(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 729 714**

(51) Int. CI.:

G16B 45/00 (2009.01)
G16B 30/00 (2009.01)
G16B 50/00 (2009.01)
G06T 3/40 (2006.01)
G09G 5/373 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **07.12.2012 PCT/US2012/068493**

(87) Fecha y número de publicación internacional: **13.06.2013 WO13086355**

(96) Fecha de presentación y número de la solicitud europea: **07.12.2012 E 12856007 (5)**

(97) Fecha y número de publicación de la concesión europea: **15.05.2019 EP 2788861**

(54) Título: **Sistema distribuido que proporciona indexado dinámico y visualización de datos genómicos**

(30) Prioridad:

08.12.2011 US 201161568478 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.11.2019

(73) Titular/es:

FIVE3 GENOMICS, LLC (100.0%)
101 Cooper Street
Santa Cruz, California 95060, US

(72) Inventor/es:

VASKE, CHARLES JOSEPH;
SANBORN, JOHN ZACHARY y
BENZ, STEPHEN

(74) Agente/Representante:

ARIAS SANZ, Juan

DESCRIPCIÓN

Sistema distribuido que proporciona indexado dinámico y visualización de datos genómicos

La presente solicitud reivindica el beneficio de la solicitud provisional de EE.UU. con el número de serie 61/568478, que se presentó el 8 de diciembre de 2011.

Campo de la invención

El campo de la invención es la genómica computacional, especialmente se refiere a representación gráfica dinámica de información genética compleja.

Antecedentes

La siguiente descripción incluye información que puede ser útil en el entendimiento de la presente invención. No es una admisión de que cualquiera de la información proporcionada en el presente documento sea estado de la técnica o relevante para la invención presentemente reivindicada, o que cualquier publicación específica o implícitamente citada sea estado de la técnica.

Con la aparición de secuenciación de alto rendimiento y la disponibilidad de conjuntos de datos del genoma completo, la velocidad de secuenciación ya no es el cuello de botella en el análisis del genoma, pero sí el almacenamiento, la recuperación y el análisis coordinado de datos. Las dificultades asociadas al almacenamiento, recuperación y análisis de datos se agravan además por los requisitos variables de información presentada de diferentes usuarios. Visto desde una perspectiva diferente, es fundamental la presentación densa en información y selectiva de datos genómicos para hacer uso de la enorme cantidad de datos ahora disponibles.

Aunque existen varios navegadores genómicos conocidos en la técnica, todos los navegadores tienen numerosas dificultades. Por ejemplo, el navegador UCSC Genome (http://genome.ucsc.edu) proporciona cuantiosos datos en una forma gráfica, sin embargo, fracasa en tener en cuenta la densidad de información especificada por un usuario como presentaciones predefinidas que son independientes del nivel de zoom. Por tanto, dichos navegadores son incapaces de responder óptimamente a peticiones a todos los niveles de zoom. Similarmente, los visualizadores gráficos como el de NCBI (http://www.ncbi.nlm.nih.gov/nuccore/) también están limitados a ciertos parámetros predefinidos y así fracasan en permitir la presentación dinámica y adaptación de contenido.

Por consiguiente, aún cuando se conocen en la técnica diversos sistemas y métodos de presentación de información genómica compleja, sin embargo, siguen existiendo numerosas desventajas. Por tanto, todavía existe la necesidad de proporcionar dispositivos y métodos mejorados para la representación gráfica de información genética compleja, y especialmente representación gráfica dinámica.

Se puede interpretar que el documento US 2003/204317 A1 desvela métodos, productos de software informático y sistemas informáticos para presentar información genética. Se usa Semantic Zooming para facilitar la visualización de información genética.

Se puede interpretar que el documento US 2009/125248 A1 desvela un sistema para el análisis y la visualización de datos genómicos. El sistema permite a un usuario seleccionar al menos una muestra individual. La muestra tiene datos cromosómicos que representan un genoma con un cromosoma y también incluye mediciones cromosómicas de al menos un evento en una localización particular sobre el cromosoma. Se genera una frecuencia de evento basándose en la muestra seleccionada. La frecuencia de evento es una frecuencia de aparición del evento en la muestra seleccionada. Se puede seleccionar al menos una anotación que incluye información específica de la región cromosómica como relacionada con el cromosoma. Finalmente, los datos cromosómicos, la anotación y la frecuencia de evento en un dispositivo de presentación se pueden presentar simultáneamente, permitiendo así que un usuario visualice información específica de la región cromosómica con respecto a un evento cromosómico particular.

Se puede interpretar que el documento US 2010/281401 A1 desvela un navegador de genoma interactivo que se ejecuta dentro de una aplicación de navegador web, configurado para presentar datos genéticos del paciente y trayectorias de datos genéticos adicionales que se alinean por par de bases. Los seguimientos adicionales pueden incluir datos, datos comunitarios, datos privados, huecos de secuencia y pruebas genéticas adicionales o sondas que están disponibles. Las pruebas o sondas se pueden ordenar seleccionándolas de una trayectoria de prueba o sonda. Los datos en una base de datos de información genética también pueden ser buscados usando el navegador de genoma interactivo. Los datos del paciente analizado se pueden publicar y poner a disposición de una comunidad de usuarios, que se pueden comunicar entre sí.

Sumario de la invención

La materia inventiva se refiere a métodos y dispositivos para la visualización dinámica de datos genómicos en la que

un sistema de visualización genómica adapta la presentación de contenido de información según anotaciones relevantes para la escala dentro de un objeto de secuencia. Así, se puede lograr la presentación de contenidos adaptativos a análisis y transferencia de datos significativamente reducidas.

Según la divulgación, se proporciona un aparato según la reivindicación independiente. Los desarrollos se exponen en las reivindicaciones dependientes.

Preferentemente, se contempla un sistema de visualización genómica que comprende una base de datos genómicos indexados que almacena un objeto de secuencia representativo de una región genómica. Preferentemente, el objeto de secuencia incluye una pluralidad de anotaciones relevantes para la escala. Se acopla un motor de escalado preferentemente al almacenamiento de datos genómicos indexados y se configura para (a) ajustar la información relevante para la escala derivada de las anotaciones relevantes para la escala del objeto de secuencia en función del nivel de zoom seleccionado por un usuario, (b) generar dinámicamente un objeto de presentación genómica representativo de la información relevante para la escala basado en el nivel de zoom, y (c) configurar un dispositivo de salida para presentar los objetos de presentación genómica a un usuario.

Preferentemente, el objeto de secuencia tiene un formato SAM/BAM o BAMBAM, y/o esa región genómica es un genoma completo, un cromosoma, un fragmento cromosómico o un alelo.

Con respecto al motor de escalado, preferentemente uno o más servidores BAM y/o servidores de visualización pueden operar como el motor de escalado. Además, se contempla que el motor de escalado se puede configurar además para ajustar la información relevante para la escala por submuestreo basado en el nivel de zoom (en el que el submuestreo puede ser una función de densidad de datos derivada del nivel de zoom). Alternativamente, o además, preferentemente el motor de escalado se configura para determinar el nivel de zoom, y opcionalmente para resumir un conjunto completo de datos del objeto de secuencia según el nivel de zoom. Preferentemente, el motor de escalado se configura para derivar la información relevante para la escala de diferencias en las anotaciones relevantes para la escala en diferentes objetos de secuencia.

Preferentemente, el objeto de secuencia comprende un objeto de secuencia de referencia, que es lo más preferentemente datos de secuencia sin procesar, datos de secuencia de homo statisticus y/o datos de secuencia de un momento especificado en el tiempo. Alternativamente, o además, el objeto de secuencia comprende preferentemente un objeto de secuencia diferencial con respecto a una región genómica de referencia (por ejemplo, región genómica de referencia de homo statisticus o para un momento especificado en el tiempo). Similarmente, las anotaciones relevantes para la escala preferentemente varían considerablemente y preferentemente incluirán información de la estructura genómica (por ejemplo, identificación de cromosomas, localización dentro de un cromosoma, alelo, etc.), información de cambio genómico (por ejemplo, una mutación, una translocación, una inversión, una deleción, una repetición y un número de copias), información de enfermedad (por ejemplo, tipo de enfermedad, un estado de enfermedad y una opción de tratamiento para la enfermedad), información relevante del gen (por ejemplo, datos de secuencia sin procesar o datos de secuencia procesados, identificación de genes, información sobre la regulación génica e información de asociación del gen con una enfermedad), información diferencial con respecto a una secuencia de referencia y/o metadatos (por ejemplo, identificación del paciente, identificación del centro, identificación del médico e información del seguro).

Preferentemente, el sistema de visualización genómica incluirá además una biblioteca gráfica genómica que almacena un objeto gráfico representativo de anotaciones relevantes para la escala. En dichos sistemas, preferentemente, el motor de escalado mapea la información relevante para la escala en los objetos gráficos de la biblioteca gráfica según el nivel de zoom, y ese objeto de presentación genómica comprende los objetos gráficos mapeados. Con respecto a dispositivos de salida adecuados, se prefieren un dispositivo de presentación, un navegador, una impresora, una impresora 3D y/o un altavoz.

Diversos objetos, características, aspectos y ventajas de la materia inventiva serán más evidentes a partir de la siguiente descripción detallada de realizaciones preferidas, junto con las figuras de dibujo adjuntas en las que números similares representan componentes similares.

Breve descripción de los dibujos

La Figura 1 proporciona una visión general de un entorno de visualización genómica distribuida.
La Figura 2 ilustra un posible sistema de visualización genómica que incluye un motor de escalado de la visualización.
La Figura 3 es una vista de presentación a modo de ejemplo al nivel de zoom base.
La Figura 4 es la vista de presentación a modo de ejemplo de la Figura 3 a un nivel de zoom de sub-kilobases.
La Figura 5 es la vista de presentación a modo de ejemplo de la Figura 4 a un nivel de zoom de kilobases.
La Figura 6 es la vista de presentación a modo de ejemplo de la Figura 5 a un nivel de zoom de cromosoma.

Descripción detallada

La materia inventiva se refiere a dispositivos y métodos de visualización dinámica de datos genómicos. Los sistemas y métodos contemplados permiten la presentación selectiva y escalable de contenido rico en información, mientras se reducen el tráfico y la agregación de datos.

Se debe observar que aunque la siguiente descripción se refiere a sistemas de visualización genómica basados en un ordenador/servidor, también se consideran adecuadas diversas configuraciones alternativas y se pueden emplear diversos dispositivos informáticos que incluyen servidores, interfaces, sistemas, bases de datos, agentes, nodos, motores, controladores, u otros tipos de dispositivos informáticos que operan individualmente o conjuntamente. Se debe apreciar que los dispositivos informáticos comprenden un procesador configurado para ejecutar las instrucciones de software almacenadas en un medio de almacenamiento legible por ordenador no transitorio tangible (por ejemplo, disco duro, unidad de estado sólido, RAM, memoria rápida, ROM, etc.). Las instrucciones de software configuran preferentemente el dispositivo informático para proporcionar los roles, responsabilidades, u otra funcionalidad como se trata más adelante con respecto al aparato desvelado. En realizaciones especialmente preferidas, los diversos servidores, sistemas, bases de datos o interfaces intercambian datos usando protocolos o algoritmos normalizados, posiblemente basados en HTTP, HTTPS, AES, intercambios de clave pública-privada, APIs de servicio web, protocolos conocidos de transacciones financieras, u otros métodos de intercambio de información electrónica. Los intercambios de datos se realizan preferentemente mediante una red de conmutación de paquetes, internet, LAN, WAN, VPN, u otro tipo de red de conmutación de paquetes.

En toda la siguiente discusión, se harán numerosas referencias referentes a servidores, servicios, interfaces, portales, plataformas, u otros sistemas formados de dispositivos informáticos. Se debe apreciar que se considera que el uso de dichos términos representa uno o más dispositivos informáticos que tienen al menos un procesador configurado para ejecutar instrucciones de software almacenadas en un medio no transitorio tangible legible por ordenador. Por ejemplo, un servidor puede incluir uno o más ordenadores que operan como un servidor web, servidor de base de datos, u otro tipo de servidor informático de un modo que cumpla los roles, responsabilidades o funciones descritas.

Como se usa en la descripción en el presente documento y en todas las reivindicaciones que siguen, el significado de "un", "una", "el" y "la" incluyen referencia al plural, a menos que el contexto dicte claramente de otro modo. Por tanto, como se usa en la descripción en el presente documento, el significado de "en" incluye "en" y "sobre", a menos que el contexto dicte claramente de otro modo.

La citación de intervalos de valores en el presente documento tiene simplemente la finalidad de servir de método abreviado de referencia individual a cada valor separado que se encuentra dentro del intervalo. A menos que se indique lo contrario en el presente documento, cada valor individual se incorpora en la memoria descriptiva como si fuera individualmente citado en el presente documento. Todos los métodos descritos en el presente documento se pueden realizar en cualquier orden adecuado, a menos que se indique lo contrario en el presente documento o se contradiga claramente de otro modo por el contexto. El uso de todos y cada uno de los ejemplos, o vocabulario a modo de ejemplo (por ejemplo, "tal como"), proporcionados con respecto a ciertas realizaciones en el presente documento tiene simplemente la finalidad de iluminar mejor la invención y no plantea una limitación al alcance de la invención de otro modo reivindicada. Ningún vocabulario en la memoria descriptiva se debe interpretar como que indica cualquier elemento no reivindicado esencial para la práctica de la invención.

No se deben interpretar como limitaciones agrupaciones de elementos alternativos o realizaciones de la invención desveladas en el presente documento. Cada miembro de grupo se puede denominar y reivindicar individualmente o en cualquier combinación con otros miembros del grupo u otros elementos encontrados en el presente documento. Uno o más miembros de un grupo se pueden incluir en, o eliminar de, un grupo por motivos de comodidad y/o patentabilidad. Cuando ocurra cualquiera de dicha inclusión o deleción, se considera que la memoria descriptiva en el presente documento contiene el grupo como se modifica, satisfaciendo así la descripción escrita de todos los grupos de Markush usados en las reivindicaciones adjuntas. Aunque cada realización representa una combinación única de elementos inventivos, se considera que la materia inventiva incluye todas las posibles combinaciones de los elementos desvelados. Así, si una realización comprende los elementos A, B y C, y una segunda realización comprende los elementos B y D, entonces también se considera que la materia inventiva incluye otras combinaciones restantes de A, B, C, o D, aunque no se desvelen explícitamente.

Como se usa en el presente documento, y a menos que el contexto dicte de otro modo, el término "acoplado a" pretende incluir tanto acoplamiento directo (en el que dos elementos que se acoplan entre sí se ponen en contacto entre sí) como acoplamiento indirecto (en el que al menos un elemento adicional se localiza entre los dos elementos). Por tanto, los términos "acoplado a" y "acoplado con" se usan sinónimamente.

Los dispositivos y métodos contemplados combinan características ventajosas de un servidor BAM y un motor de visualización del genoma que se acoplan holgadamente, tal como para permitir la integración trivial con otros motores impulsados por genómica alternativos u otras soluciones de almacenamiento de datos genómicos. Además, cada componente puede ser modificado de escala según sea necesario para acomodar múltiples servidores BAM o

múltiples motores de visualización, como se ilustra esquemáticamente y a modo de ejemplo en la **Figura 1**. Lo más preferentemente, cada servidor es lo suficientemente flexible como para mantener almacenamiento independiente, autenticación y recuperación de datos por sí mismo, así como en una naturaleza distribuida donde cada servidor puede coordinar algunas partes con otros servidores. Además, la capacidad de tanto el servidor BAM como el motor de visualización para modificar dinámicamente de escala los datos proporcionados de grandes fuentes de datos ayudarán a mitigar los significativos aumentos en los tamaños de datos de futuros formatos de datos y tipos de archivo.

La **Figura 2** ilustra el sistema de visualización genómica 200 capaz de generar una presentación visual de información genómica a diferentes escalas de observación. El sistema 200 incluye base de datos genómicos indexados 220 y motor de escalado 230. En algunas realizaciones, el sistema 200 también puede incluir biblioteca gráfica genómica 237 o incluso dispositivos 250, que posiblemente operan como clientes de los servicios ofrecidos por el sistema 200. Por ejemplo, los dispositivos 250 pueden incluir un dispositivo informático habilitado para navegador (por ejemplo, un teléfono móvil, tableta, ordenador, etc.), mediante el que un profesional sanitario o un paciente pueden acceder a la información genómica de interés mediante la red 215. El motor de escalado 230 puede proporcionar una presentación visual de la información genómica al servidor del usuario mediante HTTP, u otro protocolo adecuado.

Generalmente, se contempla que un sistema de visualización genómica 200 comprenderá una base de datos genómicos indexados 220 que almacena uno o más de objetos de secuencia 223 representativos de una región genómica, en el que el objeto de secuencia 223 incluye una pluralidad de anotaciones relevantes para la escala 225. El motor de escalado 230 se acopla con la base de datos genómicos indexados 220 y se configura para ajustar información relevante para la escala 233 que deriva de las anotaciones relevantes para la escala 225 del objeto de secuencia 223 en función del nivel de zoom seleccionado por un usuario 252. El motor de escalado 230 generará entonces dinámicamente un objeto de presentación genómica 235 que es representativo de la información relevante para la escala 233 basándose en el nivel de zoom 252, y configura un dispositivo de salida 250 para presentar el objeto de presentación genómicas 235 a un usuario.

Como se usa en el presente documento, el término "región genómica" normalmente se refiere a un nombre de secuencia y una coordenada inicial y final que especifican un intervalo cerrado dentro de esa secuencia. Una región genómica de ejemplo es: chr1:1234-5678, donde chrl especifica la secuencia del cromosoma 1 de un genoma humano de referencia, 1234 es la coordenada inicial y 5678 es la coordenada final. Sin embargo, debe ser fácilmente evidente para el experto habitual en la técnica que el formato particular de la región genómica puede variar considerablemente, y que formatos adecuados incluirán referencias particulares a la localización y/o sub-localización cromosómicas, a nombres de genes o funciones, aspectos reguladores del (de los) gen(es) en la región, aspectos estructurales de la cromatina del (de los) gen(es) en la región, longitud de secuencia, etc. Por tanto, y visto desde una perspectiva diferente, la región genómica puede ser un genoma completo, un cromosoma, un fragmento cromosómico o un alelo. Además, se debe observar que es posible la especificación de múltiples regiones genómicas en una única aplicación usando cualquier delimitador conocido entre las regiones genómicas.

Por consiguiente, se deberá reconocer que el objeto de secuencia 223 puede tener numerosos formatos de datos, y que todos los formatos conocidos se consideran adecuados, mientras que dichos formatos también incluyan una o más anotaciones relevantes para la escala. Por ejemplo, formatos particularmente preferidos para los objetos de secuencia contemplados incluyen el formato SAM/BAM y BAMBAM. Asimismo, se debe apreciar que el objeto de secuencia 223 puede representar una región genómica de un genoma de referencia (por ejemplo, de *homo statisticus*) o una región genómica de una muestra de prueba. Si el objeto de secuencia 223 es de una muestra de prueba que se va a analizar, se prefiere normalmente que el análisis se realice con respecto a un genoma de referencia y/o un genoma del mismo sujeto de prueba de un momento de tiempo diferente. Así, objetos de secuencia de referencia 223 adecuados pueden incluir datos de secuencia sin procesar, datos de secuencia de *homo statisticus* y/o datos de secuencia de un sujeto de prueba de un momento de tiempo especificado. Además, se deberá reconocer que el objeto de secuencia 223 no necesita estar necesariamente confinado a la lectura de datos sin procesar o secuencia ensamblada (por ejemplo, gen de longitud completa), sino que el objeto de secuencia 223 puede ser o comprender un objeto de secuencia diferencial 223 con respecto a una región genómica de referencia (por ejemplo, en la que solo se enumeran bases correspondientes discordantes). Como antes, dicha región genómica de referencia puede ser del mismo probando de ensayo tomado en un momento de tiempo anterior, o de un probando sano real o uno hipotético, secuencia consenso de múltiples probandos sanos (*homo statisticus*).

Con respecto a las anotaciones relevantes para la escala 225, se contempla que las anotaciones 225 pueden variar considerablemente y que todas las anotaciones conocidas en el análisis genómico se consideran adecuadas para su uso en el presente documento. Por ejemplo, las anotaciones 225 particularmente preferidas incluyen las relacionadas con la estructura genómica a diversos niveles de escala (por ejemplo, localización de secuencia sobre un cromosoma, localización dentro de un cromosoma, información de alelos, etc.) y las relacionadas con los cambios genómicos a diversos niveles de escala (por ejemplo, translocación cromosómica, repetición o número de copias, inserciones, deleciones, inversiones, diversas mutaciones tales como SNPs, transiciones, transversiones, etc.). Asimismo, las anotaciones relevantes para la escala 225 también pueden incluir información de enfermedad a diversos niveles de escala (por ejemplo, poliploidía, números de copias y/o repeticiones, opciones de

tipo/estado/tratamiento de una enfermedad asociada a mutaciones o números de copias, etc.). En aspectos adicionalmente contemplados, las anotaciones relevantes para la escala 225 también pueden incluir información relevante del gen a diversos niveles de escala (por ejemplo, gen como parte de una red funcional o reguladora de genes, nombre del gen o identificación funcional, datos de secuencia sin procesar o datos de secuencia procesados, identificación de genes, información sobre la regulación génica, e información de asociación del gen con una enfermedad).

Por supuesto, se debe apreciar que toda o parte de la información relevante también se puede expresar como información diferencial con respecto a una secuencia de referencia (por ejemplo, homo statisticus o momento anterior en el tiempo), que reducirá ventajosamente el tamaño de datos y la complejidad. Además, las anotaciones relevantes para la escala 225 normalmente también incluirán metadatos asociados al objeto de secuencia, y lo más normalmente incluyen identificación del paciente, identificación del centro, identificación del médico y/o información del seguro.

Visto desde una perspectiva diferente, las anotaciones relevantes para la escala 225 incluirán anotaciones que son adecuadas para la presentación de audiencias seleccionadas (por ejemplo, médico, investigador, paciente, seguro, etc.). Por ejemplo, si la audiencia es un médico, anotaciones relevantes para la escala 225 pueden ser relevantes para un formato de presentación de un genoma completo en formato simplificado (por ejemplo, gráfico de círculo, expansión de metafases, etc.) donde las mutaciones se indican por indicadores simples u otras herramientas gráficas. Por otra parte, si la audiencia es un investigador, anotaciones relevantes para la escala 225 pueden ser relevantes para un formato de presentación en el que se proporcionan datos de secuencia sin procesar reales y número de copias/frecuencia de alelos.

Además, e independientemente de la audiencia, se deberá reconocer que el tipo de presentación visual cambiará dinámicamente en función del nivel de zoom 252, tal que se presente contenido apropiado con respecto al zoom. Por consiguiente, las anotaciones relevantes para la escala 225 pueden incluir además datos que indican idoneidad para la anotación particular para un nivel o niveles de zoom específicos 252. Por supuesto, la idoneidad para la presentación a un nivel de zoom dado también se puede determinar independientemente de dichos datos como se trata adicionalmente más adelante. Se puede determinar el nivel de zoom 252 seleccionado por un usuario mediante diversas técnicas. En algunas realizaciones, el nivel de zoom 252 se puede determinar basándose en el perfil de usuario: profesional sanitario, paciente, compañía aseguradora, investigador, u otro tipo de perfil. Por ejemplo, se puede seleccionar el nivel de zoom 252 que representa un zoom de nivel más alto (es decir, la máxima vista de la región genómica) como por defecto cuando un paciente está viendo los datos. Alternativamente, un investigador podría tener un nivel de zoom 252 por defecto que se dirige a regiones específicas de interés. Otras técnicas para establecer el nivel de zoom 252 incluyen recibir un cuadro delimitador seleccionado por el usuario del dispositivo de visualización (por ejemplo, navegador, aplicación, etc.), tener en cuenta automáticamente regiones genómicas anómalas con respecto a una región de referencia (homo statisticus), recibir información genómica de un dispositivo de secuencia indicativo de una región de interés, u otras técnicas.

Existen numerosas opciones para representar gráficamente las anotaciones relevantes para la escala 225 y se prefiere especialmente que la representación gráfica se realice usando símbolos y notaciones conocidos. Lo más preferentemente, los símbolos y notaciones conocidos se pueden almacenar en una biblioteca gráfica genómica 237 que está configurada para almacenar objetos gráficos representativos de las anotaciones relevantes para la escala 225. En dicho caso, se prefiere particularmente que el motor de escalado se configure para mapear la información relevante para la escala 233 en objetos gráficos de la biblioteca gráfica 237 según el nivel de zoom 252, y que el objeto de presentación genómica 235 comprenda los objetos gráficos mapeados. Por ejemplo, el motor de escalado 230 recibe el nivel de zoom 252 de un profesional sanitario que está revisando la información genómica de un paciente con respecto a mutaciones conocidas. El motor de escalado 230 obtiene el objeto de secuencia 223 de la base de datos genómicos indexados 220 junto con las anotaciones relevantes para la escala 225 asociadas. El motor de escalado 230 deriva la información relevante para la escala 233 en función de las anotaciones relevantes para la escala 225, la información del profesional sanitario (por ejemplo, autorización, perfil, etc.) y el nivel de zoom 252. La información relevante para la escala 233 representa así la región genómica de objeto de secuencia 223 a un nivel de zoom apropiado, así como a un nivel apropiado de detalle con respecto al observador. En otras palabras, al nivel de zoom dado, la información relevante para la escala 233 representa la información que sería apropiada para el profesional sanitario. Si el observador fuera un paciente, la información relevante para la escala 233 llevaría probablemente una presentación diferente de la información genómica de la que sería apropiada para el paciente aún cuando el nivel de zoom 252 y el objeto de secuencia 223 fueran idénticos. El motor de escalado 230 mapea entonces la información relevante para la escala 233 en uno o más objetos gráficos en la biblioteca gráfica genómica 237 para crear objeto de presentación genómica 235.

Se debe apreciar que la biblioteca gráfica genómica 237 se configura para guardar objetos gráficos genómicos en vez de meros primitivos gráficos. La biblioteca gráfica genómica 237 puede ser actualizada con objetos gráficos genómicos adicionales según se desee, o se pueden modificar objetos gráficos genómicos existentes, posiblemente con diferentes gráficas (por ejemplo, texturas, pieles, temas, etc.). Se considera que dicho enfoque es ventajoso dentro del mercado, ya que permite el desarrollo de marca o personalización de presentaciones visuales.

Con respecto al hardware se debe observar que los dispositivos y métodos contemplados se pueden configurar y operar en numerosos modos, y se debe apreciar que la configuración particular y/o el modo de operación vendrán impuestos al menos en parte por los componentes funcionales e interconexiones. Así, la siguiente descripción de aspectos preferidos solo se debe visualizar como orientación a modo de ejemplo para el experto habitual en la técnica.

Con respecto a servidores BAM adecuados, generalmente se prefiere que el servidor BAM sea o comprenda un sistema distribuido de servidores de red capaz del eficiente acceso aleatorio a datos indexados por región genómica, que soporta el acceso protegido a datos encriptados tanto mediante conexiones seguras como mediante acceso de archivo encriptado. En un caso de uso típico, un usuario: 1. Se conectará con el servidor BAM mediante la red, 2. Emitirá una solicitud con dos parámetros - A) un archivo de datos y B) una lista de regiones genómicas, y 3. Recibirá todas las entradas de datos del archivo que solapen cualquiera de las regiones genómicas proporcionadas. Como se usa en el presente documento, el término "archivo de datos" se refiere a un conjunto de entradas de datos donde cada entrada se asocia con una región genómica. Una entrada de datos puede ser cualquier dato, que incluye un único número, una serie de caracteres, y una lista de números y/o series. Algunos ejemplos comunes de entradas de datos son una lectura de secuencia y la calidad de lectura asociada de una máquina de secuenciación, una localización de gen conocido, o una mutación detectada.

Indexado de región genómicas: Cuando se añade un archivo de datos al servidor BAM, el servidor BAM clasifica las entradas de datos por región genómica, luego crea preferentemente un árbol R como índice de discretización, como se usa comúnmente en aplicaciones de genómica y se ha descrito completamente en su uso en el navegador UCSC Genome y la biblioteca de software SAM Tools. Brevemente, una secuencia indexada se separa en cajas que se solapan. Empezando con una caja que cubre la secuencia entera, se añaden dos cajas nuevas que dividen la caja previa en la mitad. Entonces, el índice tiene indicadores desde cada caja hasta las entradas de datos que se ajustan dentro de esa caja, pero no caja más pequeña. La recuperación de entradas de datos que solapan con una consulta es entonces cuestión de examinar solo las cajas que solapan con la consulta.

Protecciones de acceso a datos: Lo más normalmente, el servidor BAM restringe el acceso a archivos de datos no públicos comprobando cada solicitud con un servidor de acceso a archivos de datos. Si el cliente no proporciona credenciales de seguridad suficientes según el servidor de acceso a archivos de datos, se deniega el acceso a cualquier resultado. Cada servidor BAM se puede configurar para un servidor de acceso a archivos de datos único, que permite esquemas de permiso flexibles y métodos de autentificación federados.

Con respecto al almacenamiento de datos, generalmente se contempla que los archivos de datos del servidor BAM se almacenen en un sistema de archivos que aparece local al servidor BAM. Este sistema de archivos puede usar discos unidos directamente al servidor BAM y/o discos accesibles mediante red. Se prefiere además que los archivos de datos protegidos se almacenen en una forma encriptada (por ejemplo, cifrado simétrico por bloques AES, usando el modo CTR). El servidor BAM normalmente no tendrá acceso a la clave de encriptación. Cuando se procesa una solicitud para un archivo de datos protegido, si el servidor de acceso a archivos de datos concede acceso, el servidor de acceso a archivos de datos proporcionará la clave de encriptación para el archivo solicitado. El servidor BAM usará la clave mientras se procesa la solicitud, y descartará la clave tan pronto como se haya procesado completamente la solicitud.

Los métodos de solicitud adecuados normalmente se hacen usando consultas RESTful (conforme a restricciones de transferencia de estado representacional) mediante HTTPS, un protocolo HTTP asegurado por SSL, o usando un mecanismo de tunelización encriptado alternativo dentro del cual se hacen las consultas de HTTPS. La naturaleza de RESTful de las consultas permite que los servidores BAM se distribuyan tanto geográficamente como locamente para proporcionar el máximo rendimiento a aplicación consumidoras. La única restricción a la localidad del servidor BAM es el acceso de archivos directos a los datos subyacentes, que se podrían incluso presentar con respecto a una red de área amplia usando los protocolos apropiados (NFS con respecto a VPN, u otras de dichas soluciones).

En aspectos preferidos adicionales, se implementa el escalado dinámico de los datos. Basándose en el tamaño de la región genómica solicitada y el conocimiento de la resolución con la que se presentarán los datos, el servidor BAM, posiblemente operando como motor de escalado 230, tiene capacidades para escalar dinámicamente ("submuestreo") los datos para proporcionar una versión más condensada que reducirá los tiempos de procesamiento y transferencia. Este submuestreo se realiza lo más preferentemente en dos mecanismos paralelos. El primer mecanismo no requiere conocimiento de los datos subyacentes, y se lleva a cabo proporcionando los archivos del servidor BAM que están pre-condensados hasta ciertos niveles. El servidor BAM puede entonces decidir dinámicamente en el momento de la consulta si debe proporcionan un nivel de datos "sin procesar", o alternativamente uno de los archivos condensados. Esta decisión se hace incluyendo un parámetro adicional en la solicitud que indica el número de puntos de datos que se utilizarán por la aplicación consumidora. Si la aplicación consumidora es un motor de visualización, que también podría operar como motor de escalado 230, un ejemplo de un recuento útil de puntos de datos se podría basar en el número de píxeles que serán dibujados en la pantalla. El segundo mecanismo para submuestreo es la sumarización dinámica de los datos completos accesibles al servidor BAM. Este mecanismo requiere proporcionar información adicional sobre el tipo de archivo para el servidor BAM de manera que pueda entender qué campos son posibles sumarizar, y el mecanismo de sumarización. Dado un archivo

con solo una única columna de datos, además del índice de coordenadas genómicas, esto se podría determinar automáticamente y se podría realizar automáticamente una mediana de la sumarización o media. Para tipos de datos más complejos o técnicas de sumarización más complejas, el servidor BAM requerirá parámetros que delineen cómo realizar esa sumarización. Un ejemplo es el submuestreo de un archivo en el formato SAM/BAM, que realizaría un submuestreo por sub-muestreo de las lecturas individuales en cada posición, proporcionando solo un número limitado de nuevo a la aplicación consumidora.

Se debe apreciar además que los sistemas y métodos contemplados son fácilmente extensibles, ya que el servidor BAM es capaz de leer archivos de múltiples formatos y entender tanto datos genómicamente indexados como formatos de almacenamiento adicionales tales como SQLite y JSON. El formato del archivo solicitado se proporciona actualmente por la aplicación consumidora, pero también se contempla la auto-detección del formato del archivo. La arquitectura del servidor BAM soporta preferentemente formatos de datos adicionales en forma de programas adicionales que pueden entender esquemas de indexado extranjeros y todavía proporcionar una interfaz unificada. Estos programas adicionales son o bien especificados mediante la solicitud REST del identificador universal de recursos (URI), o por autodetección del formato apropiado dentro del servidor BAM.

Con respecto a los motores de visualización del genoma dinámicos, generalmente se contempló que un motor de visualización del genoma dinámico era capaz de interpretar múltiples tipos de datos con el atributo común de ser mapeados en una localización en el genoma, y producir interpretaciones basadas en la imagen de los datos. Se debe observar que el concepto de un "navegador" del genoma en cierto sentido es ya conocido (por ejemplo, Universidad de California, el buscador Santa Cruz Genome, establecido en 2001 (véase URL genome.ucsc.edu)). Sin embargo, los navegadores actualmente conocidos limitan las visualizaciones de datos a densidades especificadas por el usuario y son incapaces de responder a solicitudes pasados ciertos límites de una manera oportuna y significativa. A diferencia, el motor de visualización del genoma dinámico contemplado en el presente documento es capaz de entender la cantidad de datos que se solicita por un usuario y alterar las visualizaciones presentadas para proporcionar versiones más compactas y sumarizadas cuando convenga. En el nivel uno, el nivel de submuestreo es manipulado por el servidor BAM, que entiende la región que se está intentando visualizar, y reducirá automáticamente los datos enviados al motor de visualización. A un nivel más alto, si el propio motor reconoce que está siendo solicitada una cantidad suficientemente grande de datos, las visualizaciones subyacentes producidas se alterarán de tal forma que proporcionen sumarios que son más útiles para el usuario final.

Las presentaciones pueden variar ampliamente basándose en la densidad de datos que se intentan visualizar. Las Figuras 3-6 representan algunos ejemplos de cómo estas presentaciones cambian basándose en los diversos números de bases que el usuario está visualizando en la ventana donde las presentaciones se generan a partir de objetos gráficos genómicos usados para generar los objetos de presentación genómica 235 dentro de un navegador. Es importante enfatizar que estas presentaciones se generan dinámicamente y no se calculan previamente, aunque para ciertos casos de uso no se excluyen imágenes estáticas pregeneradas y son soportadas por los dispositivos y métodos contemplados. En la **Figura 3**, se muestran 52 bases del genoma humano a través de aproximadamente 1000 píxeles horizontales, con representaciones gráficas del número total de copias, número de copias específicas de alelo, datos de secuenciación sin procesar de BAM y una trayectoria de anotación de UCSC Known Genes. Cada uno de estas trayectorias es sacada dinámicamente de la arquitectura del servidor BAM brevemente expuesto anteriormente, y cada trayectoria puede consultar un servidor BAM independiente para obtener los datos necesarios. Debido a que se está mostrando dicho pequeño número de bases, no se está realizando submuestreo ni en el servidor BAM ni el motor de visualización. Así, se prefiere particularmente que el nivel de zoom más bajo sea en la lectura base de la secuencia sin procesar o calculada.

La **Figura 4** representa un nivel de zoom de sub-kilobases que muestra aproximadamente 1000 bases de la misma región del genoma. A esta resolución y número de bases no está teniendo lugar submuestreo en el servidor BAM, sin embargo, el motor de visualización ha empezado a alterar la presentación de cada fuente de datos para acomodar el puerto de visualización aumentado. En particular, las letras en cada base ya no aparecen tanto en la barra de bases de referencia superior como dentro de las lecturas de bam individuales, en lugar de recurrir a simples colores para representar los cambios identificados.

La **Figura 5** está visualizando aproximadamente 2 megabases (2 millones de bases) a un nivel de zoom de kilobases mientras que el número de píxeles se mantiene constante. Como resultado, tanto el servidor BAM como el motor de visualización han submuestreado los datos que se dibujan. El servidor BAM ha reducido la cantidad de datos de números de copias que proporciona el motor de visualización, y el motor de visualización ha ignorado la trayectoria de datos sin procesar debido a que la visualización no sería práctica. Además, el motor de visualización ha empezado a sumarizar uno de las trayectorias de variantes (la trayectoria más inferior) que produce un histograma gráfico en la parte superior. Finalmente, el motor de visualización ha promediado juntos los múltiples puntos de datos para la variación del número de copias que están debajo de cada píxel para producir una imagen más precisa.

La resolución final, Figura 6, representa todo el cromosoma 12 a un nivel de zoom de cromosoma. Todo el submuestreo previo está ocurriendo a esta resolución, siendo el submuestreo adicional reducido para retirar el texto y presentar una representación más gráfica de tanto UCSC Known Gene como las trayectorias de variantes

COSMIC en la parte inferior de la imagen. Aunque se ha representado un claro ejemplo en estos diagramas, este motor proporciona un marco para la visualización dinámica que no se limita a niveles de resolución predeterminados y previamente dibujados, y además puede acomodar muchos tipos diferentes de datos subyacentes, además de los que se han mostrado aquí.

Debe ser evidente para los expertos en la técnica que son posibles muchas más modificaciones, además de las ya descritas, sin apartarse del alcance de las reivindicaciones adjuntas. Todos los términos se deben interpretar en el modo más amplio posible, de acuerdo con el contexto. En particular, los términos "comprende" y "que comprende" se deben interpretar como con referencia a elementos, componentes, o etapas de una manera no excluyente, que indica que los elementos, componentes o etapas referenciados pueden estar presentes, o ser utilizados, o combinados con otros elementos, componentes o etapas que no se referencian expresamente. Si una definición o uso de un término en una referencia incorporada es incoherente o contrario a la definición de ese término proporcionado en el presente documento, se aplica la definición de ese término proporcionada en el presente documento y no se aplica la definición de ese término en la referencia. Si las reivindicaciones de la memoria descriptiva se refieren a al menos uno de algo seleccionado del grupo que consiste en A, B, C .... y N, lo siguiente se debe interpretar como que requiere solo un elemento del grupo, no A más N, o B más N, etc.

REIVINDICACIONES

1. Un sistema de visualización genómica que comprende:

una base de datos genómicos indexados (220) configurada para almacenar un objeto de secuencia (223) representativo de una región genómica, comprendiendo el objeto de secuencia una pluralidad de anotaciones relevantes para la escala; y
un motor de escalado (230) acoplado al almacenamiento de datos genómicos indexados y configurado para:

- ajustar la información relevante para la escala (233) derivada de las anotaciones relevantes para la escala del objeto de secuencia en función del nivel de zoom seleccionado por un usuario (252);
- generar dinámicamente un objeto de presentación genómica (235) representativo de la información relevante para la escala basado en el nivel de zoom, en el que la generación dinámica comprende una alteración de la visualización presentada del objeto de secuencia;
- submuestrear dinámica y automáticamente la cantidad de datos que comprende el objeto de presentación genómica en respuesta a una solicitud que indica varios puntos de datos que se utilizarán por una aplicación consumidora, y en el que el submuestreo dinámico se basa además en el tamaño de la región genómica solicitada y el conocimiento de la resolución gráfica de un dispositivo de salida (250) con el que se presentarán los datos; y
- configurar el dispositivo de salida para presentar los objetos de presentación genómica a un usuario.

2. El sistema de la reivindicación 1, en el que

(i) el objeto de secuencia tiene un formato SAM/BAM o BAMBAM; o
(ii) la región genómica es una de las siguientes: un genoma completo, un cromosoma, un fragmento cromosómico y un alelo.

3. El sistema de la reivindicación 1, que comprende además un servidor BAM que opera como el motor de escalado.

4. El sistema de la reivindicación 1, que comprende además un servidor de visualización que opera como el motor de escalado.

5. El sistema de la reivindicación 1, en el que el dispositivo de salida comprende al menos uno de los siguientes: un dispositivo de presentación, un navegador, una impresora, una impresora 3D y un altavoz.

6. El sistema de la reivindicación 1, en el que el motor de escalado se configura además para ajustar la información relevante para la escala por submuestreo basado en el nivel de zoom.

7. El sistema de la reivindicación 1, en el que el motor de escalado se configura además para determinar el nivel de zoom, y en el que el motor de escalado se configura además para resumir un conjunto completo de datos del objeto de secuencia según el nivel de zoom.

8. El sistema de la reivindicación 1, en el que el motor de escalado se configura además para derivar la información relevante para la escala de diferencias en anotaciones relevantes para la escala en diferentes objetos de secuencia.

9. El sistema de la reivindicación 1, en el que el objeto de secuencia comprende un objeto de secuencia de referencia.

10. El sistema de la reivindicación 1, en el que el objeto de secuencia comprende un objeto de secuencia diferencial con respecto a una región genómica de referencia.

11. El sistema de la reivindicación 1, en el que las anotaciones relevantes para la escala incluyen al menos una de las siguientes: información de estructura genómica, información de cambio genómico, información de enfermedad, información relevante del gen, información diferencial con respecto a una secuencia de referencia y metadatos.

12. El sistema de la reivindicación 11, en el que

(i) la información de estructura genómica incluye al menos una de las siguientes: identificación de cromosomas, localización dentro de un cromosoma o localización dentro de un alelo;
(ii) la información de cambio genómico incluye al menos una de las siguientes: una mutación, una translocación, una inversión, una deleción, una repetición y un número de copias;
(iii) la información de enfermedad incluye al menos una de las siguientes: un tipo de enfermedad, un estado de enfermedad y una opción de tratamiento para la enfermedad;
(iv) la información relevante del gen comprende datos de secuencia sin procesar o datos de secuencia procesados, identificación de genes, información sobre la regulación génica, e información de asociación del gen con una enfermedad; o

(v) los metadatos incluyen al menos uno de los siguientes: identificación del paciente, identificación del centro, identificación del médico e información del seguro.

13. El sistema de la reivindicación 1, que comprende además una biblioteca gráfica genómica (237) configurada para almacenar un objeto gráfico representativo de anotaciones relevantes para la escala.

14. El sistema de la reivindicación 13, en el que el motor de escalado se configura además para mapear la información relevante de escala a objetos gráficos de la biblioteca gráfica según el nivel de zoom.

15. El sistema de la reivindicación 14, en el que el objeto de presentación genómica comprende los objetos gráficos mapeados.

FIG. 1

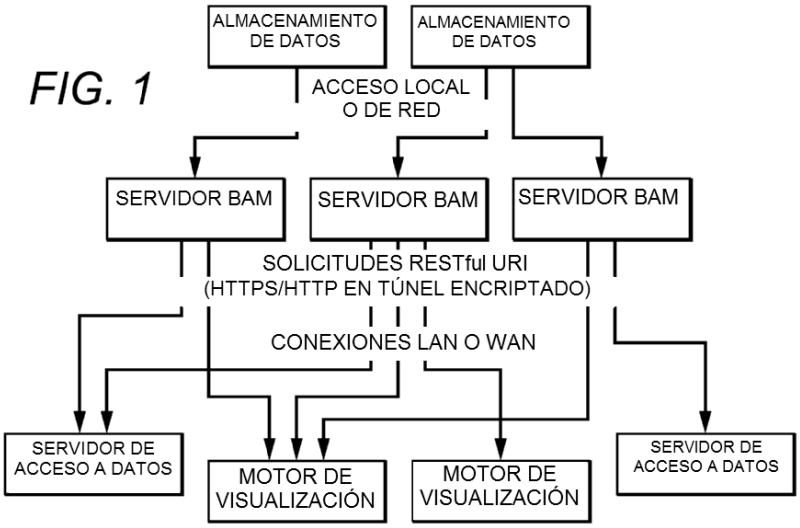

ALMACENAMIENTO DE DATOS
ALMACENAMIENTO DE DATOS
ACCESO LOCAL O DE RED
SERVIDOR BAM
SERVIDOR BAM
SERVIDOR BAM
SOLICITUDES RESTful URI (HTTPS/HTTP EN TÚNEL ENCRIPTADO)
CONEXIONES LAN O WAN
SERVIDOR DE ACCESO A DATOS
MOTOR DE VISUALIZACIÓN
MOTOR DE VISUALIZACIÓN
SERVIDOR DE ACCESO A DATOS

FIG. 2

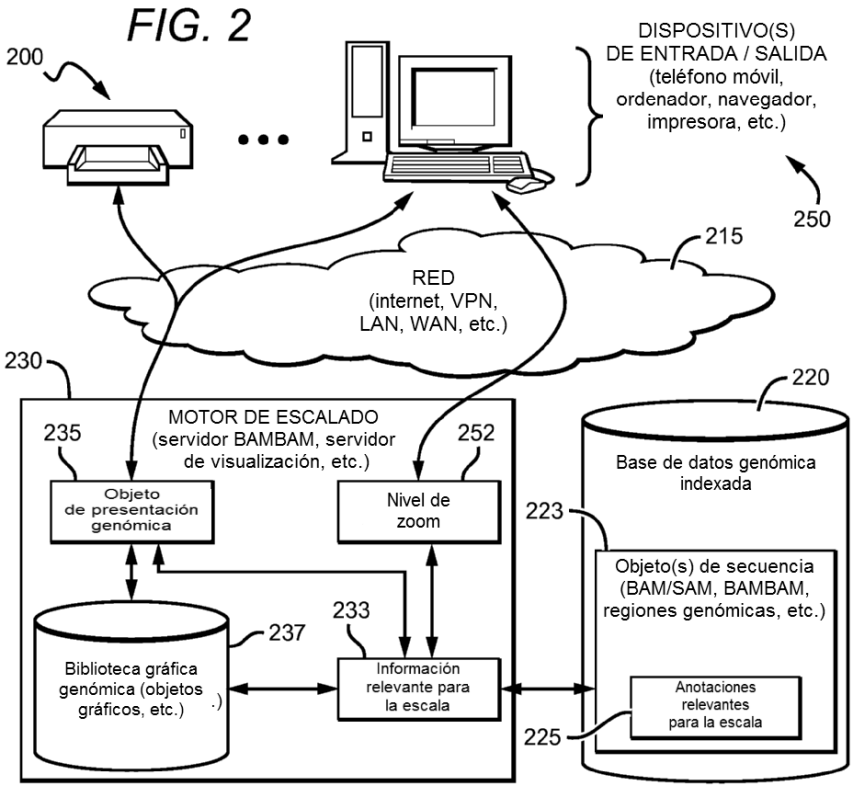
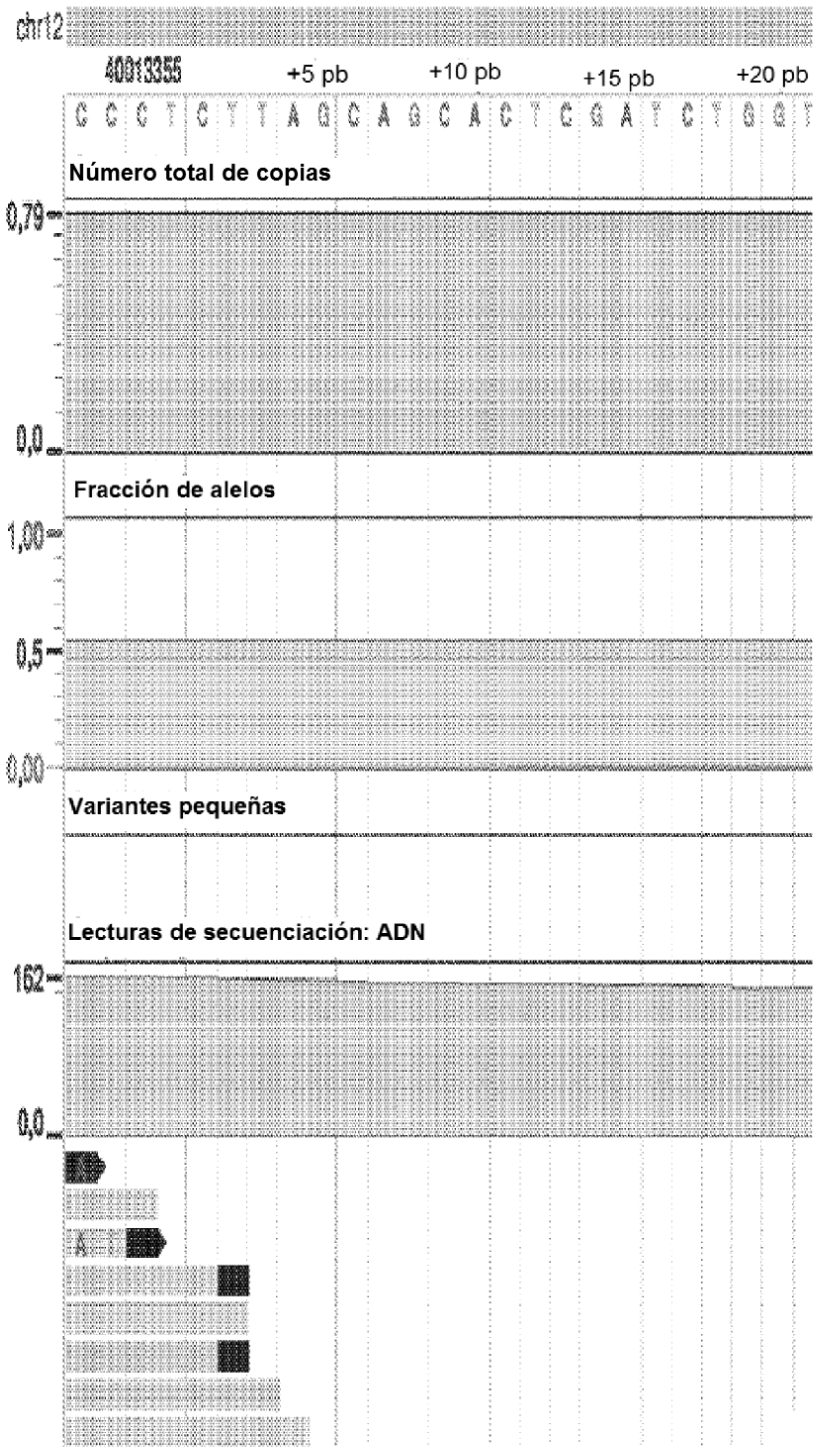

200
DISPOSITIVO(S) DE ENTRADA / SALIDA (teléfono móvil, ordenador, navegador, impresora, etc.)
250
215
RED (internet, VPN, LAN, WAN, etc.)
230
220
235
MOTOR DE ESCALADO (servidor BAMBAM, servidor de visualización, etc.)
252
Objeto de presentación genómica
Nivel de zoom
Base de datos genómica indexada
223
Objeto(s) de secuencia (BAM/SAM, BAMBAM, regiones genómicas, etc.)
233
237
Biblioteca gráfica genómica (objetos gráficos, etc.)
Información relevante para la escala
Anotaciones relevantes para la escala
225

FIG. 3-1


chr12
40913355
+5 pb
+10 pb
+15 pb
+20 pb
Número total de copias
0,79
0,0
Fracción de alelos
1,00
0,5
0,00
Variantes pequeñas
Lecturas de secuenciación: ADN
162
0,0

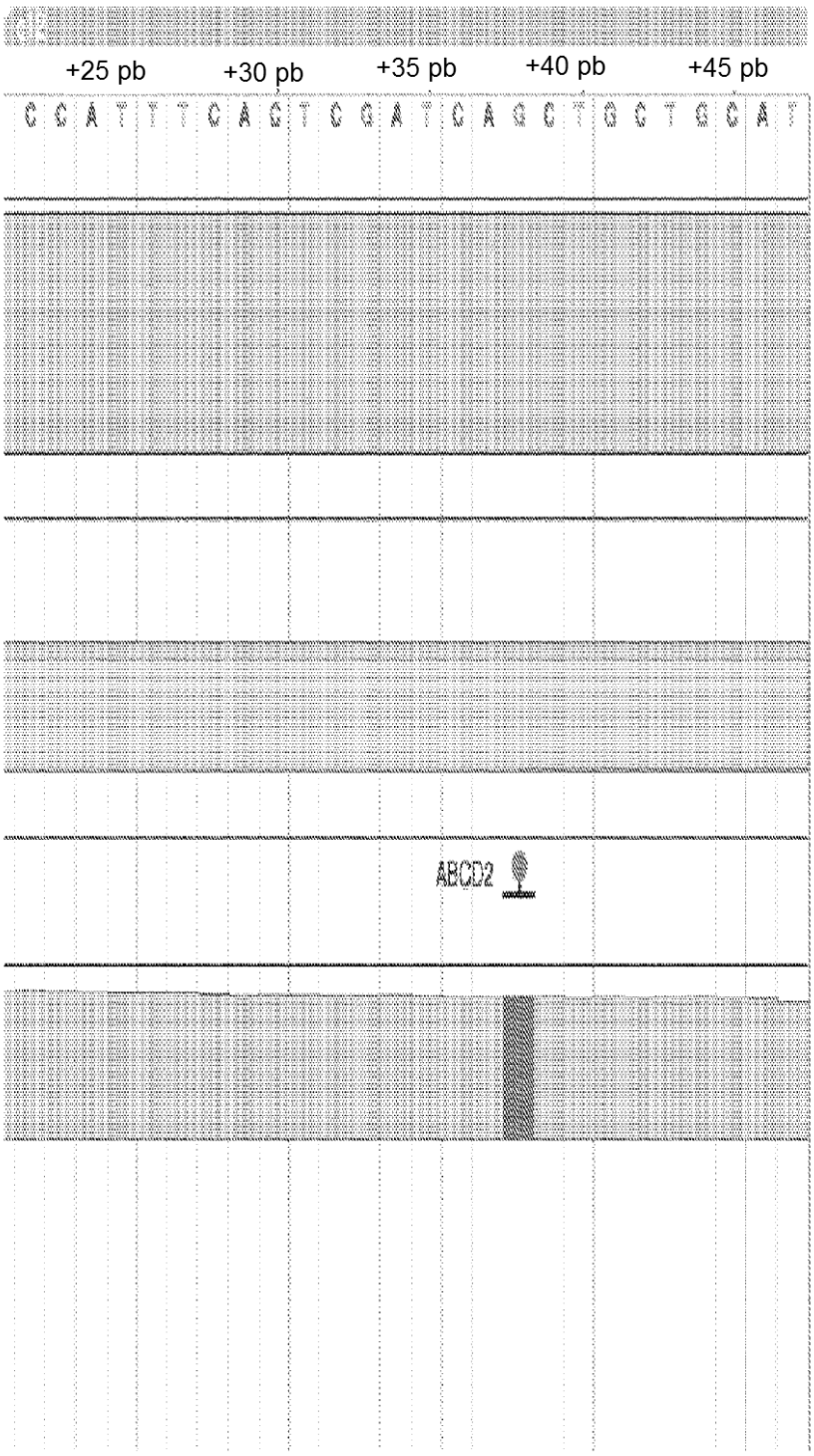
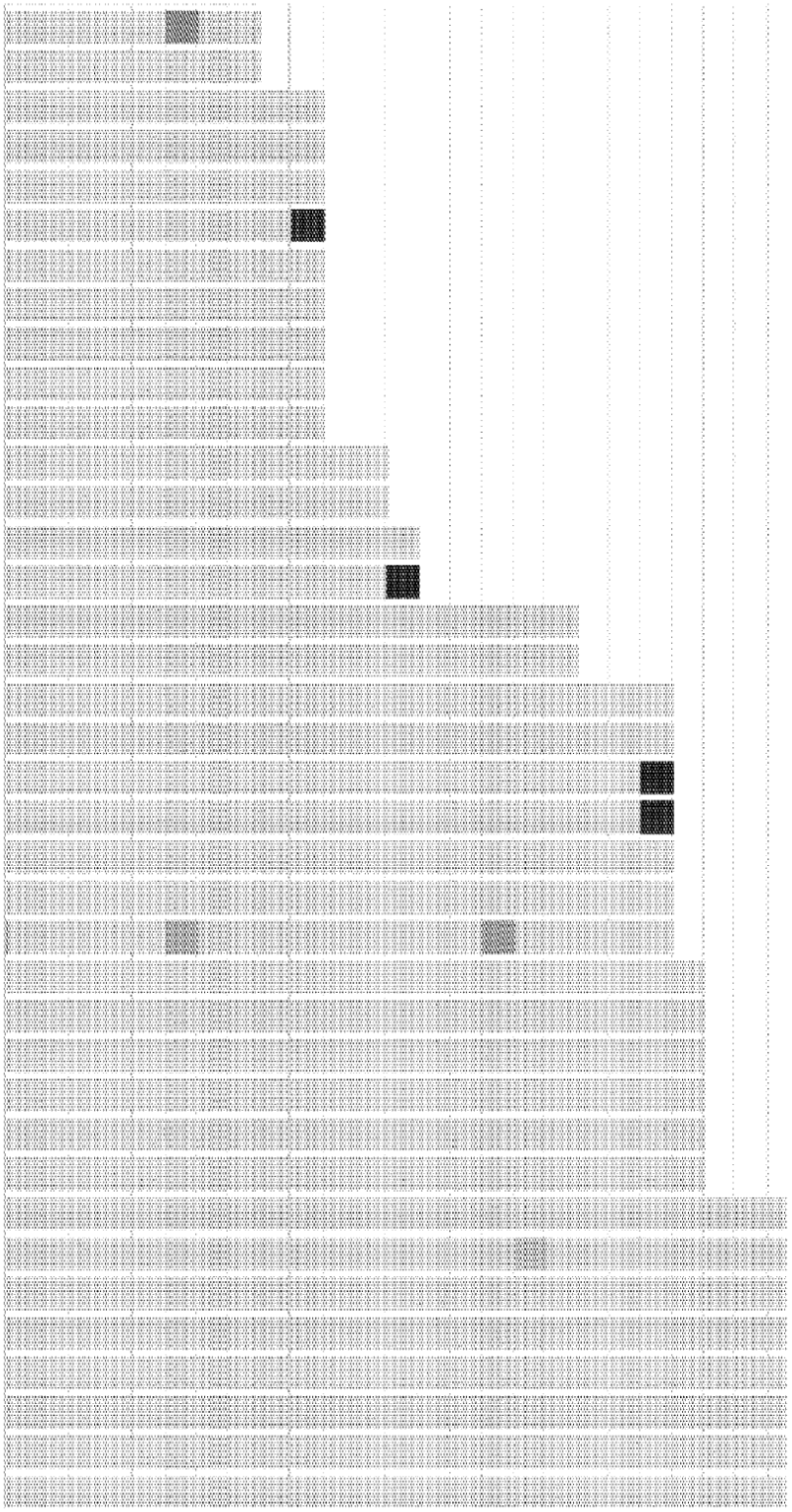

+25 pb
+30 pb
+35 pb
+40 pb
+45 pb
C C A T T T C A C T C G A T C A G C T G C T G C A T
ABCD2

FIG. 3-2

FIG. 3-3



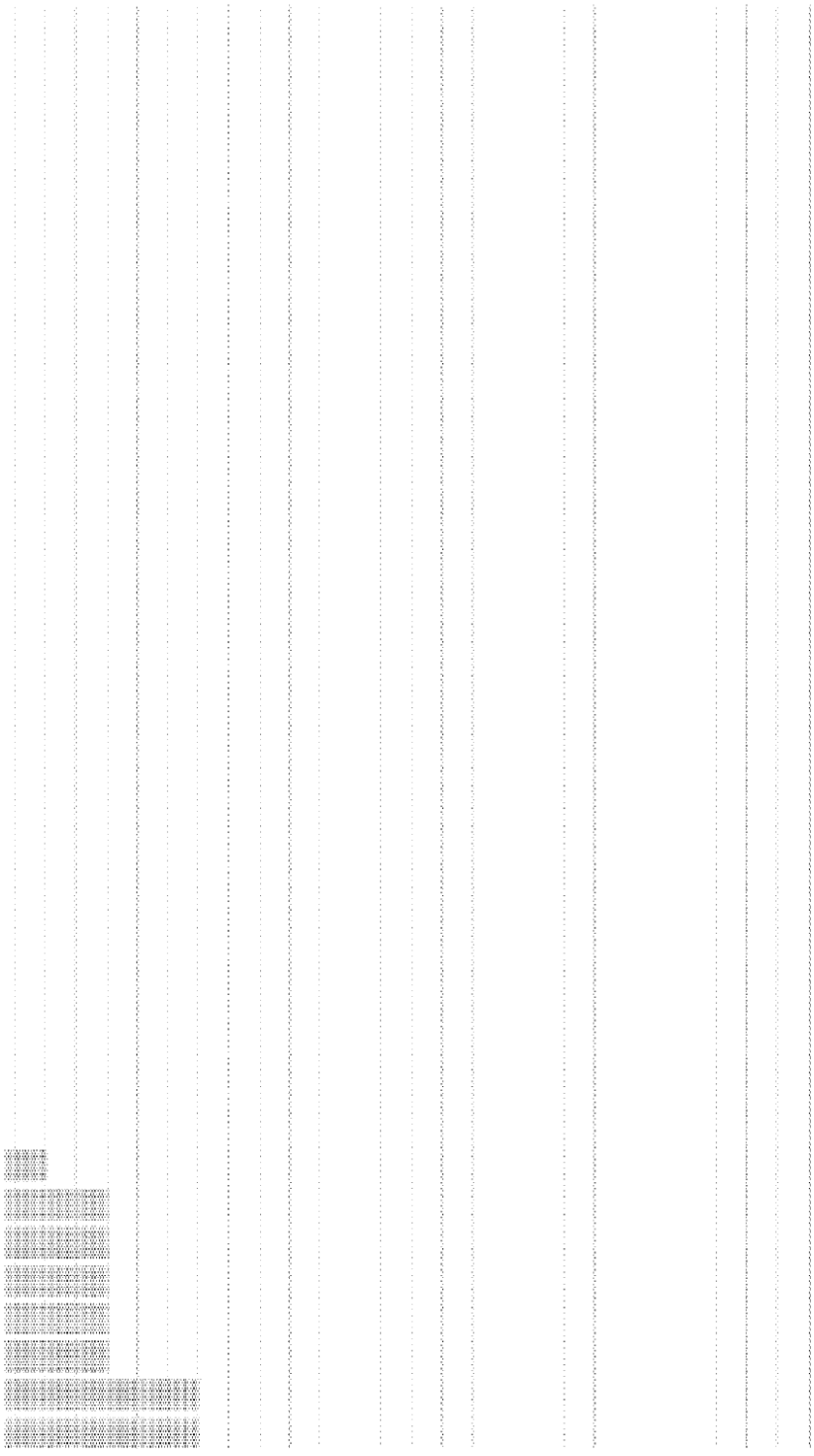
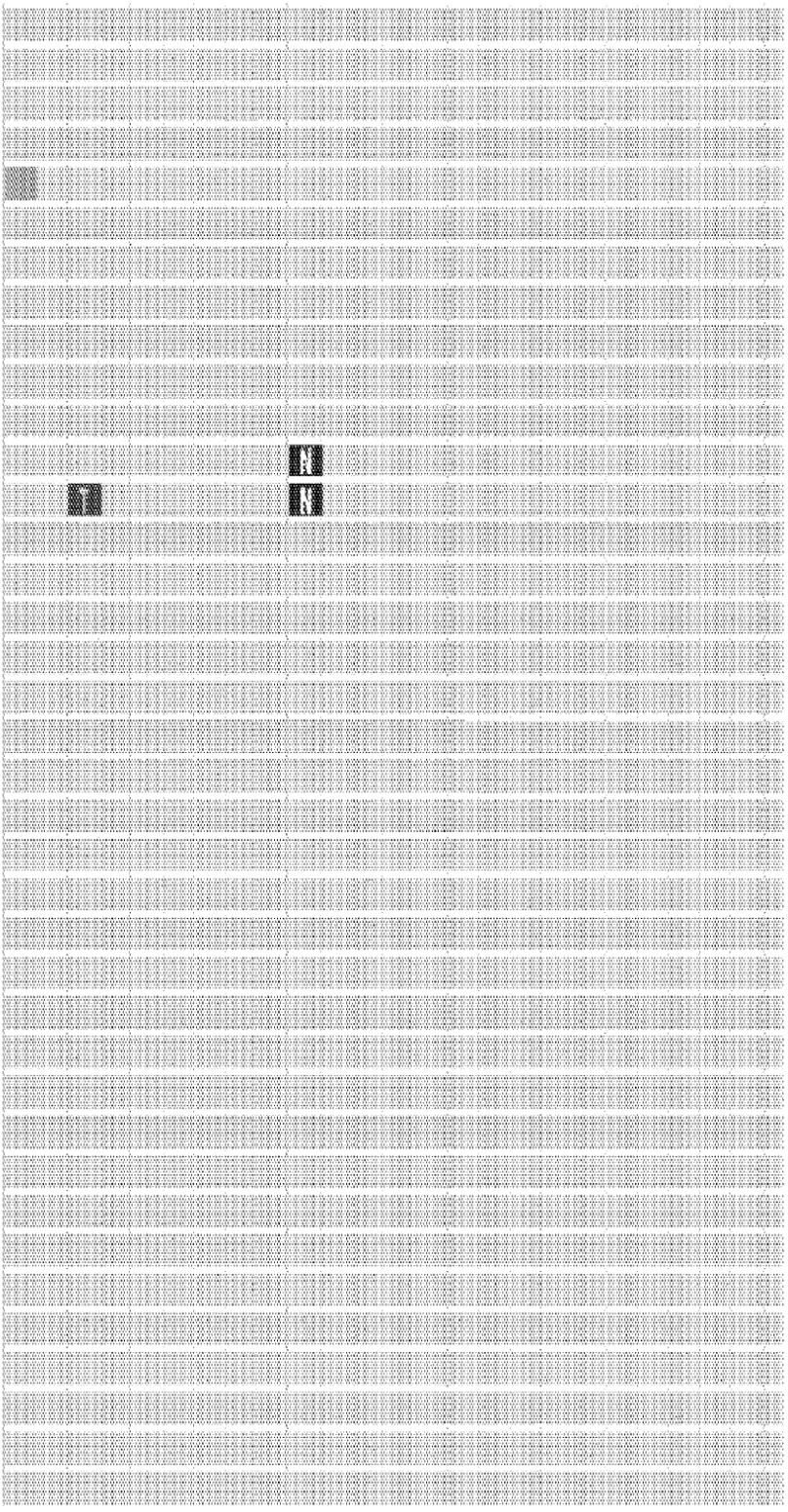


FIG. 3-4

FIG. 3-5



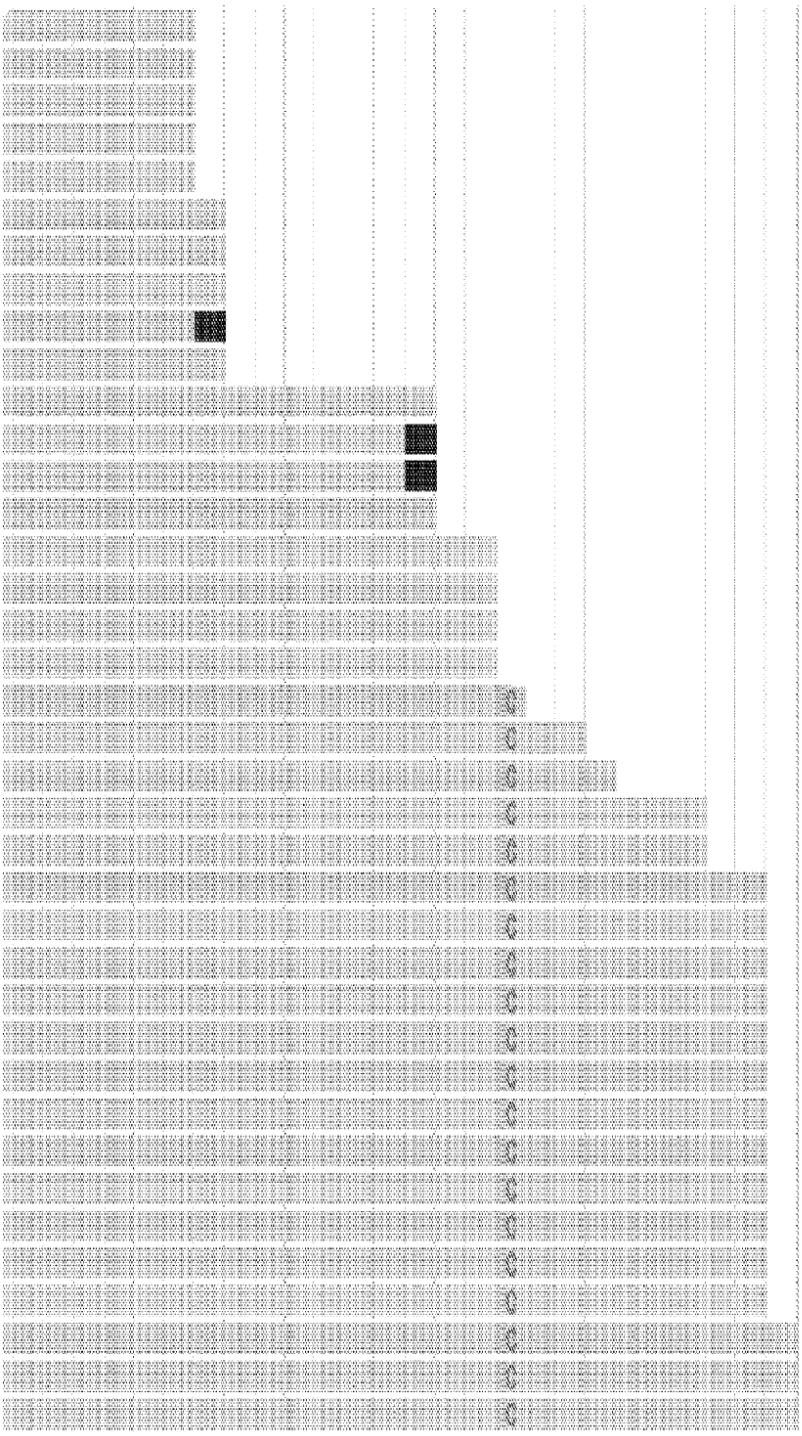
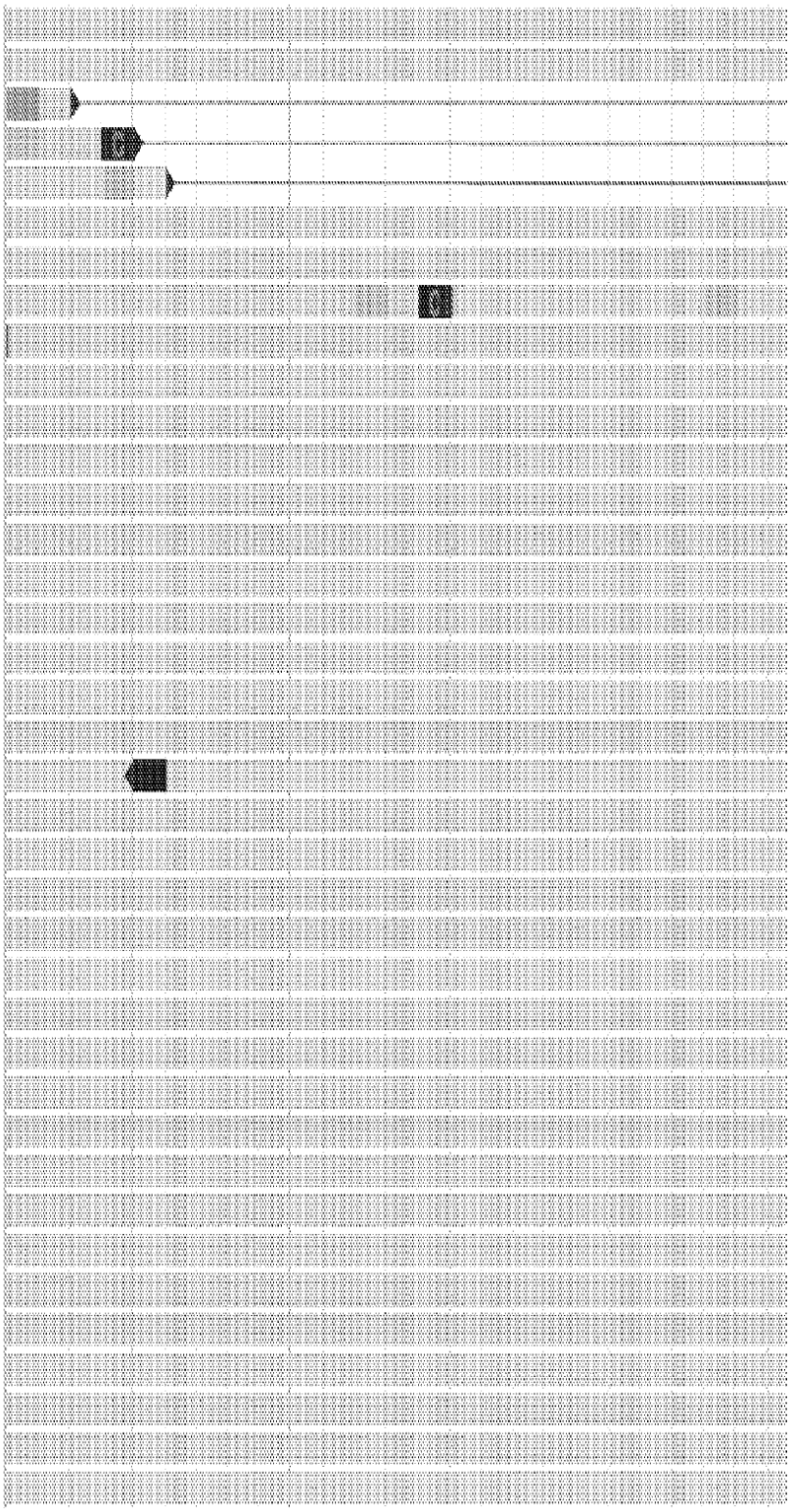


FIG. 3-6

FIG. 3-7



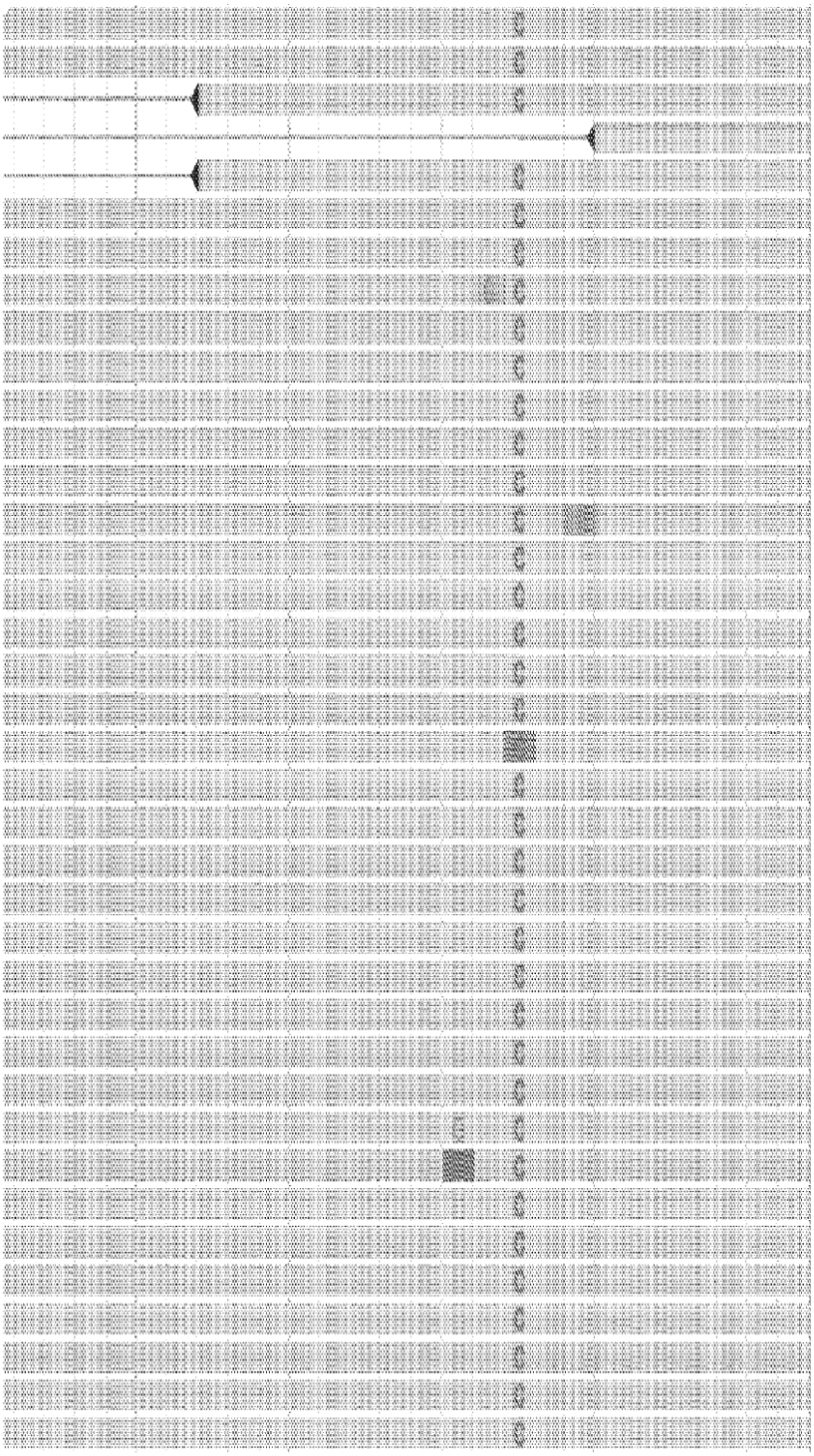
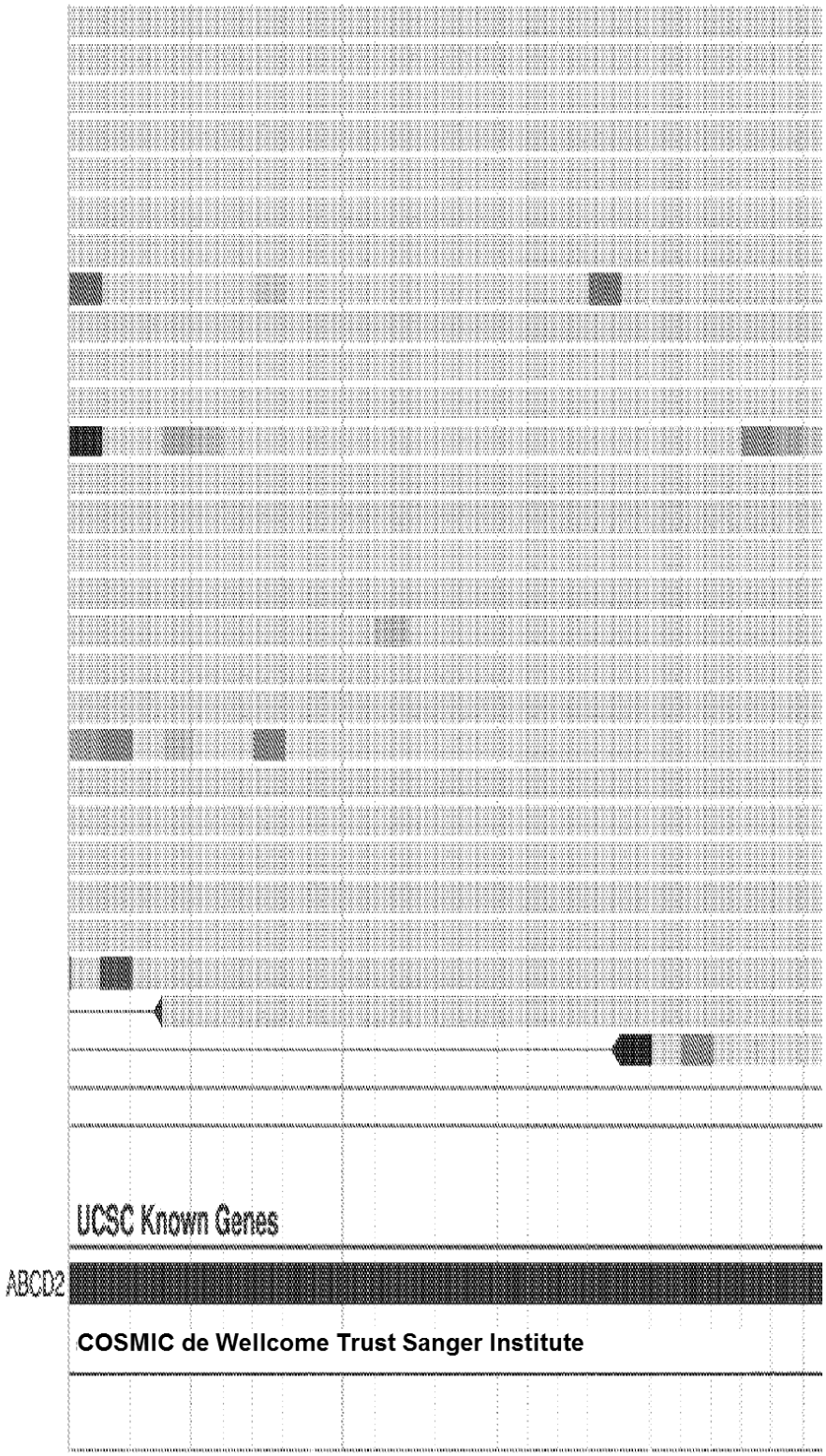


FIG. 3-8

FIG. 3-9


UCSC Known Genes
ABCD2
COSMIC de Wellcome Trust Sanger Institute

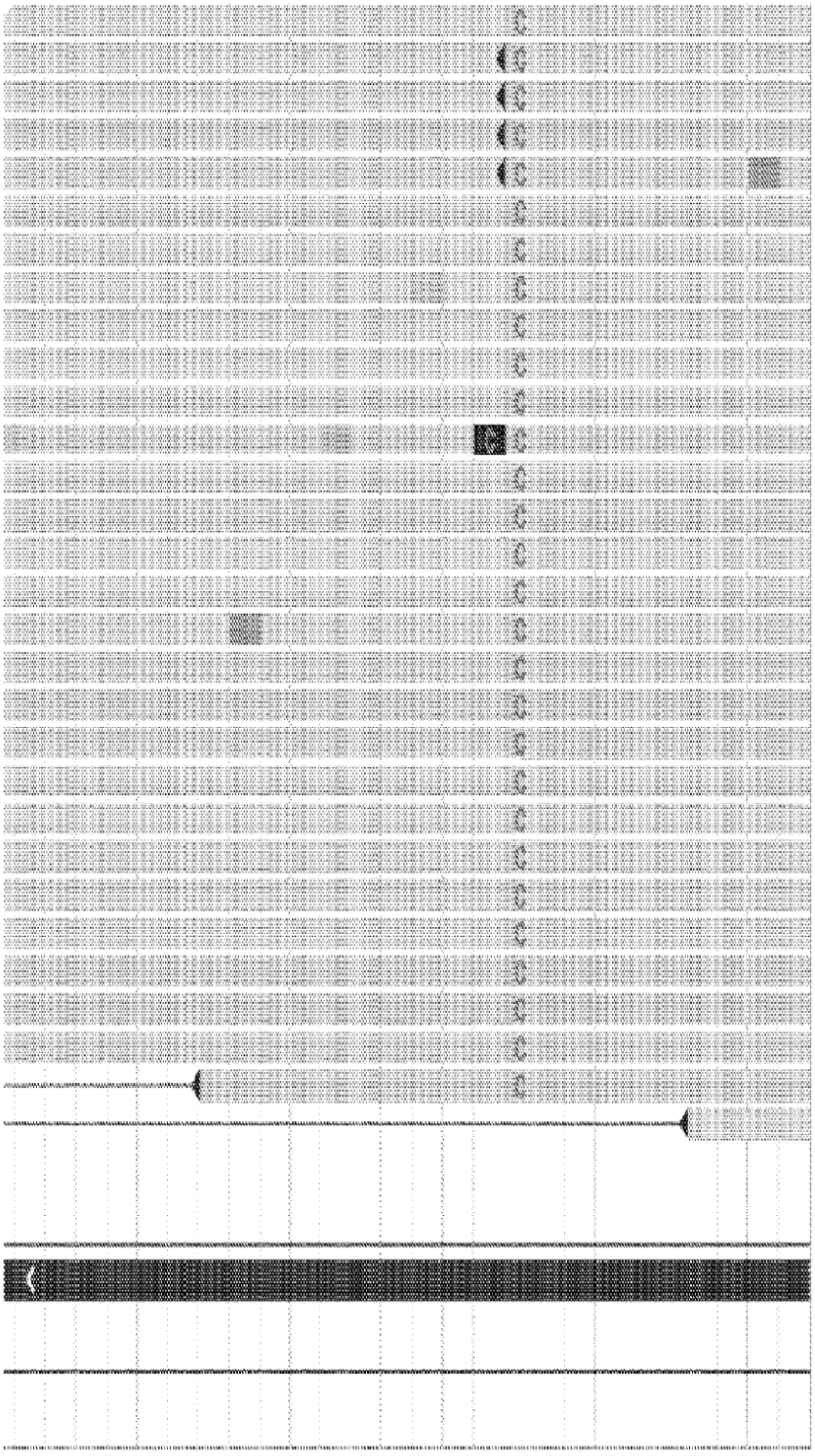
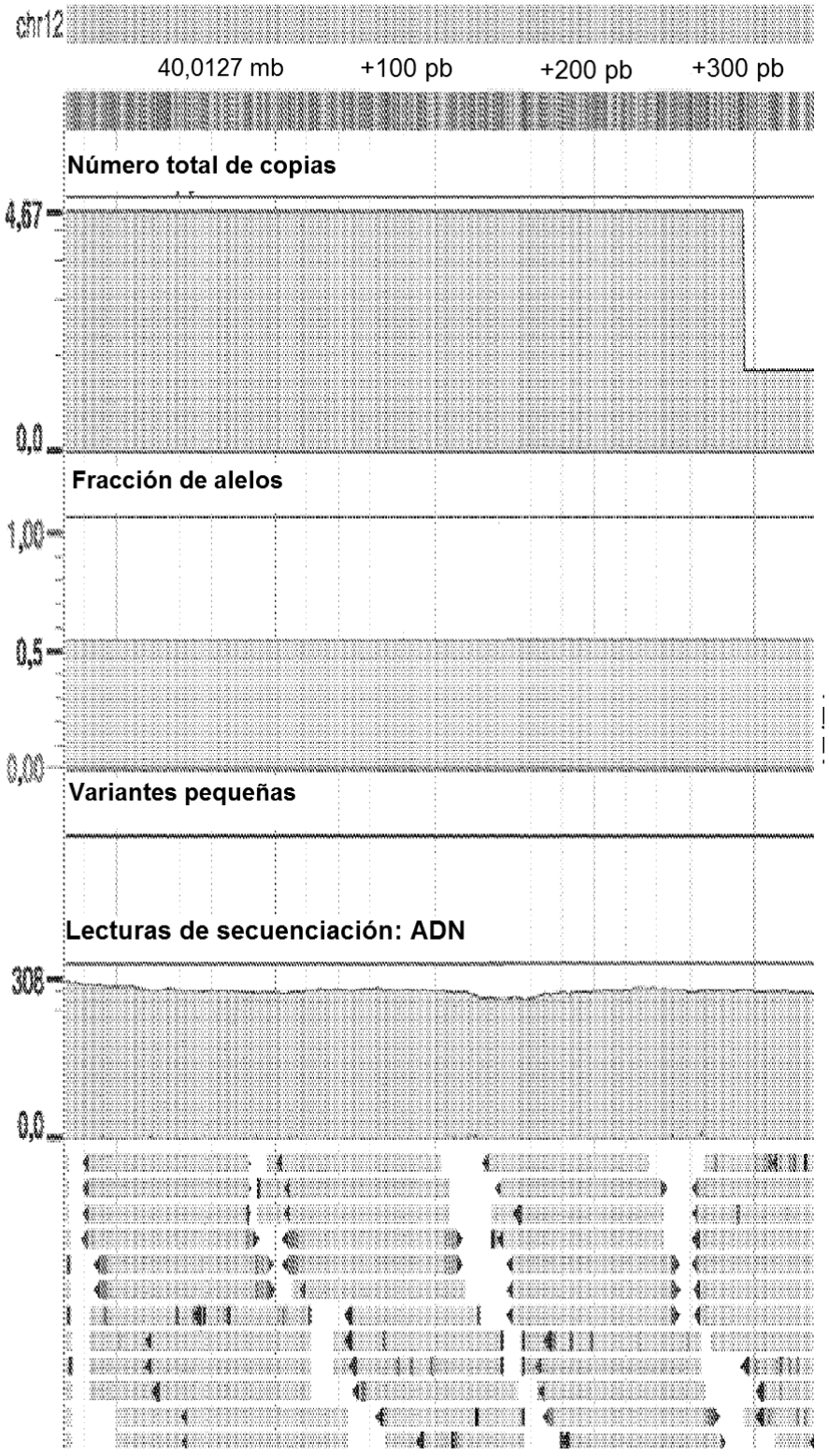


FIG. 3-10

FIG. 4-1


chr12
40,0127 mb
+100 pb
+200 pb
+300 pb
Número total de copias
4,67
0,0
Fracción de alelos
1,00
0,5
0,00
Variantes pequeñas
Lecturas de secuenciación: ADN
308
0,0

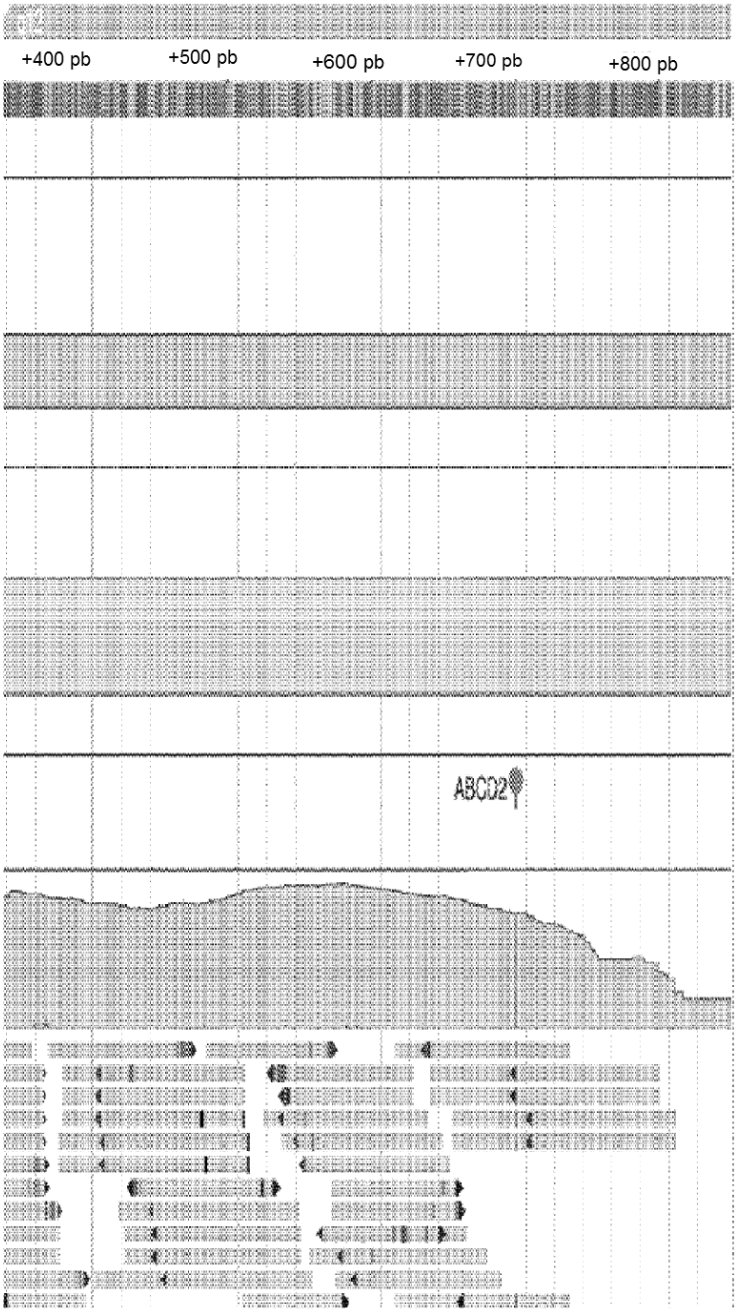
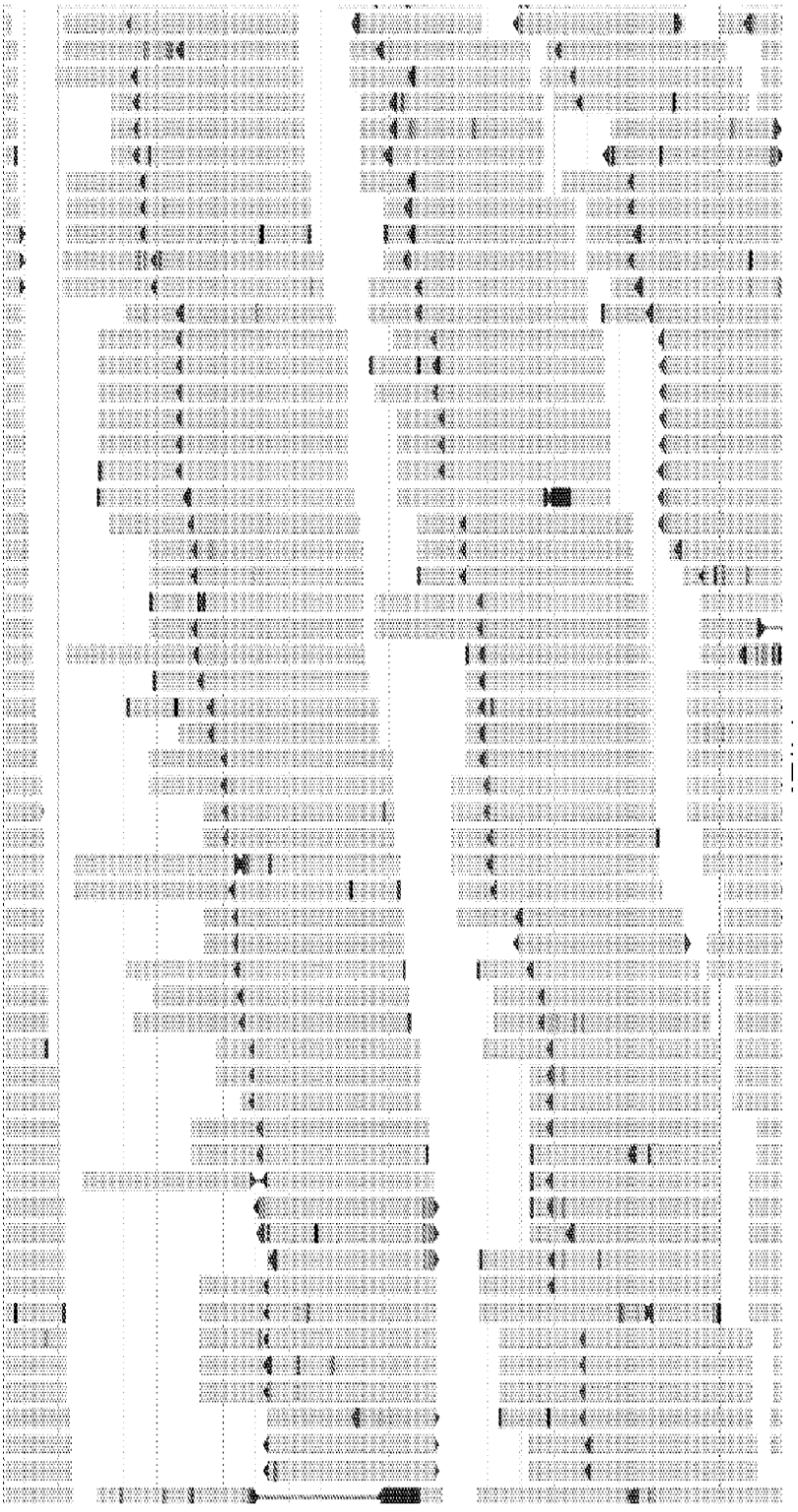


FIG. 4-2

FIG. 4-3



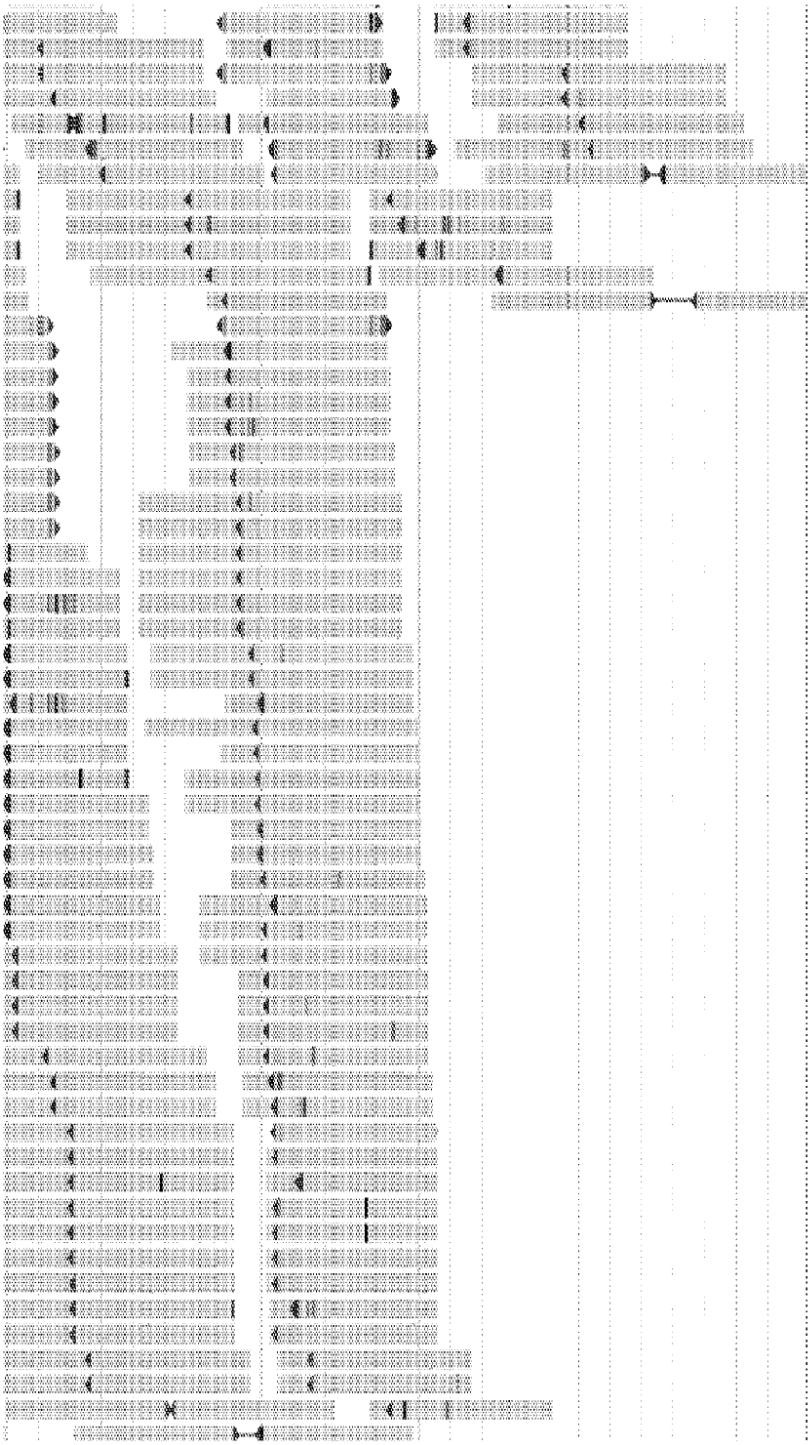
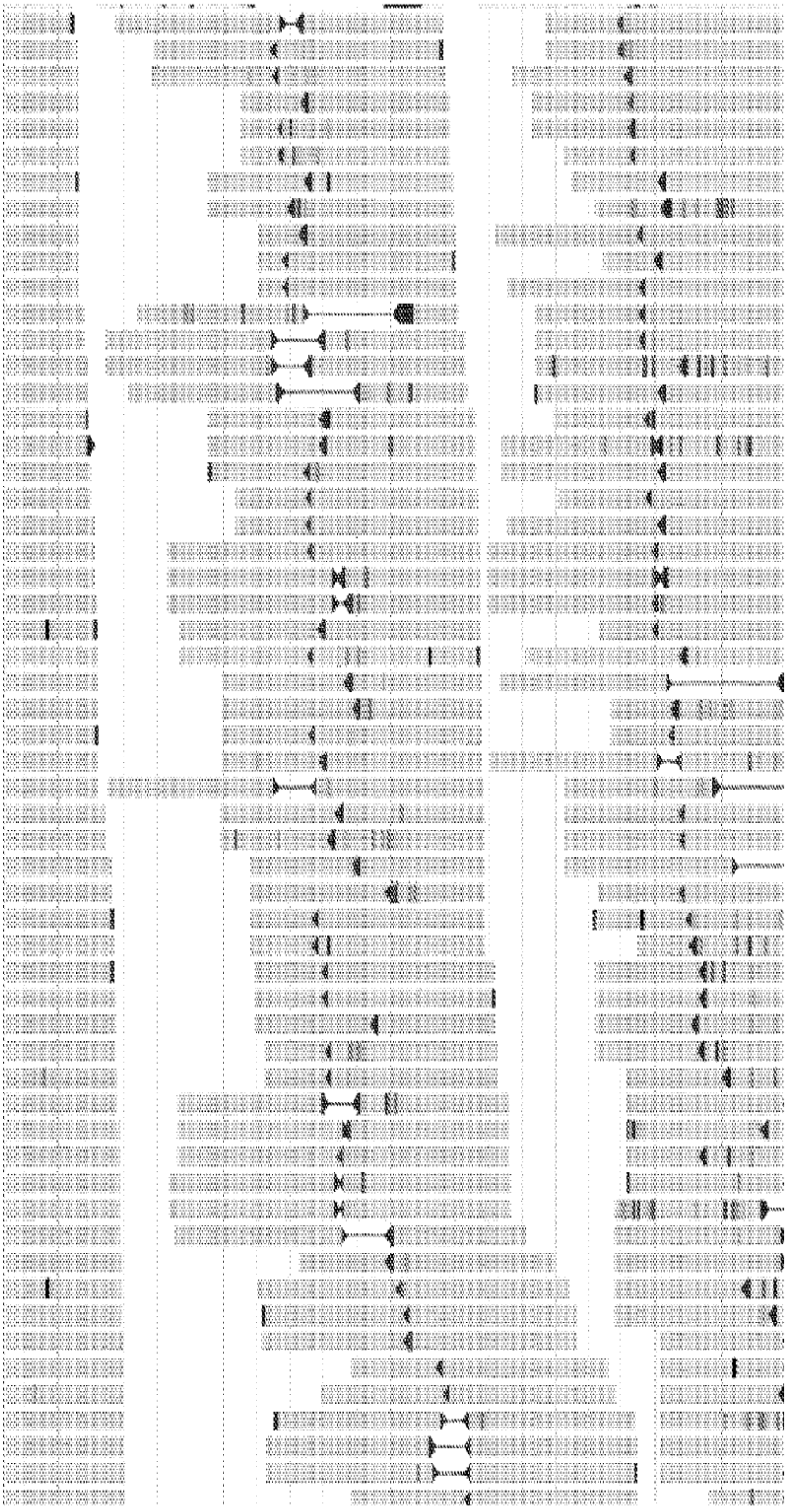


FIG. 4-4

FIG. 4-5



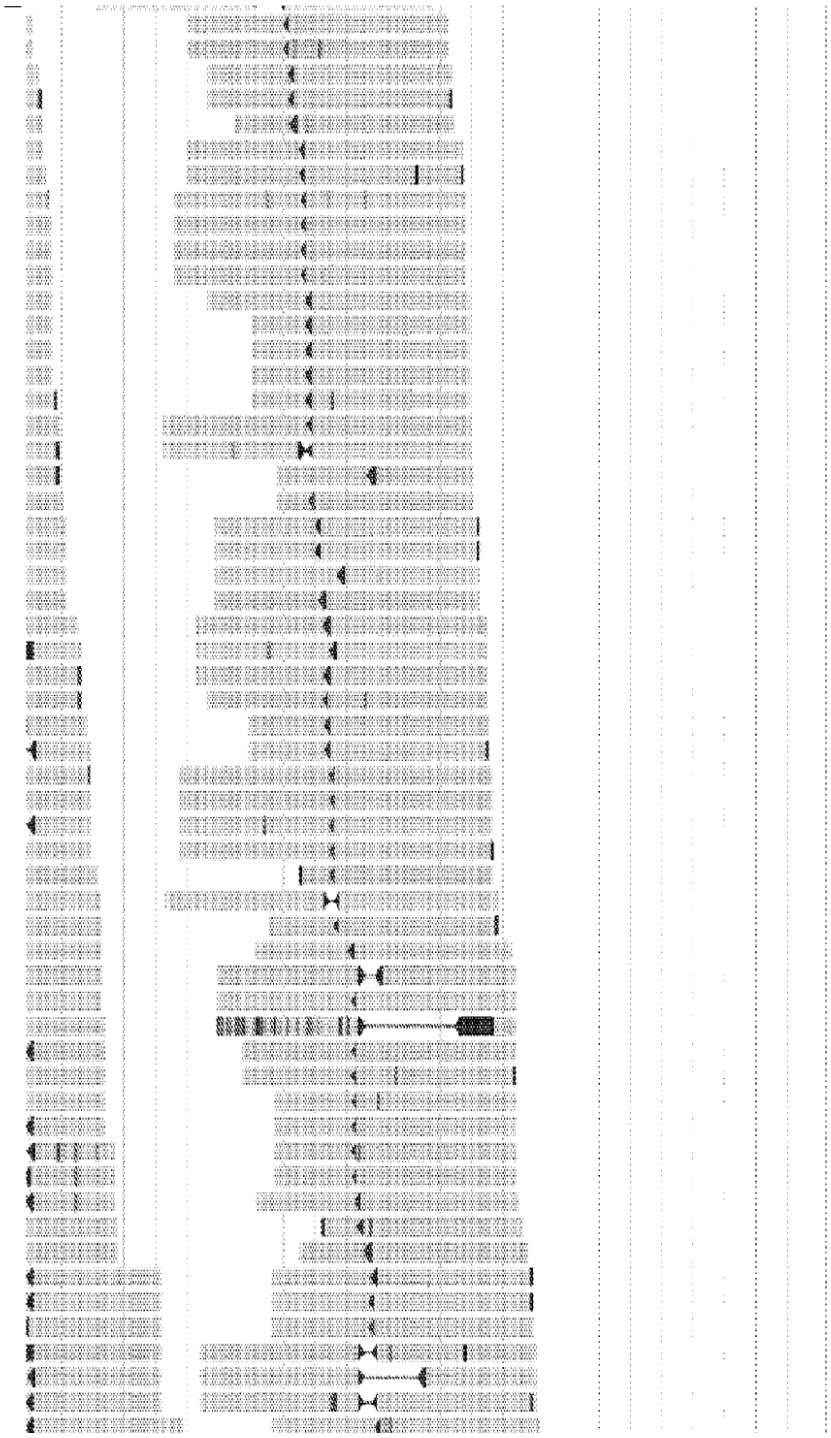
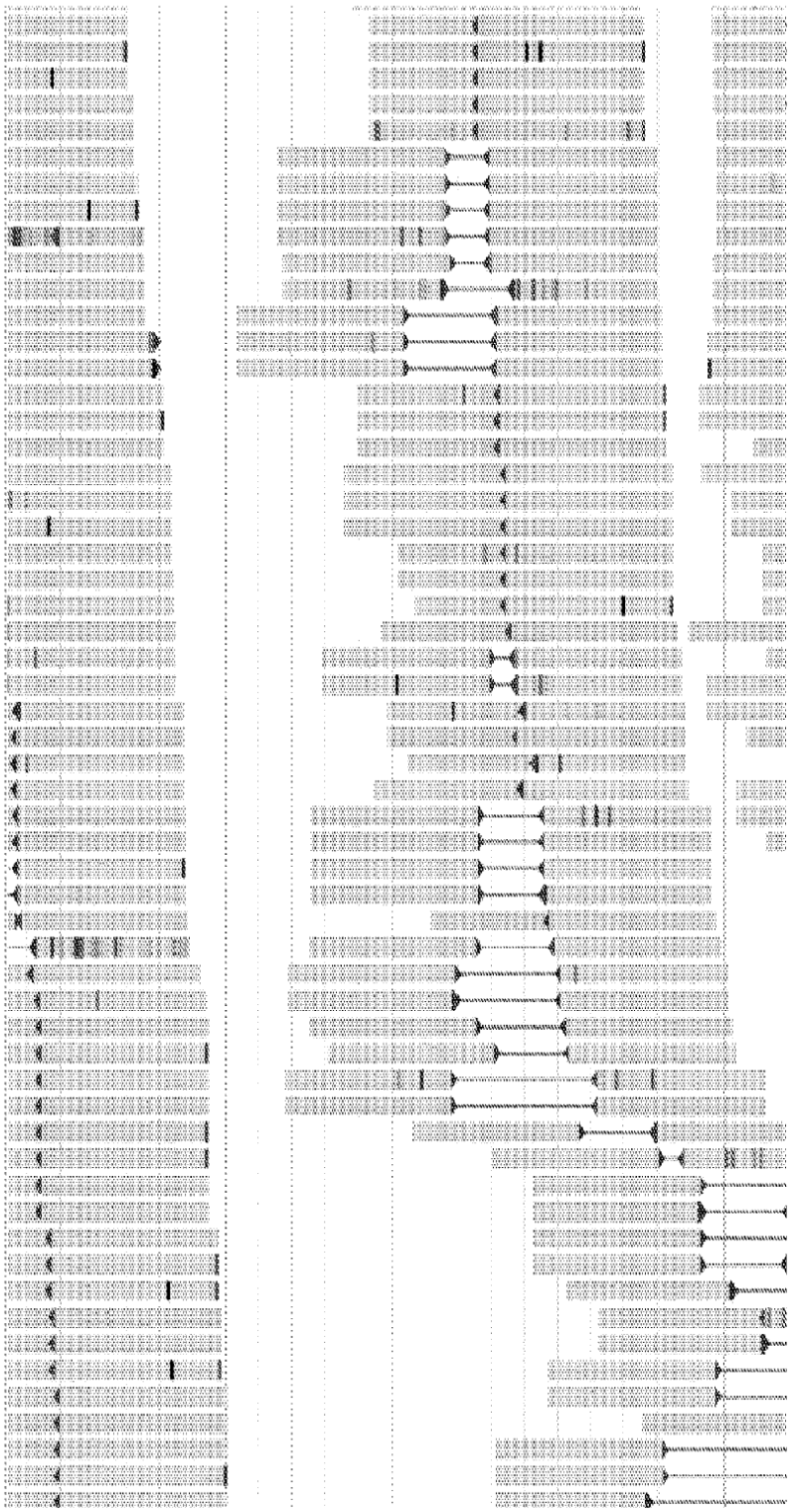


FIG. 4-6

FIG. 4-7



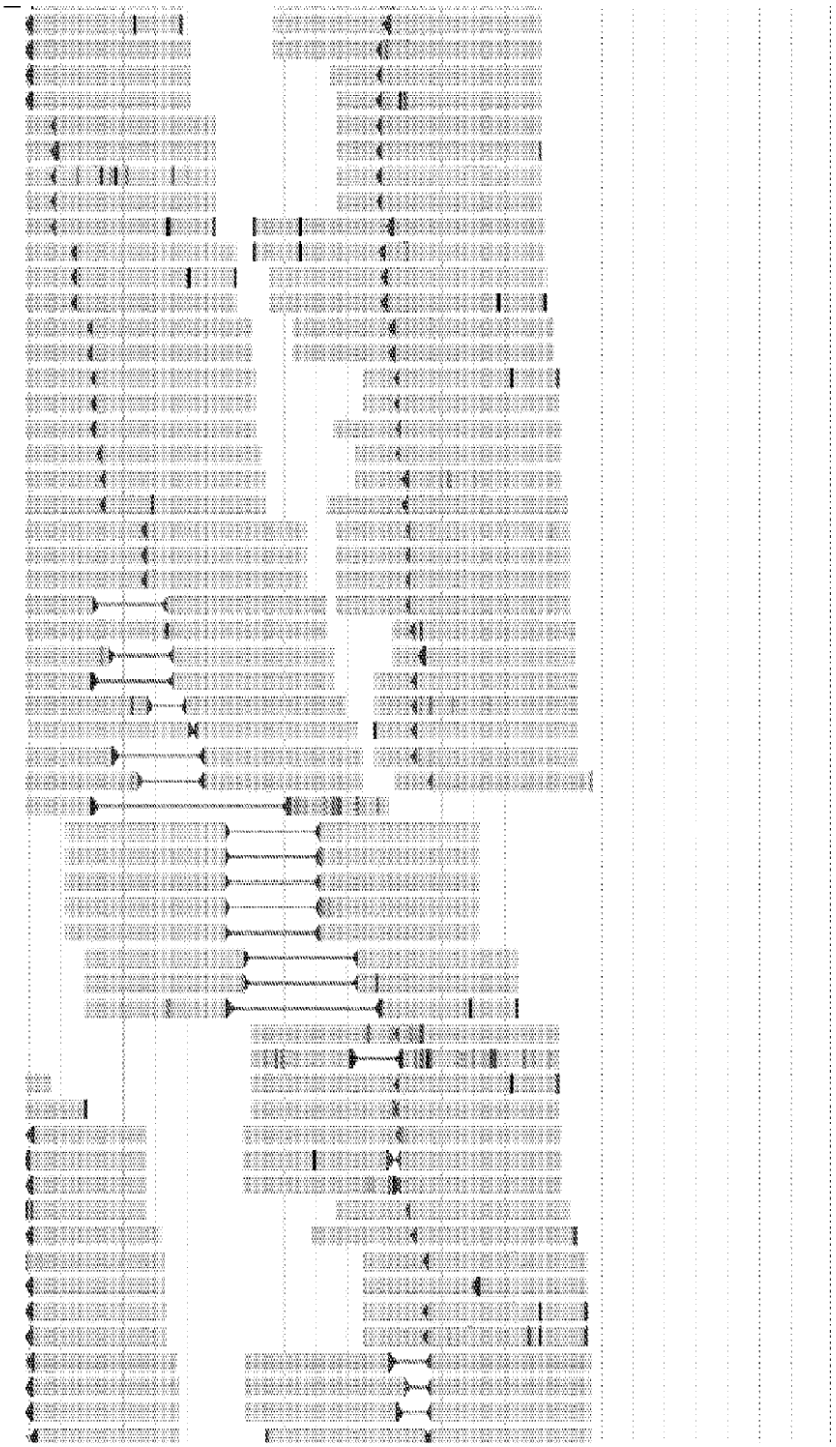
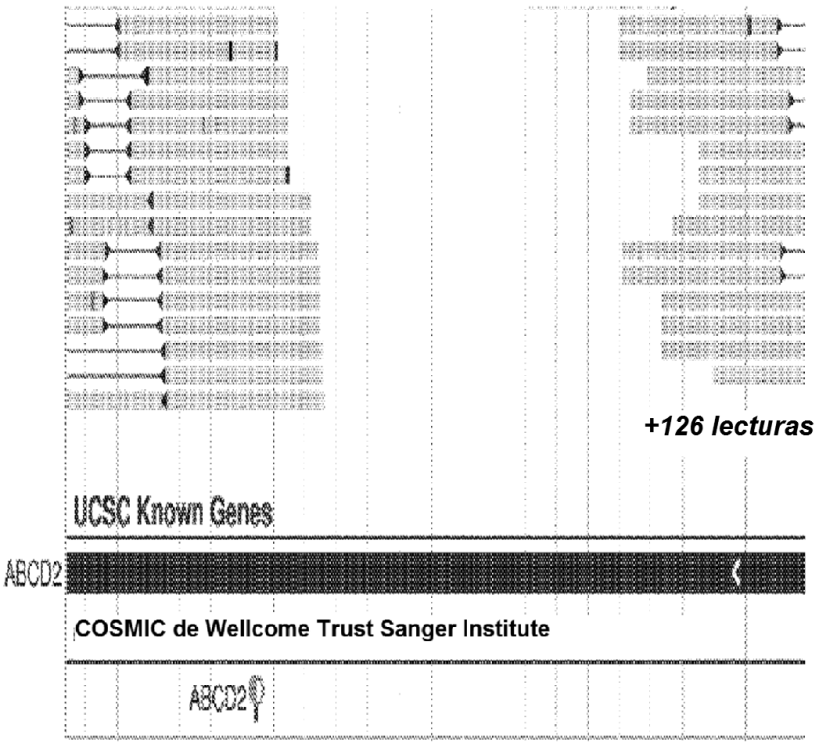


FIG. 4-8

FIG. 4-9


+126 lecturas
UCSC Known Genes
ABCD2
COSMIC de Wellcome Trust Sanger Institute
ABCD2

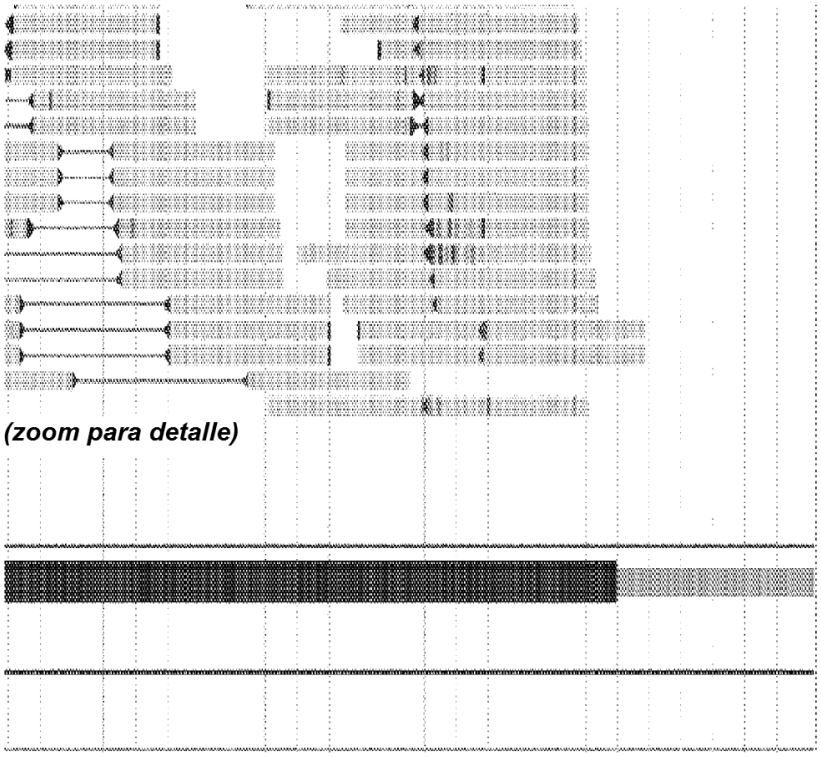
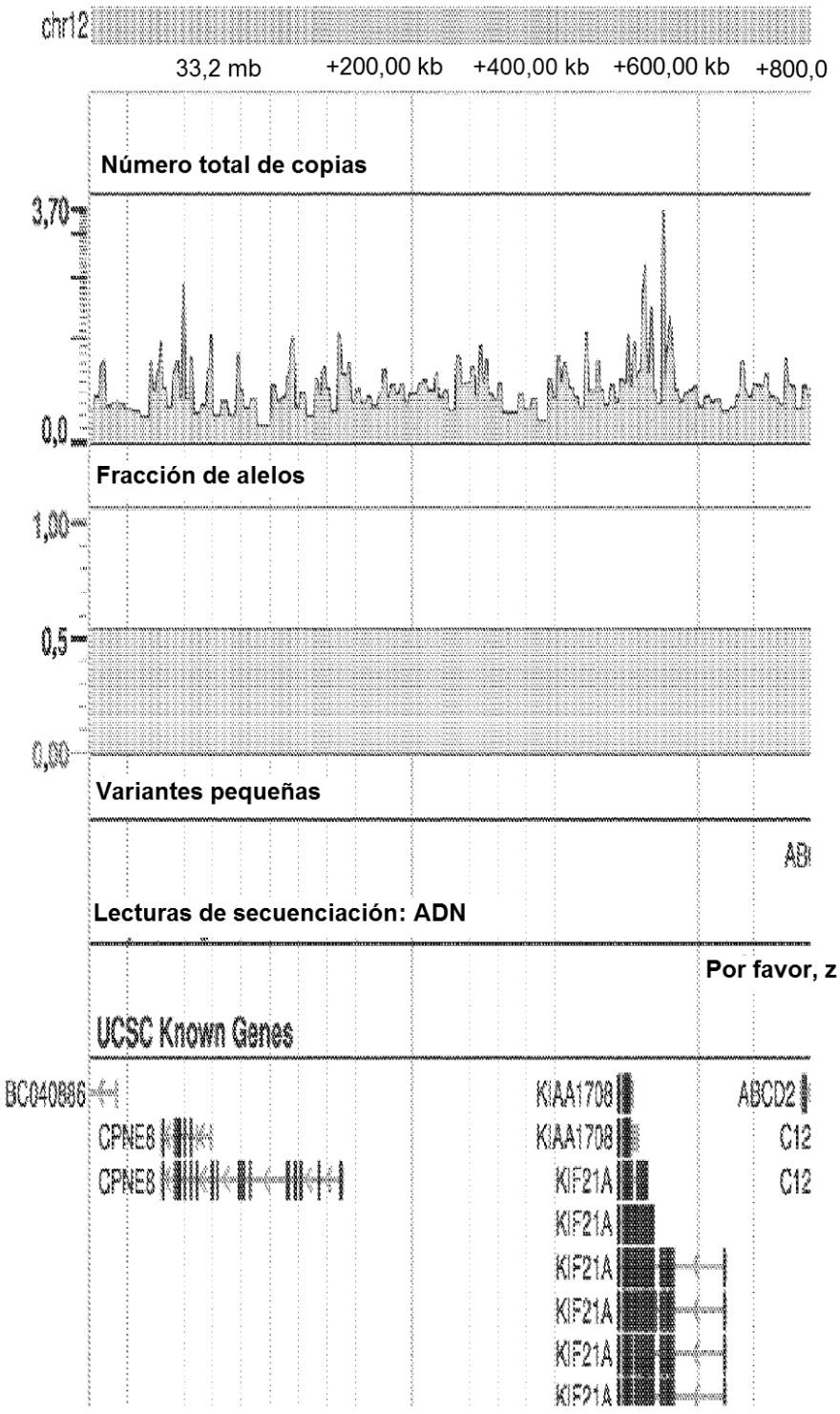


FIG. 4-10

FIG. 5-1


chr12
33,2 mb
+200,00 kb
+400,00 kb
+600,00 kb
+800,0
Número total de copias
3,70
0,0
Fracción de alelos
1,00
0,5
0,00
Variantes pequeñas
ABI
Lecturas de secuenciación: ADN
Por favor, z
UCSC Known Genes
BC040886
CPNE8
CPNE8
KIAA1708
KIAA1708
KIF21A
KIF21A
KIF21A
KIF21A
KIF21A
KIF21A
ABCD2
C12
C12

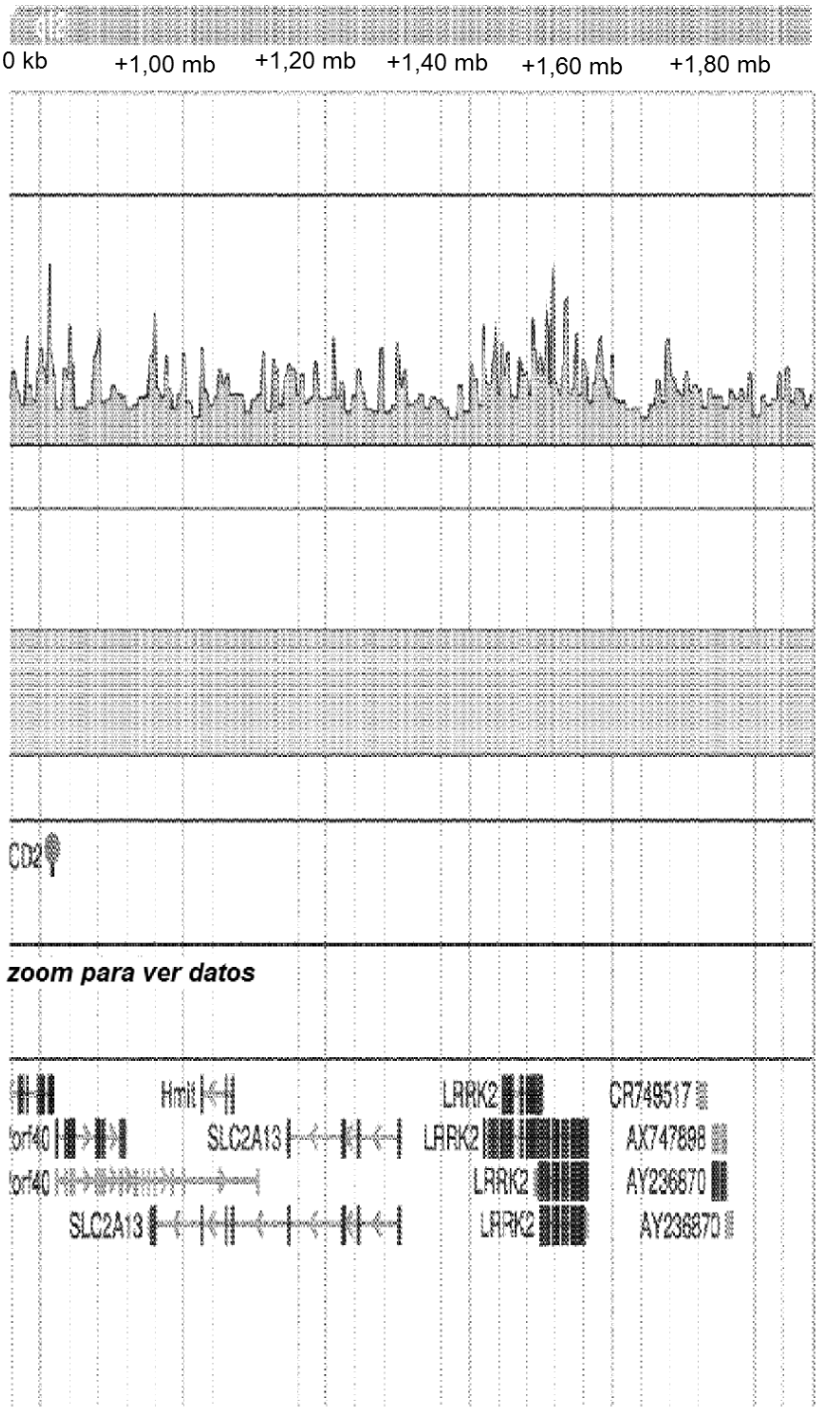
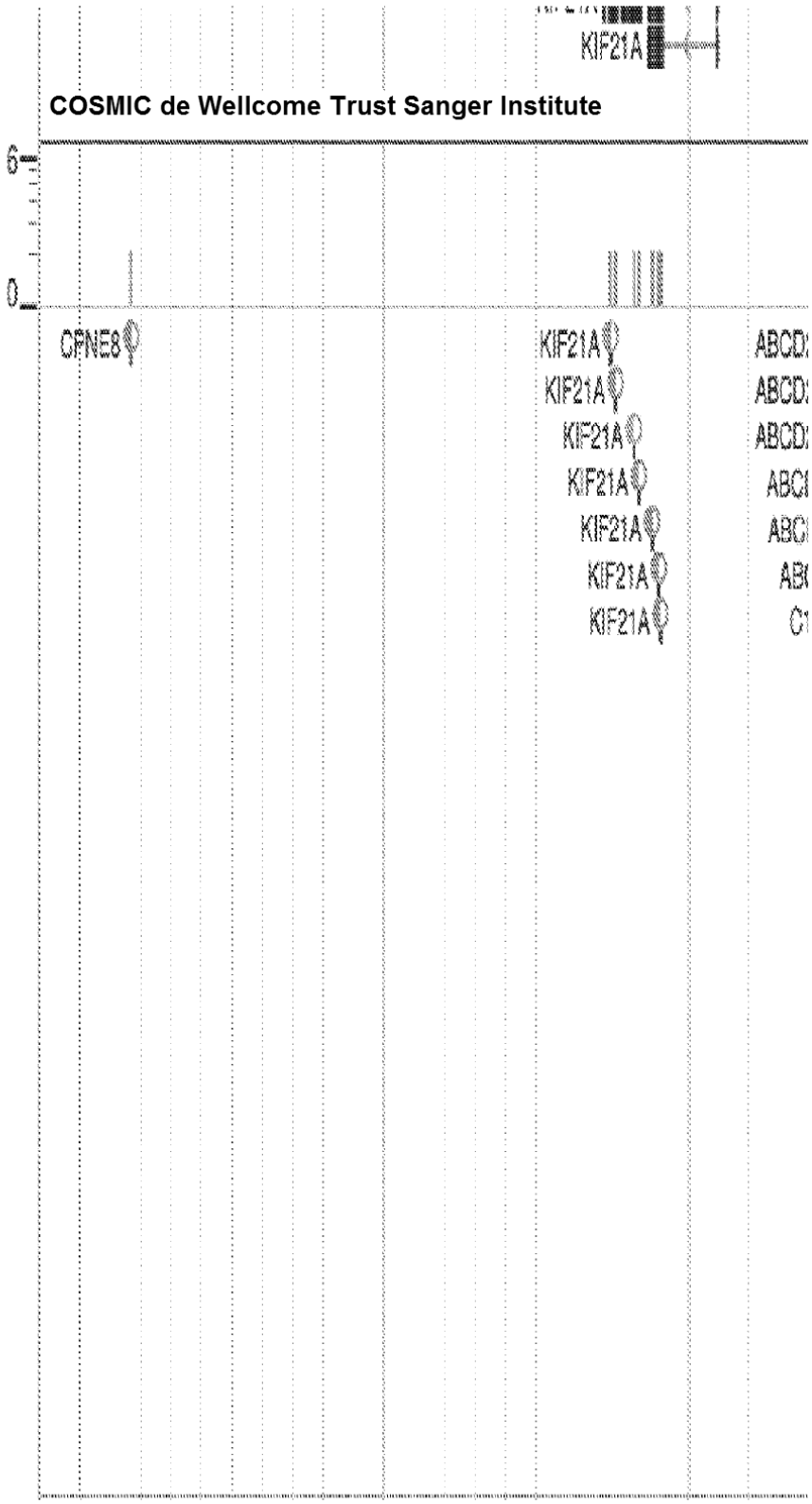

0 kb
+1,00 mb
+1,20 mb
+1,40 mb
+1,60 mb
+1,80 mb
CD2
zoom para ver datos
Hmlt
SLC2A13
LRRK2
LRRK2
LRRK2
LRRK2
CR749517
AX747898
AY236870
AY236870

FIG. 5-2

FIG. 5-3


KIF21A
COSMIC de Wellcome Trust Sanger Institute
6
0
CFNE8
KIF21A
KIF21A
KIF21A
KIF21A
KIF21A
KIF21A
KIF21A

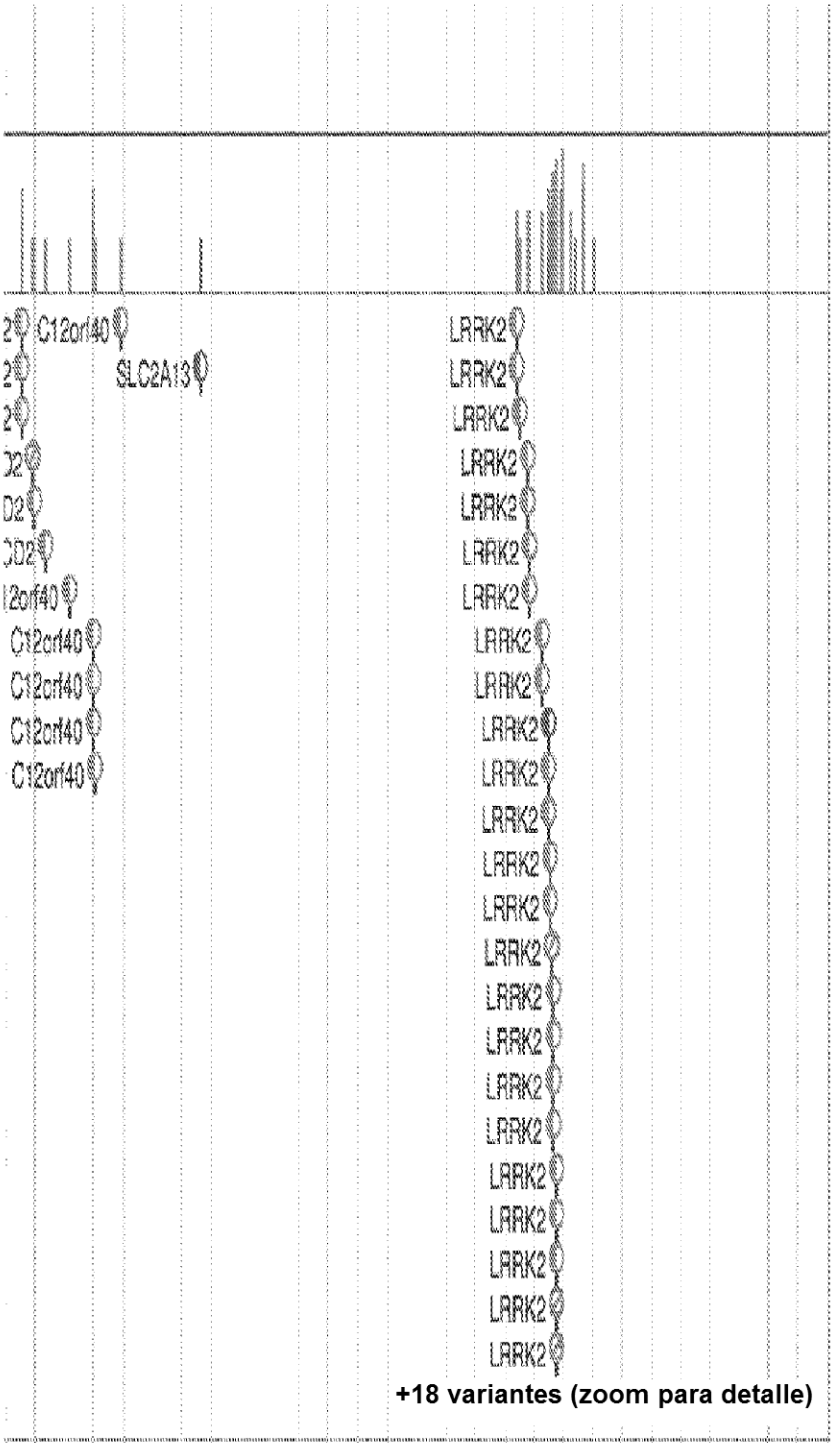


FIG. 5-4

FIG. 6-1

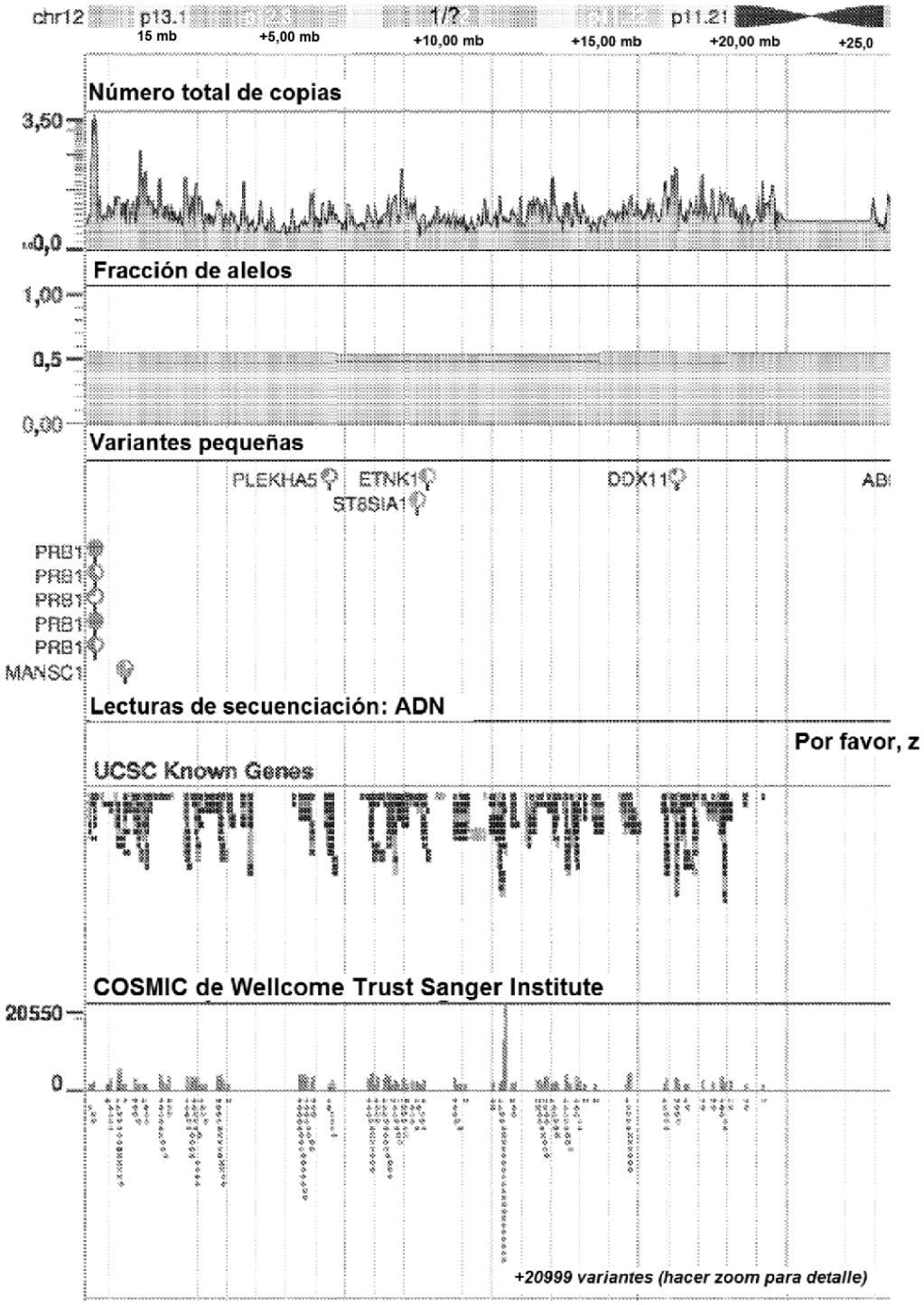
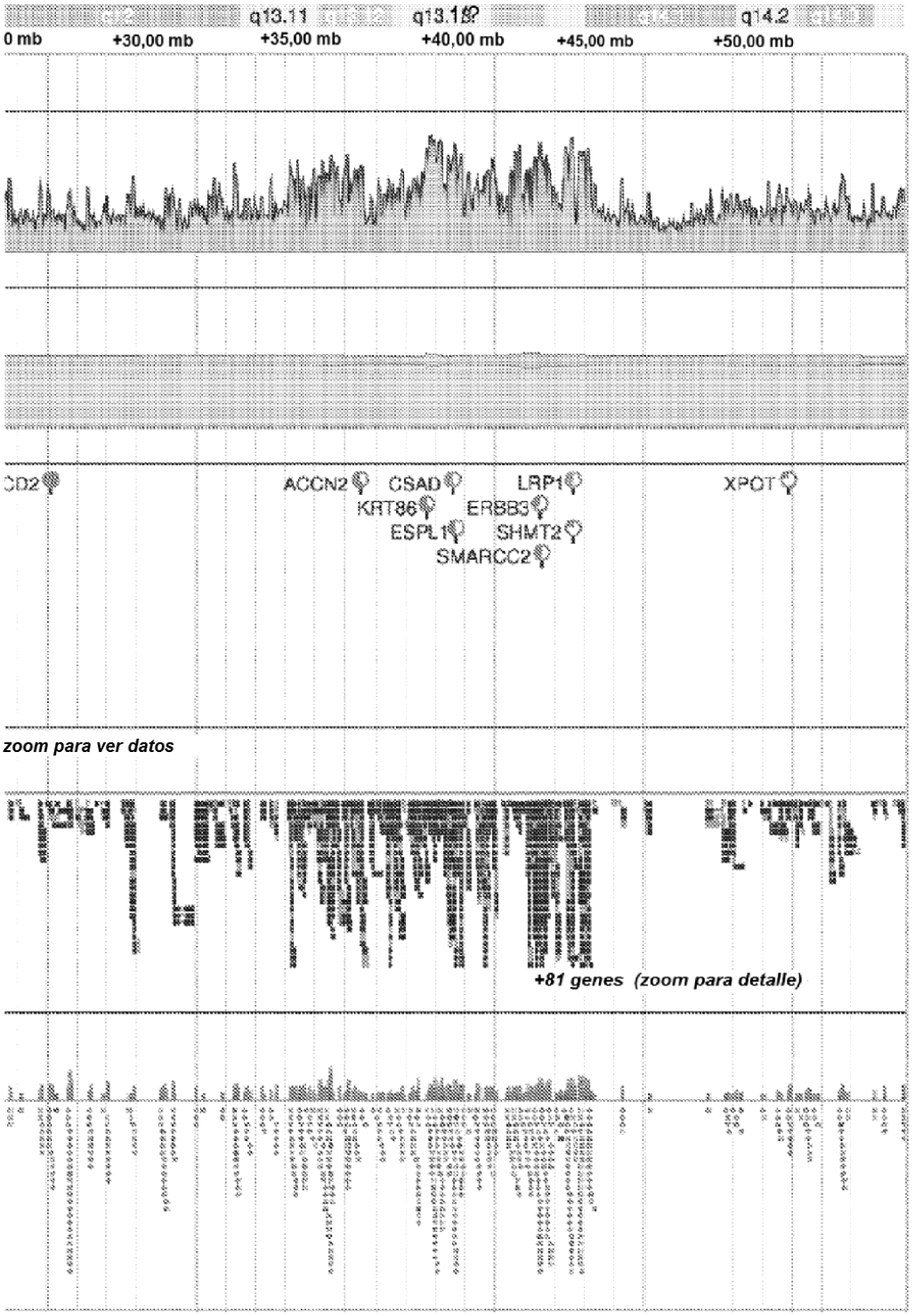

chr12
p13.1
1/?
p11.21
15 mb
+5,00 mb
+10,00 mb
+15,00 mb
+20,00 mb
+25,0
Número total de copias
3,50
0,0
Fracción de alelos
1,00
0,5
0,00
Variantes pequeñas
PLEKHA5
ETNK1
ST8SIA1
DDX11
PRB1
PRB1
PRB1
PRB1
PRB1
MANSC1
Lecturas de secuenciación: ADN
Por favor, z
UCSC Known Genes
COSMIC de Wellcome Trust Sanger Institute
20550
0
+20999 variantes (hacer zoom para detalle)

FIG. 6-2


q13.11
q13.13
q14.2
0 mb
+30,00 mb
+35,00 mb
+40,00 mb
+45,00 mb
+50,00 mb
CD2
ACCN2
CSAD
LRP1
XPOT
KRT86
ERBB3
ESPL1
SHMT2
SMARCC2
zoom para ver datos
+81 genes (zoom para detalle)