El formato GFF, en inglés *gene-finding format* o *generic feature format*, es un archivo con extensión ".gff" creado con el objetivo de identificar y describir genes, sus características y secuencias proteicas.

La estructura de este tipo de archivos es muy similar a la de los GTF: tiene, del mismo modo, 9 campos separados por tabulador, que coinciden exactamente hasta la columna 7. Se diferencian solamente en el último campo (columna 9) que contiene los atributos de la secuencia, como se puede apreciar en la Figura 15.

Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

Figura 15. Ejemplo visual de un archivo formato GFF (ENSEMBL, s.f.).

ANEXO II

CONCEPTOS

SECUENCIACIÓN *SINGLE-END*

La secuenciación *single-end* se refiere al proceso en el que se secuencia desde un único extremo del fragmento "problema". Este tipo de secuenciación es la más sencilla.

SECUENCIACIÓN CON LECTURAS PAREADAS

Secuenciación empleando la estrategia de leer ambos extremos de la lectura al mismo tiempo, siendo esta técnica más compleja que la *single-end* y genera bibliotecas de alta calidad. La mayoría de las técnicas de secuenciación de nueva generación (NGS) obtienen lecturas pareadas, existen dos estrategias que conforman este tipo de secuenciación: mate pairs y paired-end (ILLUMINA, 2020b).

ESTRATEGIA *MATE PAIRS*

La estrategia *mate pair* consiste en fragmentar las lecturas de DNA en fragmentos largos (>600 bp, hasta 5 kb). Estos se circularizan gracias a la biotinilización de sus extremos, y se re-fragmentan en fragmentos de entre 200 y 600 pb. Se purifican aquellos que tienen biotina – los extremos- por afinidad. Posteriormente se produce la ligación de los adaptadores de secuenciación y se secuencia, como se detalla en la Figura 16.

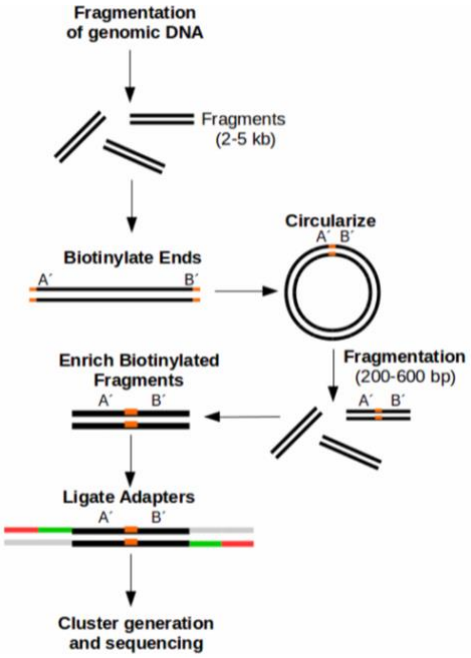

Figura 16. Esquema de la estrategia mate pair (ECSEQ BIOINFORMATICS, 2017).

Una ventaja de esta estrategia es que es capaz de cubrir rangos de mayor tamaño (ILLUMINA, s.f. b).

ESTRATEGIA *PAIRED END*

La estrategia *paired end* es sencilla, como se refleja en la Figura 17, y consiste en fragmentar en segmentos de un tamaño pequeño (<300pb) las lecturas a secuenciar; y posteriormente, se lee desde amos extremos del fragmento. El resultado es la posibilidad de cubrir tamaños de inserto más pequeños.

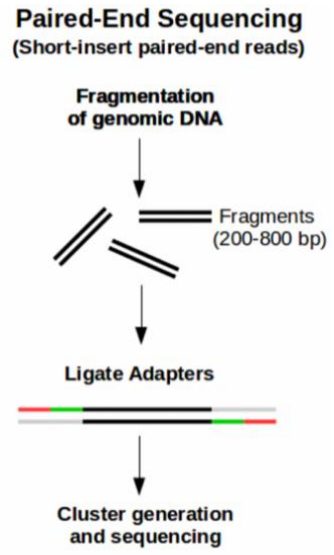

Figura 17. Esquema de la estrategia *paired end* (ECSEQ BIOINFORMATICS, 2017).

***SINGLE CELL SEQUENCING* (SCS)/ *SINGLE MOLECULE SEQUENCING* (SMS)**
La secuenciación de célula única o de células individuales en un tipo de secuenciación que abarca una variedad de tecnologías, y se emplea para estudiar células en solitario. A consecuencia, estas técnicas tienen una mayor resolución y se enmarcan en los métodos denominados HTS (*High Throughput Sequencing*) o secuenciación de alto rendimiento. La secuenciación por síntesis (SBS) y la secuenciación a tiempo real (SMRT) son tipos de SCS (Eberwine *et al*., 2014).

***SEQUENCING BY SYNTHESIS* (SBS)**
La secuenciación por síntesis es un tipo de secuenciación masiva (NGS) que fue lanzado al mercado por Illumina. La peculiaridad de la SBS es el uso de la amplificación clonal *in vitro* gracias a una PCR en puente. El enfoque de la secuenciación por sínstesis se emplea para efectuar la SCS (ILLUMINA, s.f. c).

***SINGLE MOLECULE REAL TIME SECUENCING* (SMRT)**
En castellano secuenciación a tiempo real de una única molécula de DNA, este tipo de secuenciación de tercera generación (TSG) fue desarrollada por Pacific Biosciences. Su funcionamiento consiste en el empleo de un chip – que presenta una estructura ZWM (guía de onda "modo cero") – en el que se distribuyen moléculas de DNA polimerasa de forma individual y sDNA (DNA de cadena sencilla). En otras palabras, el chip está dividido en pozos microscópicos en los que se aloja una DNA polimerasa activa y un sDNA molde. Dicha tecnología la secuenciación se acelera y se obtienen fragmentos bastante largos (>7kb) (Eid *et al*., 2009).

***POLYMERASE CHAIN REACTION* (PCR)**
La PCR o reacción en cadena de la polimerasa es una técnica de laboratorio, perteneciente al campo de la biología molecular, cuyo objetivo es la amplificación de secuencias de DNA con fines de investigación y diagnósticos. El procedimiento requiere de diferentes reactivos como son primers (cebadores), una solución tampón (buffer), DNA polimerasa (enzima que produce la amplificación) y sus cofactores, desoxirribonucleótidos-trifosfato (dNTP) y, por supuesto, la muestra de DNA molde. El resultado son amplificar el número de copias de la hebra molde para otros estudios (Saiki *et al*., 1998).

***OPEN READING FRAME* (ORF)**
Un ORF, en castellano marco de lectura abierto, es la secuencia que se enmarca entre un codón de inicio (en eucariotas predomina el codón "AUG") y un codón de terminación. Es decir, es un

fragmento que al pasar por el proceso de traducción no posee codones de terminación. Los ORFs suelen ser exones que forman parte de un gen. En concreto existen seis sentidos en los que aparecn marcos abiertos de lectura (+1, +2, +3, -1,-2,-3), lo que se debe a la naturaleza de la lectura del código genético que ocurre en codones de 3 bp (NHGRI, s.f. a).

cDNA
El cDNA o DNA complementario es una molécula sintetizada – artificialmente – por medio de la enzima transcriptasa reversa y que corresponde a una copia proveniente del RNA mensajero. Como consecuencia, estos transcritos no tienen intrones.

***EXPRESSED SEQUENCE TAG* (EST)**
En castellano marcador de secuencia expresada, "es una pequeña sub-secuencia de una secuencia nucleotídica transcripta (codificante de una proteína o no). Se pueden usar para identificar genes que se transcriben y en el descubrimiento de genes, y para determinación de secuencias" (Adams *et al.*, 1991).
Estas "etiquetas" se producen mediante la secuenciación de un ARNm clonado, por lo que se consideran de baja calidad. GenBank (base de datos pública) alberga más de 50 millones de ESTs diferentes.

ADAPTADORES
En el contexto de las nuevas técnicas secuenciación masiva (NGS) los adaptadores se definen como secuencias consenso que reconocen los *primers* (cebadores) que se emplean para la amplificación y la secuenciación en sí. En concreto son fragmentos/secuencias de doble cadena de DNA y son específicos para cada plataforma de secuenciación (Rehm *et al*., 2013).

RNA-seq
RNA-seq también conocida como *Whole Transcriptome Shotgun Sequencing* o, en castellano, Secuenciación del Transcriptoma para Clonación al Azar, es una técnica de secuenciación que permite secuenciar el producto cDNA de RNA mensajeros de una muestra.
Actualmente, esta técnica se usa para identificar los transcriptos de una célula, para experimentos de expresión diferencial o para conocer los modelos e isoformas de las proteínas (Chu y Corey, 2012).

CONTIG
Conjunto de lecturas de DNA que se solapan, gracias a las cuáles se pueden conseguir secuencias más largas. Los contigs se obtienen tras el proceso de ensamblaje y representan una región consenso del genoma (NHGRI, s.f. b).

SCAFFOLD
El término tiene su análogo en castellano, denominándose en esta lengua andamio. No obstante, es comúnmente empleada la palabra *scaffold* en el campo de las ciencias ómicas.
En concreto los *scaffold* son un conjunto de *contigs* ordenados, pero en este caso no tienen porqué solaparse entre sí, siendo normal la aparición de huecos entre unos y otros (NCBI, 2021).

BINNING
El *data binning* consiste en un proceso de pre-procesamiento de los datos a analizar, con el objetivo de facilitar el análisis posterior. En metagenómica se refiere a la agrupación de lecturas o contigs en bins para su posterior asignación a un genoma individual. Dependiendo el tipo de enfoque, se da el binning antes o después del alineamiento: si se produce antes el objetivo del proceso son las lecturas, gracias al agrupamiento se reducen las necesidades computacionales, pero el resultado del alineamiento se vuelve menos fiable por la pequeña longitud de los bins.
El binning se puede dar con diferentes estrategias: agrupación por identidad (similitud), agrupación por composición, o ambas (Sangwan *et al*., 2016).

FM-INDEX

El índice FM es un índice que "aprovecha la relación entre la transformada Burrows-Wheeler y la estructura de datos de matriz de sufijos" (Ferragina y Manzini, 2005). El objetivo de esta herramienta es conseguir pasar de un texto comprimido a un índice que represente todo el contenido del archivo gracias a la Transformación Burrows-Wheeler (BWT).

K-MER

Un k-mer es una subcadena de longitud k que se encuentra contenida dentro de una secuencia biológica.

GRAFOS DE BRUJIN

"Un grafo De Bruijin se describe como un grafo dirigido que representa los solapamientos de longitud (n-1) entre todas las palabras de longitud n con un alfabeto dado" (Bruijn, 1946), Figura 18.

En el campo de la bioinformática se emplean estos grafos en herramientas de ensamblado. Las lecturas para ensamblar se fragmentan en "palabras" de longitud n que se solapan mediante estos grafos.

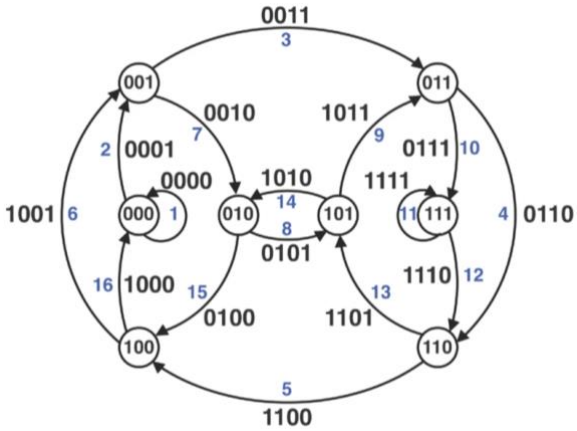

Figura 18. Ejemplo visual del grafo de Bruijn. El grafo de Bruijn B para n = 4 y un alfabeto de dos caracteres compuesto por los dígitos 0 y 1. Este grafo tiene un ciclo euleriano (un camino euleriano se define como el camino que pasa por cada arista una vez) porque cada nodo tiene un grado igual a 2. Siguiendo las aristas numeradas en azul en orden del 1 al 16 se traza un ciclo euleriano 0000, 0001, 0011, 0110, 1100, 1001, 0010, 0101, 1011, 0111, 1111, 1110, 1101, 1010, 0100, 1000. Al registrar el primer carácter (en negrita) de cada etiqueta de borde deletrea la supercadena cíclica 0000110010111101 (Compeau, P. *et al*., 2011).

GRAFOS EMPAREJADOS DE BRUJIN (PDBG)

"Los grafos de Bruijn emparejados son una generalización de los grafos de Brujin mediante la incorporación de la información que proporciona la secuenciación *mate pair*" (Medvedev *et al*., 2011). La Figura 19 aclara las diferencias entre los grafos de Bruijn y los PDGB.

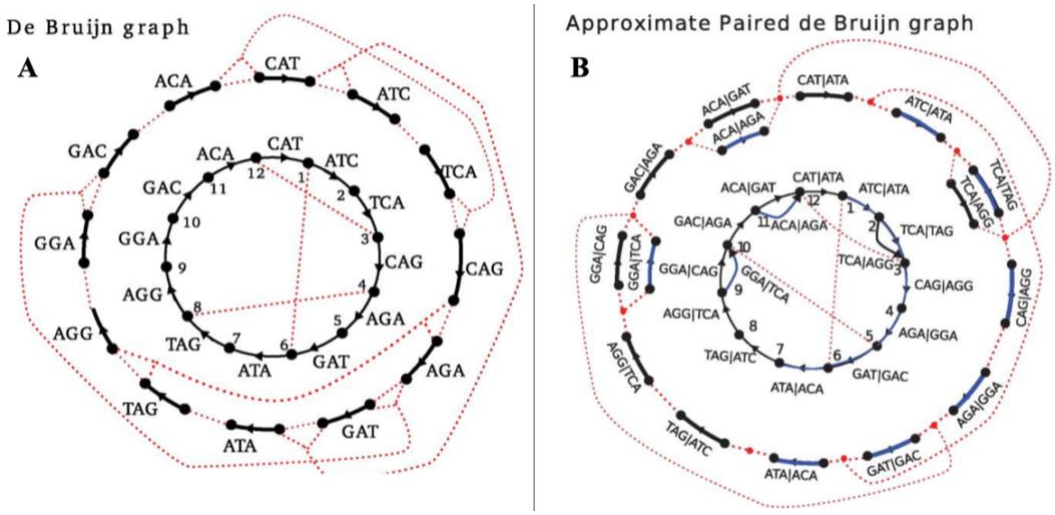

Figura 19. Comparación entre un grafo de Bruijn (A) y un grafo emparejado de Bruijn (B). En la imagen A, el círculo exterior muestra un borde negro separado para cada k-mer (k = k-mers de 3 nucleótidos de longitud). Las líneas rojas punteadas indican los vértices que se solapan. El círculo interior muestra el resultado de aplicar algunas de las colas. En la imagen B se muestra un posible espectro de cobertura en el círculo exterior, con aristas negras para los elementos con distancia entre *mate pairs* de 6 y aristas azules para la distancia de 5 (Medvedev *et al*., 2011).

GRAFO ACÍCLICO DIRIGIDO (DAG)

Un grafo acíclico dirigido es un grafo dirigido pero que no contiene ciclos directos; es decir, no existe una ruta directa que empiece y termine en un mismo vértice del grafo como queda representado en la Figura 20 (Weisstein, s.f.).

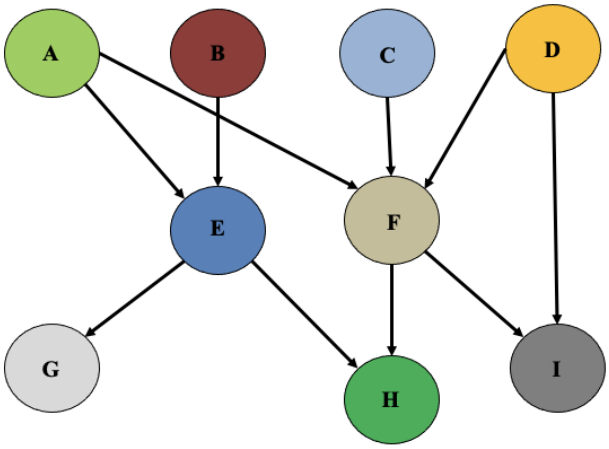

Figura 20. Representación de un DAG, los nodos se han nombrado aleatoriamente.