Formatos de archivo utilizados en secuenciación de nueva generación: Una revisión bibliográfica

Ángel Canal-Alonso[1], Pedro Jiménez[1] and Noelia Egido[1], Javier Prieto[1], Juan Manuel Corchado[1]

[1] Departamento de Bioinformática y Biología Computacional, AIR Institute, Carbajosa de la Sagrada, España

E-mail: acanal@air-institute.com

Resumen

La secuenciación de nueva generación (NGS) ha revolucionado el campo de la genómica, permitiendo una mirada detallada y precisa del ADN. A medida que esta tecnología avanzó, surgió la necesidad de formatos de archivo estandarizados para representar, analizar y almacenar los vastos conjuntos de datos producidos. En este artículo, revisamos los formatos de archivo clave utilizados en NGS: FASTA, FASTQ, BED, GFF y VCF.

El formato FASTA, uno de los más antiguos, proporciona una representación básica de secuencias genómicas y proteicas, identificables por encabezados únicos. FASTQ es esencial para NGS, ya que almacena tanto la secuencia como la información de calidad asociada. BED ofrece una representación tabular de loci genómicos, mientras que GFF detalla la localización y estructura de características genómicas en secuencias de referencia. Finalmente, VCF ha emergido como el estándar predominante para documentar variantes genéticas, desde simples SNPs hasta variantes estructurales complejas.

La adopción y adaptación de estos formatos han sido fundamentales para el progreso en la bioinformática y la genómica. Proporcionan una base sobre la cual se construyen análisis sofisticados, desde el descubrimiento de genes y la predicción de funciones, hasta la identificación de variantes asociadas con enfermedades. Con una comprensión clara de estos formatos, los investigadores y profesionales están mejor equipados para aprovechar el poder y el potencial de la secuenciación de nueva generación..

Palabras Clave: Next-Generation sequencing, File format, Data sharing

---

Introducción

El ADN, ácido desoxirribonucleico, es la molécula portadora de la información genética que define y regula las características de todos los seres vivos. Está compuesto por una secuencia específica de nucleótidos, que son las unidades básicas del ADN. Cada nucleótido consiste en una de las cuatro bases: adenina (A), timina (T), citosina (C) y guanina (G). El orden o secuencia en que estas bases aparecen en la molécula de ADN es lo que codifica la información genética. La determinación del orden exacto de los nucleótidos en un fragmento de ADN es lo que se conoce como secuenciación de ADN.

La secuenciación del ADN ha jugado un papel fundamental en la biología y la medicina desde su desarrollo inicial en la década de 1970. Durante décadas, el método de secuenciación por terminación de cadena de Sanger fue el enfoque predominante. Sin embargo, a medida que la demanda de secuenciación creció, surgieron nuevas técnicas y tecnologías que permitieron secuenciar más rápidamente y a menor costo.

En este contexto, la secuenciación de nueva generación (NGS, por sus siglas en inglés) ha emergido en el siglo XXI como una técnica revolucionaria. A diferencia de la secuenciación de Sanger, que analiza un fragmento de ADN a la vez, NGS permite la secuenciación masiva de millones de fragmentos de ADN en paralelo. Esta capacidad de "alta

capacidad" ha abierto puertas a numerosas aplicaciones, desde la genómica del cáncer hasta la microbiología ambiental y la genética evolutiva.

Con el auge de la NGS, también ha surgido la necesidad de manejar, analizar e interpretar el vasto volumen de datos generados. Esto ha llevado al desarrollo de diversos formatos de archivo específicos para garantizar el almacenamiento eficiente, el análisis preciso y la compartición de datos genómicos. En esta revisión, examinaremos detenidamente estos formatos, proporcionando una visión comprensiva de su estructura, utilidad y software asociado.

Tecnologías de secuenciación en NGS

Desde el surgimiento de la NGS, han aparecido varias tecnologías que han revolucionado el ámbito de la secuenciación genómica. A continuación, presentamos un resumen de las principales tecnologías utilizadas en NGS:

1. Illumina (Secuenciación por síntesis):
- Principio: Utiliza la detección de nucleótidos fluorescentes incorporados durante la síntesis de ADN.
- Características:
- Alta precisión.
- Genera grandes volúmenes de lecturas cortas.
- Ampliamente utilizado en genómica, transcriptómica y epigenómica.

2. Ion Torrent (Secuenciación por semiconductor):
- Principio: Detecta protones liberados durante la incorporación de nucleótidos en la síntesis de ADN.
- Características:
- No requiere fluorescencia.
- Capaz de generar lecturas de longitud media.
- Adecuado para genotipado y secuenciación dirigida.

3. PacBio (Secuenciación de moléculas individuales, en tiempo real):
- Principio: Observa la incorporación de nucleótidos en tiempo real en moléculas individuales.
- Características:
- Genera lecturas extremadamente largas.
- Alta tasa de errores, pero errores aleatorios que pueden ser corregidos con coberturas adecuadas.
- Ideal para ensamblaje de genomas y detección de variantes estructurales.

4. Oxford Nanopore Technologies (ONT):
- Principio: Detecta cambios en la corriente eléctrica mientras una molécula de ADN pasa a través de un poro nanométrico.
- Características:
- Capaz de producir las lecturas más largas de todas las tecnologías.
- Portabilidad (por ejemplo, el dispositivo MinION).
- Aplicaciones en genómica de campo, monitoreo ambiental y diagnóstico en tiempo real.
- Mayor tasa de errores en comparación con otras tecnologías, pero la longitud de las lecturas y las mejoras en software ayudan en la corrección de errores.

5. 10X Genomics:
- Principio: Combina la microfluídica para particionar células o fragmentos de ADN y luego aplica la secuenciación Illumina.
- Características:
- Proporciona información sobre las fases y estructuras cromosómicas.
- Utilizado para genómica, transcriptómica a nivel de célula única y análisis epigenómicos.

La elección de la tecnología NGS a utilizar dependerá del objetivo del estudio, el tipo de información requerida y el presupuesto disponible. Cada tecnología tiene sus propias ventajas, desventajas y nichos de aplicación. Lo que es constante es la rápida evolución y mejora de estas tecnologías, que continúan ampliando las fronteras de lo que es posible en la investigación genómica.

La información en NGS

Introducción a los formatos de archivo en NGS

En el entorno de la secuenciación de nueva generación (NGS), la producción de datos genómicos ha crecido de manera exponencial. Estos datos, caracterizados por su volumen y complejidad, requieren de formatos de archivo especializados que faciliten su almacenamiento, análisis e interpretación. Estos formatos no solo sirven como contenedores de información, sino que también establecen estándares que aseguran la coherencia, interoperabilidad y reproducibilidad de los análisis realizados en distintas plataformas y software.

Desde las etapas iniciales, donde se almacenan las secuencias crudas y sus calidades, hasta las fases posteriores, donde se registran alineaciones, variantes genómicas o anotaciones funcionales, existe una variedad de formatos diseñados específicamente para cada tipo de dato y propósito.

Algunos de estos formatos se han convertido en estándares de facto en la comunidad científica debido a su versatilidad y amplia adopción. Es esencial para cualquier profesional en genómica familiarizarse con estos formatos, ya que forman la

base sobre la cual se construyen los análisis y las interpretaciones en la era moderna de la genómica.

En las siguientes secciones, exploraremos en detalle los principales formatos de archivo utilizados en NGS, brindando una visión integral de su estructura, funcionalidad y aplicaciones en el mundo de la secuenciación genómica

FASTA:

El formato FASTA, a menudo reconocido por su extensión `.fasta` o `.fa`, es uno de los formatos más antiguos y ampliamente utilizados en bioinformática. Aunque en el contexto actual de NGS, el formato FASTQ suele ser más comúnmente asociado con datos de secuenciación cruda, el formato FASTA sigue siendo fundamental en muchos análisis genómicos.

Historia:

El formato FASTA debe su nombre al programa FASTA, que fue desarrollado en la década de 1980 por David J. Lipman y William R. Pearson. Originalmente, este programa fue diseñado para realizar búsquedas de secuencias de proteínas en bases de datos. A medida que el programa ganó popularidad para la alineación de secuencias y la búsqueda en bases de datos, el formato de archivo asociado, que era simple y eficaz para representar secuencias de nucleótidos o aminoácidos, también se popularizó.

La estructura simple del formato FASTA fue una de las razones principales de su amplia adopción. Una secuencia en formato FASTA comienza con una línea de descripción, precedida por el símbolo “>”, seguida de líneas que representan la secuencia propiamente dicha. Esta simplicidad permitió que el formato fuese fácilmente legible tanto por humanos como por computadoras.

Con la expansión de la genómica y la bioinformática en las décadas siguientes, el formato FASTA se convirtió en el estándar de facto para representar secuencias de ADN, ARN y proteínas. Las bases de datos genómicas más grandes, como GenBank, EMBL y la Protein Data Bank, adoptaron el formato FASTA para la distribución y búsqueda de secuencias, solidificando aún más su posición en la comunidad científica.

Aunque el paisaje de la secuenciación ha evolucionado enormemente desde los días del programa FASTA original, y nuevos formatos como FASTQ han surgido para satisfacer las necesidades específicas de NGS, el formato FASTA sigue siendo esencial. Se utiliza para representar genomas de referencia, secuencias de genes individuales, conjuntos de proteínas y mucho más. Su legado y relevancia perduran en la era moderna de la genómica, subrayando la importancia de las soluciones simples y eficaces en la ciencia..

El formato FASTA es apreciado por su simplicidad y claridad. Aunque su estructura básica ha permanecido relativamente constante a lo largo del tiempo, ha demostrado ser flexible y adaptable a diversas aplicaciones en genómica y proteómica.

Estructura básica:

1. Línea de encabezado: Cada entrada en un archivo FASTA comienza con una línea de encabezado, que se distingue por el símbolo “>” al inicio. Esta línea de encabezado proporciona una descripción o identificación de la secuencia. A menudo, esta línea contiene un identificador único para la secuencia, y puede incluir información adicional, como la fuente del organismo, el número de acceso de la base de datos, entre otros.

Ejemplo:
```
>NM_001301717.2 Homo sapiens actin alpha cardiac muscle 1 (ACTC1), transcript variant 2, mRNA
```

2. Secuencia: Después de la línea de encabezado, sigue la secuencia propiamente dicha, escrita en líneas consecutivas. En el caso de secuencias de ADN o ARN, la secuencia estará compuesta por las letras que representan los nucleótidos (A, T, C, G, U). Para las secuencias de proteínas, se utilizan letras que representan los aminoácidos.

Ejemplo:
```
ATGGAGACAGAAGTCTTCACTGCTGAGGAGGAGGA
AGAGGAAGCAGATGGAGAAGAAGCTG

TCAACATCAAGTCTGACCTAATCACTGAGAAGCTCG
GGAAGGAAACTGACCGAAGGCAAGA
```

Consideraciones adicionales:

- Un archivo FASTA puede contener múltiples secuencias. Cada secuencia se demarcará por su propia línea de encabezado seguida por su secuencia.

- La longitud de cada línea de la secuencia suele estar restringida a un número específico de caracteres (por ejemplo, 60 o 80) por razones de legibilidad, aunque esto no es una regla estricta.
- No hay restricción sobre el número total de caracteres en la secuencia, por lo que un archivo FASTA puede representar desde pequeños fragmentos de ADN hasta genomas completos.
- Aunque la estructura básica del formato FASTA es simple, las convenciones específicas para la línea de encabezado pueden variar según la base de datos o el recurso de donde provenga el archivo.

En resumen, el formato FASTA proporciona una representación simple y clara de secuencias biológicas. Su estructura intuitiva ha facilitado que sea ampliamente adoptado y utilizado en diversas aplicaciones, desde búsquedas de homología hasta estudios de evolución molecular. Aunque carece de la capacidad de almacenar información de calidad que se requiere en la NGS moderna (algo que el formato FASTQ proporciona), el FASTA sigue siendo un pilar en bioinformática y genómica.

FASTQ:

Historia

A medida que la bioinformática y la genómica avanzaban en la primera década del siglo XXI, también lo hicieron las tecnologías de secuenciación. La llegada de la secuenciación de nueva generación (NGS) marcó un punto de inflexión, no solo en términos de la cantidad de datos producidos sino también en la calidad y la resolución de esos datos. Este cambio en el paisaje tecnológico exigía una evolución en los formatos de datos. Es en este contexto donde el formato FASTQ nació y se consolidó.

El formato FASTQ fue creado inicialmente para adaptarse a las demandas de las primeras plataformas de secuenciación de nueva generación, como la tecnología Solexa de Illumina. Estos nuevos métodos generaban, por primera vez, lecturas individuales con asociaciones de calidad para cada base secuenciada, algo que las técnicas anteriores, como la secuenciación Sanger, no producían en la misma escala.

El nombre "FASTQ" se deriva de la combinación de "FASTA", reflejando su similitud estructural con el formato de secuencia original, y "phred quality scores", que son las calificaciones que se utilizan para representar la calidad de las bases. Estas puntuaciones de calidad, originalmente desarrolladas para la secuenciación Sanger en los años 90, se adaptaron y expandieron para su uso en la NGS.

Aunque el formato FASTQ empezó como una solución específica para las tecnologías Solexa, rápidamente fue adoptado y adaptado por otras plataformas de NGS. Sin embargo, este rápido crecimiento llevó a ciertas inconsistencias, especialmente en la codificación de las puntuaciones de calidad. Por ejemplo, mientras que la plataforma Solexa utilizaba un rango de codificación de calidad diferente, las versiones posteriores de Illumina y otras tecnologías adoptaron el estándar Phred +33, que es el mismo sistema utilizado en la secuenciación Sanger.

Esta variabilidad en la codificación de calidad y otras particularidades, si bien reflejaba la rápida evolución y competencia entre tecnologías de secuenciación, también presentó desafíos para los bioinformáticos y desarrolladores de software. Se hizo evidente la necesidad de estandarizar y clarificar el formato. Afortunadamente, con el tiempo y el esfuerzo de la comunidad, se ha alcanzado un consenso más amplio sobre cómo se debe utilizar el formato FASTQ, aunque es crucial que los usuarios sean conscientes de la versión específica y de las peculiaridades asociadas a la plataforma de secuenciación que están utilizando.

Desde sus humildes inicios adaptándose a una nueva era de tecnología de secuenciación, el formato FASTQ ha crecido para convertirse en un pilar en el mundo de la NGS. Su historia refleja la naturaleza dinámica y a menudo desafiante del campo en rápido movimiento de la genómica, y subraya la importancia de la adaptabilidad y colaboración en la ciencia.

Funcionamiento y Estructura

El formato FASTQ es esencial en el contexto de la secuenciación de nueva generación (NGS). A diferencia del formato FASTA, que almacena solamente secuencias, el FASTQ incluye información sobre la calidad de cada base secuenciada. Esta información de calidad es crucial para determinar la confiabilidad de las lecturas y para llevar a cabo análisis posteriores de NGS, como el alineamiento y la identificación de variantes.

Estructura básica:

El archivo FASTQ consta de cuatro líneas por registro:

1. Línea de encabezado: Comienza con el símbolo “@” y suele contener información sobre el instrumento de

secuenciación, el número de corrida, el identificador único de la lectura, entre otros.

Ejemplo:
```
@SEQ_ID MISEQ:7:000000000-A4K08:1:1101:12469:2217 1:N:0:
```

2. Secuencia: La línea siguiente presenta la secuencia de nucleótidos en sí, al igual que en un archivo FASTA.

Ejemplo:
```
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

3. Separador: Una línea que solo contiene el símbolo "+". Ocasionalmente, esta línea puede repetir la información de la línea de encabezado, aunque en la mayoría de los casos simplemente se presenta el símbolo "+" por sí solo.

Ejemplo:
```
+
```

4. Calidades: Esta línea codifica la calidad de cada base en la secuencia utilizando caracteres ASCII. Cada carácter representa un valor de calidad, que indica la probabilidad de que una base en particular haya sido secuenciada de manera incorrecta. Los programas de análisis de NGS utilizan estos valores para ponderar la confiabilidad de las lecturas.

Ejemplo:
‘’’ !"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65 ```

- El formato FASTQ permite representar lecturas de secuenciación de diferentes longitudes. Por lo tanto, no todas las secuencias en un archivo FASTQ tendrán la misma longitud.
- El sistema de codificación de calidad ha variado entre las versiones y las plataformas de secuenciación. Las versiones más comunes son el código ASCII 33 (Sanger) y el código ASCII 64 (Illumina 1.3+), y es vital saber cuál se está utilizando para interpretar correctamente los valores de calidad.
- Al igual que con FASTA, un archivo FASTQ puede contener muchos registros, es decir, múltiples grupos de estas cuatro líneas.
- Los archivos FASTQ pueden ser muy grandes, especialmente en estudios de secuenciación de alta cobertura, por lo que a menudo se comprimen utilizando herramientas como gzip.

En resumen, el formato FASTQ es fundamental en la era de la secuenciación de nueva generación. Al proporcionar información tanto de secuencia como de calidad, el FASTQ permite a los investigadores y bioinformáticos evaluar y analizar datos de secuenciación con un nivel de detalle sin precedentes. Su estructura, aunque un poco más compleja que la del FASTA, se ha convertido en un estándar en el campo de la NGS.

BED:

En el contexto de la genómica y la bioinformática, el formato BED (Browser Extensible Data) se ha establecido como un estándar para representar regiones de interés en genomas y otros conjuntos de datos relacionados con la secuencia. La simplicidad, versatilidad y capacidad extensible del formato BED lo ha hecho indispensable en diversas aplicaciones.

Orígenes del BED:

El formato BED fue inicialmente desarrollado por el equipo del UCSC Genome Browser, una herramienta online ampliamente utilizada para visualizar anotaciones y características genómicas en contextos genómicos de referencia. Dado que uno de los principales objetivos del Genome Browser es permitir a los usuarios explorar y visualizar regiones genómicas específicas junto con una variedad de datos y anotaciones asociadas, había una necesidad evidente de un formato que pudiera representar de manera eficiente estas regiones y anotaciones en el genoma.

El equipo de UCSC desarrolló el formato BED para facilitar la carga y visualización de datos personalizados en el navegador. Al hacerlo, crearon una estructura que no solo era adecuada para sus propias necesidades, sino que también se adaptaba bien a muchos otros usos en genómica.

Expansión y adopción en la comunidad:

Lo que distingue al formato BED es su estructura simple pero altamente extensible. En su forma más básica, un archivo BED necesita solo tres columnas: cromosoma, posición de

inicio y posición final. Sin embargo, el formato puede extenderse para incluir información adicional en columnas adicionales, como nombres de elementos, puntuaciones y dirección de la cadena, entre otros.

Dada esta flexibilidad, no es sorprendente que el formato BED haya sido rápidamente adoptado por la comunidad genómica más amplia. Se ha utilizado en una variedad de aplicaciones, desde la identificación de regiones reguladoras hasta la anotación de variantes genómicas. Además, muchos softwares y herramientas bioinformáticas, como BEDTools, se han desarrollado específicamente para trabajar con archivos BED, reforzando aún más su posición como un estándar en la genómica.

El formato BED es un ejemplo de cómo una solución desarrollada para una necesidad específica puede, con el tiempo, convertirse en un estándar de la industria. Desde su creación en el UCSC Genome Browser, ha evolucionado y se ha adaptado a una variedad de aplicaciones en genómica, y su legado subraya la importancia de la simplicidad y la extensibilidad en el diseño de formatos de datos.

BED: Estructura y Funcionamiento

El formato BED (Browser Extensible Data) es una forma sencilla y flexible de representar regiones en un genoma. A pesar de su simplicidad, es poderoso y ampliamente utilizado en bioinformática. A continuación, se describen la estructura básica y el funcionamiento del formato BED.

Estructura básica:

Un archivo BED consta de una serie de líneas, cada una de las cuales describe una región en particular del genoma. Las columnas en una línea BED están separadas por tabulaciones. El número de columnas puede variar, pero las primeras tres son esenciales y siempre deben estar presentes:

1. cromosoma (chrom): El nombre del cromosoma o secuencia de referencia. Por ejemplo: `chr1`, `chr2`, `chrX`, etc.

2. posición de inicio (chromStart): La posición de inicio de la región en el cromosoma. Es importante notar que BED utiliza una indexación basada en cero, lo que significa que la numeración comienza desde 0.

3. posición final (chromEnd): La posición final de la región en el cromosoma. A diferencia del inicio, esta posición es exclusiva, es decir, la base señalada por chromEnd no está incluida en la región BED.

Estos tres campos son obligatorios en cualquier archivo BED. Sin embargo, hay nueve campos adicionales opcionales que pueden estar presentes, extendiendo el formato:

4. name: El nombre de la región BED.
5. score: Una puntuación entre 0 y 1000. Se puede usar para almacenar cualquier valor numérico asociado con la región.
6. strand: La hebra del genoma. Puede ser '+' o '-'.
7. thickStart y thickEnd: Estos campos son útiles para representar regiones codificantes en genes.
8. itemRgb: Color para mostrar el elemento.
9. blockCount: Número de bloques (exones, por ejemplo) en la región.
10. blockSizes: Tamaño de los bloques.
11. blockStarts: Posiciones de inicio de los bloques.

Funcionamiento y aplicaciones típicas:

El formato BED se utiliza principalmente para definir y manipular conjuntos de coordenadas genómicas. Es especialmente útil para:

- Visualizar regiones genómicas en navegadores de genomas, como el UCSC Genome Browser.
- Anotar regiones de interés, como sitios de unión de proteínas o regiones diferencialmente metiladas.
- Manipular y analizar conjuntos de coordenadas, como intersecciones, uniones y complementos, especialmente usando herramientas como BEDTools.
- Definir regiones codificantes o no codificantes en genes.

Ejemplo:

Suponiendo que queremos representar un gen situado en el cromosoma 1, que comienza en la base 100 y termina en la base 500, con el nombre "MiGen", una puntuación de 960, en la hebra positiva, podría verse así:

```
chr1  99    500    MiGen  960    +
```

El formato BED es una herramienta esencial en genómica y bioinformática debido a su simplicidad y versatilidad. Aunque su estructura básica es sencilla, su diseño extensible le permite representar una amplia variedad de información genómica, haciendo de él una elección popular para muchas aplicaciones en el campo. Es crucial entender la indexación basada en cero y otros detalles específicos del formato para garantizar un uso correcto y una interpretación precisa de los datos.

GFF

El Formato General de Características (General Feature Format, GFF) ha sido una herramienta esencial en genómica y bioinformática durante las últimas décadas. Provee un mecanismo para representar la estructura de genes y otras anotaciones en genomas. Veamos cómo ha evolucionado el GFF y cómo ha servido a la comunidad científica desde su concepción.

Orígenes del GFF:

La primera versión de GFF fue desarrollada en el ámbito del proyecto WormBase, una base de datos dedicada a la anotación y análisis genómico del nematodo Caenorhabditis elegans, uno de los primeros organismos en ser completamente secuenciados. Con el auge de proyectos de secuenciación en la década de 1990, era evidente la necesidad de un formato estandarizado que permitiera describir la localización y estructura de genes, exones, intrones y otras características genómicas relevantes en secuencias de referencia.

Evolución y adaptaciones:

El formato original de GFF, ahora conocido como GFF1, era simple y cumplía con las necesidades básicas. Sin embargo, a medida que avanzó la genómica y se identificaron más características genómicas y tipos de anotaciones, fue necesario adaptar y expandir el formato.

Este impulso llevó al desarrollo de GFF2, que introdujo cambios y mejoras, pero aún tenía limitaciones en cuanto a la representación de relaciones complejas entre características, como las jerarquías en las estructuras de genes.

Por lo tanto, la versión 3, GFF3, fue lanzada para abordar estas y otras cuestiones. GFF3 introdujo una estructura más rica, permitiendo representar relaciones entre características usando identificadores y referencias. Además, se incorporaron reglas más estrictas para garantizar la uniformidad y consistencia en las anotaciones.

Adopción y usos en la comunidad:

Desde su introducción, el formato GFF ha sido ampliamente adoptado en la genómica. Muchas bases de datos genómicas y navegadores de genomas, como UCSC y Ensembl, han ofrecido la posibilidad de descargar o subir datos en formato GFF. También ha sido esencial para herramientas de anotación de genes y pipelines de análisis genómico.

Con el tiempo, variantes y formatos relacionados, como el GTF (Gene Transfer Format), surgieron para satisfacer las necesidades específicas de ciertos proyectos o plataformas. Aunque estos formatos derivados comparten muchas similitudes con GFF, también poseen diferencias clave en su estructura y especificaciones.

La historia del GFF es un testimonio del dinamismo y adaptabilidad de la comunidad genómica. A medida que la ciencia avanzaba, el formato evolucionó en respuesta a las crecientes demandas y complejidades del campo. Hoy en día, el GFF y sus variantes siguen siendo herramientas indispensables, reflejando su valor y relevancia en un campo que sigue evolucionando rápidamente.

GFF: Funcionamiento y Estructura

El Formato General de Características (General Feature Format, GFF) es un estándar utilizado para describir la localización y estructura de características genómicas en secuencias de referencia. En su versión más reciente, GFF3, este formato ha sido diseñado para ofrecer una representación detallada y extensible de anotaciones genómicas. A continuación, exploramos su estructura y funcionamiento.

Estructura básica:

El GFF es un formato basado en texto con columnas delimitadas por tabulaciones. Cada línea del archivo representa una característica genómica, como un gen, un exón, un intrón, etc. Las columnas en GFF son:

1. seqid: El identificador de la secuencia de referencia (por ejemplo, un cromosoma o contig) donde se encuentra la característica.
2. source: La fuente o programa que generó esta característica. Esto puede ser un programa de predicción de genes, una base de datos de anotaciones, etc.
3. type: El tipo de la característica (por ejemplo, gen, mRNA, exón, CDS, etc.).
4. start: Posición de inicio de la característica en la secuencia de referencia.
5. end: Posición final de la característica.
6. score: Una puntuación asociada con la característica. Si no hay puntuación, se utiliza un punto ('.') como marcador de posición.
7. strand: La hebra en la que se encuentra la característica. Puede ser '+', '-' o '.' (si la hebra es desconocida o no aplicable).
8. phase: Para características del tipo "CDS", indica dónde comienza la siguiente codificación completa. Puede ser '0', '1', '2' o '.' (si no aplica).

9. attributes: Una lista de pares clave-valor separados por punto y coma. Estos atributos pueden incluir identificadores únicos, relaciones con otras características (como un exón asociado con un mRNA específico) y otra metadata.

Ejemplo de una entrada GFF3:

```
chr1 . gene 1000 5000 . + . ID=gene0001;Name=my_gene
chr1 . mRNA 1000 5000 . + . ID=mRNA0001;Parent=gene0001;Name=my_transcript
chr1 . exon 1000 1500 . + . ID=exon0001;Parent=mRNA0001
chr1 . exon 2000 2500 . + . ID=exon0002;Parent=mRNA0001
chr1 . CDS 1200 1500 . + 0 ID=cds0001;Parent=mRNA0001
```

Funcionamiento:

- El GFF permite representar anotaciones genómicas jerárquicas. Por ejemplo, un gen puede tener múltiples tránscritos (mRNAs), y cada tránscrito puede tener varios exones y regiones codificantes (CDS). Estas relaciones se representan usando los atributos "ID" y "Parent" en la columna de atributos.

- Es común que los archivos GFF3 estén acompañados de archivos FASTA que contienen las secuencias de las características descritas. Esto puede estar al final del archivo GFF3, separado por una línea "##FASTA".

El formato GFF3, siendo versátil y detallado, es esencial para representar la complejidad inherente a las anotaciones genómicas. Aunque puede requerir cierta curva de aprendizaje para entender y usarlo eficientemente, su estructura está diseñada para capturar con precisión las relaciones y características genómicas en el contexto de secuencias de referencia. Es crucial para muchos procesos de análisis y herramientas bioinformáticas y es una piedra angular en genómica moderna.

VCF

El formato VCF (Variant Call Format) ha revolucionado la manera en la que representamos y analizamos variantes genéticas en secuenciación de alta profundidad. Desde su creación, ha sido un estándar clave para documentar variantes como SNPs, indels y variantes estructurales más complejas. Veamos cómo surgió y evolucionó el VCF en respuesta a las necesidades emergentes de la genómica.

Con el inicio de la era de la secuenciación de nueva generación (NGS) a mediados de la década de 2000, la comunidad científica comenzó a generar grandes volúmenes de datos de secuencia. A medida que más genomas humanos y de otras especies eran secuenciados, se hizo evidente que existía una diversidad genética masiva entre individuos. Estas variaciones podían ser desde simples cambios de un solo nucleótido (SNPs) hasta deletiones, inserciones o reordenamientos cromosómicos. Era necesario un formato estandarizado para documentar estas variantes de forma coherente y estructurada.

Nacimiento del VCF:

El formato VCF fue introducido por primera vez alrededor de 2008, principalmente por el 1000 Genomes Project, un consorcio internacional que tenía como objetivo secuenciar los genomas de un amplio rango de individuos para catalogar la variación genética humana. VCF fue diseñado para ser flexible, permitiendo la representación de variantes a través de diferentes genomas y ensamblajes de referencia.

Evolución y mejoras:

El VCF original (VCFv1) era funcional pero limitado en su capacidad para representar la creciente complejidad de los datos genómicos. Esto llevó al desarrollo de VCFv2 y, más tarde, VCFv3. Finalmente, VCFv4 (y sus subversiones) incorporaron características adicionales, incluida la capacidad de representar variantes estructurales más complejas y metadata detallada sobre los llamados de variantes.

Una mejora clave en las versiones posteriores de VCF fue la introducción de campos INFO y FORMAT, que proporcionan metadatos adicionales sobre las variantes y permiten una mayor personalización en la descripción de las variantes genéticas.

Adopción y usos en la comunidad:

Desde su introducción, el formato VCF ha sido ampliamente adoptado en la comunidad genómica y se ha convertido en el estándar de facto para la representación de variantes genéticas. Ha sido esencial para grandes proyectos genómicos, bases de datos de variantes y herramientas de bioinformática.

La historia del VCF refleja la rápida evolución y adaptabilidad de la comunidad genómica. Fue desarrollado en respuesta a una necesidad urgente y ha continuado

adaptándose a las demandas cambiantes del campo. En la actualidad, VCF sigue siendo una herramienta fundamental en genómica, evidencia de su robustez y versatilidad. Su desarrollo y adopción masiva es testimonio de la colaboración y visión colectiva de la comunidad científica para abordar desafíos emergentes en genómica.

VCF (Variant Call Format): Funcionamiento y Estructura

El VCF es un formato de texto que se utiliza para describir variantes genéticas en el contexto de secuencias de referencia. Está estructurado para ser claro y legible, pero a la vez proporciona un nivel detallado de información sobre las variantes detectadas. Vamos a desglosar la estructura y el funcionamiento del VCF:

Encabezado (Header):

El archivo VCF comienza con un encabezado, que permite a los usuarios y programas interpretar adecuadamente el contenido del archivo. Las líneas del encabezado comienzan con "##" y proporcionan metadatos sobre el archivo. Estos metadatos pueden incluir versiones del formato VCF, información sobre los programas o pipelines utilizados para generar las variantes, definiciones de campos INFO y FORMAT, y más. También hay una línea de encabezado que especifica las columnas en el archivo, que comienza con "#CHROM".

Columnas principales:

1. CHROM: El nombre del cromosoma o la secuencia de referencia.
2. POS: La posición de inicio de la variante en el cromosoma o secuencia de referencia.
3. ID: Identificador de la variante, por lo general desde bases de datos como dbSNP.
4. REF: El alelo de referencia en esa posición.
5. ALT: El(los) alelo(s) alternativo(s) observado(s) para la variante.
6. QUAL: Un valor que representa la calidad del llamado de variante.
7. FILTER: Indica si la variante ha pasado los filtros establecidos. Si es así, generalmente se muestra "PASS".
8. INFO: Contiene metadatos adicionales sobre la variante, definidos en el encabezado.
9. FORMAT (opcional): Especifica el formato de los datos individuales en las columnas siguientes.
10. Columnas de muestra (si están presentes): Proporcionan información específica de cada muestra sobre la variante, como genotipos, profundidad de lectura, etc.

Ejemplo de una entrada VCF:

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  Sample1 Sample2
chr1    12345   rs1234  A       G       99      PASS    AF=0.5  GT:DP   0/1:20  1/1:18
```

Funcionamiento:

- Cada línea después del encabezado representa una variante genética en una posición específica. Puede ser un SNP, un indel o incluso una variante estructural, dependiendo de las entradas REF y ALT y de la información en la columna INFO.

- La columna INFO proporciona detalles adicionales sobre la variante, como la frecuencia alélica (AF), impactos funcionales, evidencia de asociación a enfermedades, entre otros. Los campos dentro de INFO están definidos en el encabezado y pueden variar dependiendo de la herramienta o base de datos.

- Las columnas de muestra, cuando están presentes, ofrecen información genotípica y otros datos relacionados con las muestras individuales. Por ejemplo, "0/1" podría indicar un genotipo heterocigoto, mientras que "1/1" indica un homocigoto para el alelo alternativo.

Consideraciones finales:

El VCF es una herramienta poderosa que encapsula una amplia gama de información genética en un formato estandarizado. Es fundamental para muchos aspectos de la genómica, desde la investigación básica hasta aplicaciones clínicas. Su diseño modular y extensible lo ha hecho apto para manejar la creciente complejidad y volumen de datos en la genómica moderna. Al comprender su estructura y funcionamiento, los investigadores y clínicos pueden extraer, comparar y analizar de manera efectiva la rica información contenida en sus entradas.

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in

Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6

Agradecimientos

El presente estudio ha sido financiado por el proyecto AIR Genomics (con número de expediente CCTT3/20/SA/0003), mediante la convocatoria 2020 PROYECTOS I+D ORIENTADOS A LA EXCELENCIA Y MEJORA COMPETITIVA DE LOS CCTT por el Instituto de Competitividad Empresarial de Castilla y León y fondos FEDER