Del genoma humano a su pangenoma pasando por los consorcios ENCODE y T2T

por M. GONZALO CLAROS

Catedrático de Bioquímica y Biología Molecular. Universidad de Málaga.

contacto: claros@uma.es

Enviado: 25/4/2022

Resumen: Llevamos 21 años explotando un genoma humano quimérico lleno de huecos. Afortunadamente, sabemos que hay mucho más que genes codificantes gracias a que el consorcio ENCODE lo ha escudriñado a fondo. Entre los hallazgos más inesperados están que más del 75 % del genoma se transcribe, que cada gen sufre 6,3 tipos de ayuste alternativo y fabrica 4 transcritos distintos, que tenemos muchísimas regiones reguladoras repartidas por el genoma y, lo más sorprendente de todo, que desde el punto de vista transcripcional solo tenemos cinco tipos de células diferentes. En estos 21 años no solo hemos mejorado aquella primera secuencia y conocemos mejor para qué sirven las secuencias genómicas, sino que también hemos pasado de una secuenciación indiscriminada preamplificada de lecturas cortas (75 a 600 nt) a una secuenciación molécula a molécula que proporciona miles de bases contiguas. Gracias a ello se acaba de publicar el primer genoma humano completo (salvo el cromosoma Y), no quimérico y sin huecos, que está abriendo nuevas puertas a la genómica y la medicina, incluida la obtención del pangenoma humano que recoja toda la diversidad genética de nuestra especie.

Abstract: We have been exploiting a chimeric human genome full of gaps for 21 years. Fortunately, we know that the genome has more than coding genes because the ENCODE Consortium scrutinised it in depth. Among the most unexpected findings are that more than 75 % of the genome is transcribed, that each gene undergoes 6.3 types of alternative splicing and makes 4 different transcripts, that there are many, many regulatory regions scattered throughout the genome and, the most surprising of all, that there are only five transcriptionally different cell types. In these 21 years, the initial sequence has been improved and a better understanding of what genomic sequences are for was achieved. But research has also moved from indiscriminate pre-amplified sequencing of short reads (75 to 600 nt) to single molecule sequencing that provides thousands of contiguous bases. Consequently, the first complete, non-chimeric, human genome (except for the Y chromosome) without gaps has just been published, opening new doors to genomics and medicine, including obtaining the human pangenome to gather the whole genetic diversity of our species.

> *Las verdades de la razón, una vez consolidadas en nuestra mente, debían prevalecer sobre las creencias irracionales que convenían a los ignorantes y les proporcionaba consuelo.*
>
> – Ramón Muñoz-Chápuli, *El sueño del anticristo*

Cómo llegamos al primer genoma humano:

El 25 de abril de 1953, hace nada menos que 69 años, Watson y Crick publicaron en *Nature* el artículo donde se describían las cuatro características incuestionadas de la estructura del DNA [35]. Comenzaba con un párrafo que para algunos sería hoy una herejía al usar la primera persona del plural en inglés en lugar de la infame voz pasiva:

> *We wish to suggest a structure for the salt of deoxyribose acid (D.N.A.). This structure has novel features which are of considerable biological interest.*

Hace «tan solo» 21 años, el 15 de febrero de 2001, impulsado por el mismo Watson, se publicó en *Nature* el primer borrador del genoma humano gracias al Consorcio Internacional para la Secuenciación del Genoma Humano [18]. Ese mismo día se publicaba en *Science* otro borrador de otro genoma elaborado por Celera Genomics de la mano de Craig Venter [33]. Ambos borradores estaban llenos de huecos e incertidumbres. Muy poco después, en 2004, se dio el primero por completado [15] a pesar de que solo abarcaba el 96 % de la eucromatina. Nota deprimente: ni un solo laboratorio español contribuyó en nada a este primer genoma humano.

Poco a poco se ha ido mejorando y refinando gracias a los avances de la tecnología de secuenciación y la bioinformática, hasta obtener la versión actual (GRCh38) elaborada por el Genome Reference Con-

sortium (GRC) en 2013 y revisada por última vez en 2019. El GRC ya veía poco margen de mejora porque la tecnología de secuenciación indiscriminada (→ *shotgun sequencing*) de lecturas cortas no daba más de sí.

Cualquier investigador tiene claro que la secuencia de un genoma no basta: para que resulte útil debemos conocer qué hay en ella y, a ser posible, con una resolución nucleotídica. Desde el primer borrador de 2001, quedamos sorprendidos (y los homocentristas deprimidos) porque:

1. solo teníamos entre 30 000 y 40 000 genes, muchos de los cuales se parecerían a los de otros vertebrados e incluso de las bacterias (¡a ver si vamos a ser transgénicos!);
2. la mitad del genoma derivaba de transposones;
3. se detectaron 1,4 millones de polimorfismos (para unos, mucho, pero poco para otros); y
4. había más seudogenes y RNA no codificantes (que por entonces no se entendían muy bien qué pintaban ahí) de los que nos parecían razonables.

Cuando se completó el primer ensamblaje en 2004[15] la situación fue a peor porque se estimó que la parte codificante de nuestro genoma no superaba el 2 %, muy lejos del 10 % que los más optimistas proponían a finales del siglo XX. También se rebajó significativamente el número de genes codificantes a tan solo 20 000 (nada que ver con los 100 000 que se suponía que teníamos antes de ponernos a secuenciar).

La versión más reciente del GRCh38 es de 2019 (GRCh38.p13) y sus principales rasgos son:

1. 3 096 649 726 pb de longitud;
2. 20 442 genes codificantes (aunque se le han predicho 51 756);
3. 23 982 genes no codificantes (la mayoría [16 896] largos);
4. 15 228 seudogenes;
5. 237 081 transcritos diferentes sintetizados; y
6. 84 277 proteínas distintas traducidas, un número curiosamente cercano a la creencia inicial de que teníamos 100 000 genes distintos porque, ilusos de nosotros, estábamos convencidos de que un gen solo daba una proteína.

En febrero de 2022 se acaba de liberar una nueva revisión (y van 29) que responde a la secuencia GCA_000001405.29 del futuro GRCh38.p14 con 208 688 nucleótidos menos. O igual no...

¿Que hay en un genoma con tan pocos genes?:

Como un ridículo 2 % del genoma codificaba una cantidad tan irrisoria de genes, los científicos se preguntaron rápidamente qué había en ese 98 % no codificante que denominábamos peyorativamente DNA basura (→ *junk DNA*) con ese homocentrismo típico que nos hace pensar que lo que no conocemos no tiene ningún valor, como ocurría en su día con las amígdalas, la apéndice, las malas hierbas, las alimañas, y tantas otras cosas (ojo, que hay quien propone recobrar este concepto y añadir otro denominado *spam DNA* para esas secuencias que perduran en el genoma de las especies sin ninguna utilidad aparente[9]). De ahí nació en setiembre de 2003, impulsado por el National Human Genome Research Institute (NHGRI) estadounidense, el consorcio ENCODE (Encyclopedia of DNA Elements → Enciclopedia de Elementos del DNA) con el objetivo de catalogar los elementos funcionales y entender qué razones hay para replicar una y otra vez tanto genoma lleno de «basura». En 2007, el puñado de investigadores de laboratorio y bioinformáticos que intervino en la **fase piloto (ROADMAP)** del proyecto ENCODE publicó los resultados sobre un 1 % del genoma (unas 30 Mb) de unas pocas líneas celulares en cultivo[7].

En primer lugar, catalogaron las regiones del genoma que se transcribían en algún tipo de RNA (tanto codificante como no codificante). Llamó la atención que prácticamente todo el genoma se transcribiera (incluidas regiones que se creían silenciosas), y que había zonas donde se transcribían ambas hebras. A continuación buscaron las secuencias que regulaban tanto transcrito y vieron que, como cabía esperar, eran muy accesibles a las exonucleasas, a factores de transcripción y a enzimas modificadoras del DNA, y recibieron el nombre de **CRE** (*cis-regulatory elements* → regiones reguladoras en *cis*). No hay que confundir estas CRE con las secuencias CRE procariotas (*cAMP response elements* → elementos de respuesta al AMPc[25]), aunque yo creo que en el subconsciente de los autores sí había alguna relación entre ellas. Como comprobaron que los CRE estaban distribuidos simétricamente en torno a los inicios de transcripción, sin ningún tipo de sesgo hacia un lado u otro, los clasificaron en dos grandes grupos:

- promotores (→ *promoters*) cuando se sitúan sobre el inicio de transcripción;
- potenciadores (o intensificadores, → *enhancer*) cuando están muy alejados del inicio y son más sensibles a la DNasa I.

Otra importante consecuencia de este estudio piloto la sufrió el concepto de gen, que obligó a redefinir como 'unidad mínima heredada' debido a que gran parte del genoma que hasta entonces se creía que no valía para nada se dedicaba a regular, con consecuencias fenotípicas nada despreciables (incluidas las enfermedades). Todas estas revelaciones despertaron aún más el interés por ese genoma no codificante, así que se acabaron implicando en el proyecto 440 investigadores repartidos por 32 centros de investigación del planeta (esta vez sí participaron grupos españoles).

El éxito del estudio piloto de ENCODE también impulsó el proyecto GENCODE (*encyclopaedia of genes and gene variants* → enciclopedia de genes y variantes génicas) liderado por The Wellcome Sanger Institute en Hinxton (Reino Unido) para identificar y localizar todos los posibles genes del genoma humano y del ratón mediante la combinación de análisis computacionales, anotación manual y validación experimental. Los primeros resultados aparecieron en 2006[13], gracias a lo cual llevan encadenados varios exitosos proyectos, con la incorporación de investigadores de todo el mundo. Ambos consorcios siguen activos, se intercambian información, y todas sus anotaciones se ofrecen desde sus portales de la Enciclopedia ENCODE y del proyecto GENCODE, y se incorporan en las bases de datos de Ensembl y el UCSC Genome Browser, así como en su espejo europeo.

ENCODE entra en la fase de producción:

En 2012, la sociedad y los científicos nos quedamos sorprendidos por la publicación simultánea de 5 artículos en *Nature*, más otros 6 en *Genome Biology* y otros 18 en un número especial de *Genome Research*, sobre la segunda fase de ENCODE[6]. En ella, extendieron lo puesto a punto en fase piloto por todo el genoma con muchas más (147) líneas celulares de humano y se generaron 1 640 conjuntos de datos.

Lo más sobresaliente fue la confirmación de que se transcribe el 75 % del genoma, una transcripción que se solapa tanto en las regiones codificantes como en las no codificantes. Estimaron que, de promedio, cada gen sufría 6,3 tipos de ayuste alternativo (→ *alternative splicing*) y fabricaba 4 transcritos distintos. También le dieron función (principalmente reguladora) al 80,4 % del DNA que hasta entonces se consideraba ⊗~~basura~~, lo que supuso otro empujón al destierro de este término del vocabulario genómico). Se estimó que había 70 292 promotores para 20 687 genes (estos valores van variando en las nuevas versiones), cuya regulación dependía de 399 124 potenciadores (de los que solo unos 200 000 parecen activos en cada tipo celular). Se localizaron 8 801 ncRNA y 9 640 lncRNA, así como 11 224 seudogenes, de los que se transcribían solo 863 (¡o nada menos!). Resultó también sorprendente que muchísimos polimorfismos asociados a una enfermedad no se hallaran en las regiones codificantes, sino fuera de ellas, cerca o dentro de los CRE. De hecho, se identificaron más de 8,4 millones de secuencias a las que se unía algún factor de transcripción (para que nos hagamos una idea de su tamaño, ocupan el doble de secuencia que el exoma humano). Hasta ENCODE, para explicar los parecidos entre especies y las enfermedades se ponía el foco en las regiones codificantes de los genes, pero ahora empieza a centrarse en las secuencias reguladoras.

Impulsados por unos hallazgos tan reveladores y llamativos, muchos investigadores se plantearon que se deberían aplicar las mismas estrategias a otros seres vivos. Así, en 2014 aparecen los resultados para el genoma de ratón[38] basados en 100 tipos de células y tejidos. La gran noticia fue que no solo se confirmaba todo lo novedoso del genoma humano sobre las secuencias funcionales, sino que la organización a gran escala de ambos genomas era muy parecida. Resultó especialmente llamativo que se conservara bastante la red de factores de transcripción reguladores, mientras que divergieran los CRE reconocidos en cada especie. También se confirmó que las secuencias con las que se distinguía mejor un genoma humano del de ratón descansaban en las secuencias repetitivas, no en las codificantes ni en las únicas. Por tanto, ya **empezamos a explicarnos cómo actúan las fuerzas evolutivas sobre los genes y su regulación**, y qué mecanismos de las enfermedades humanas compartimos con otros mamíferos.

Surgen los primeros disidentes:

No toda la comunidad científica admite las conclusiones de ENCODE tal como se cuentan[11]. Muchos defienden que buena parte de esa enorme cantidad de transcritos carece de actividad o utilidad biológica, que no son más que errores de la RNA—polimerasa, que reconoce como promotores secuencias que no lo son. La crítica más sólida indica que se abusa del concepto de función al afirmar que basta con que se cumpla alguna de las siguientes 'circunstancias':

1. se transcribe (eso implica que todos los intrones son funcionales),
2. está en una zona accesible de la cromatina,
3. se le acoplan factores de transcripción, o
4. contiene dinucleótidos CpG metilados

Los detractores de ENCODE afirman que esta visión simplista de una secuencia reguladora no implica que sirva para regular, ni tan siquiera que lo haga. A diferencia de los negacionistas de la COVID-19, estos críticos aportan un respaldo experimental serio[11]: en el último siglo, la genética ha demostrado que sólo el 10 % del genoma humano se ha conservado evolutivamente gracias a la selección, por lo que hay un 70 % de dudosa utilidad dado que las mutaciones en él no parecen alterar el funcionamiento de la célula.

ENCODE 3 durante la pandemia:

En julio de 2020, en plena pandemia de COVID-19, se publicó la tercera fase del proyecto ENCODE para humano y ratón en la que han intervenido unos 500 científicos de todo el mundo, incluida España[8]. Como en el caso anterior, se publican de golpe todos los artículos, 9 de ellos en la revista *Nature*, y 5 más en otras del grupo. La lista completa junto con el resumen de los hallazgos se puede consultar en `http://go.nature.com/encode`. El estudio se ha completado con datos procedentes de 5 992 experimentos realizados con 13 069 muestras de células tomadas directamente de 503 tipos de células y tejidos de humano y de ratón. También hay otra novedad: se caracterizan **los CRE que aparecen en los ARN**, así como la **estructura 3D de la cromatina celular** (formación de bucles de cromatina que acercan los CRE a los genes que regulan). Pero por mucho que se expanda el catálogo de CRE, dado que muchos actúan solo en algunos tipos celulares o en momentos concretos, seguimos sin saber si ya los conocemos todos o si siguen faltándonos algunos. No obstante, las anotaciones de ENCODE ya se han convertido en la herramienta por antonomasia con la que conocer la regulación génica y la predisposición genética a las enfermedades. Por eso ya está ya en marcha la cuarta fase del proyecto en la que se incorporarán nuevas tecnologías (sobre todo las basadas en analizar las células una a una y la genómica funcional de alto rendimiento) y más tipos de células (incluidos los tejidos y las enfermedades poco frecuentes). Mientras tanto, vamos a comentar brevemente estos últimos hallazgos tan relevantes.

Registro exhaustivo de elementos funcionales del genoma. Este catálogo se puede consultar y descargar en el portal SCREEN, donde se recogen más de 1 200 000 cCRE (*candidate cis-regulatory elements* → candidatos a regiones reguladoras en *cis*) en el genoma de humano (929 535) y de ratón (339 815), que representan el 7,9 % y el 3,4 % del genoma de cada especie, respectivamente. Además, con la integración del estudio *in vivo* e *in vitro* de la interacción de las proteínas con el RNA han podido determinar el efecto que algunos CRE de ARN tienen sobre la estabilidad del transcrito y sobre el ayuste alternativo. También fue sorprendente observar que casi la mitad de las proteínas que se fijaban al RNA interaccionaban también con el DNA, aunque no sobre la misma secuencia. Seguro que ya estás pensando en la cantidad de conocimientos básicos sobre los procesos biológicos que aporta este registro, que también servirá para conocer mejor la salud y las enfermedades, y a destripar cómo los gobiernan los cCRE.

Los CRE y la topología del genoma. La mayoría de los CRE están ocupados por varios factores de transcripción que se unen de manera independiente y bien espaciada. Pero no todos los CRE están igual de ocupados, sino que entre promotores y potenciadores hay unos 5 000 CRE «calientes» (por el acrónimo HOT (*highly occupied target* → dianas muy ocupadas) en claro juego de palabras con los conocidos puntos calientes (→ *hot spot*) de los procariotas) en los que se fijan muchísimos factores de transcripción. Las interacciones de los CRE de la cromatina y los RNA con las proteínas forman bucles que pueden ser transitorios, específicos o característicos. Estos bucles acercan los CRE a los genes que regulan, por lo que su alteración cambiará la expresión génica. También se ha demostrado que los genes de mantenimiento están sujetos a la regulación de pocos CRE: su expresión estable y constante se ve favorecida por los circuitos de regulación sencillos. En cambio, se necesita la coordinación más compleja de muchos CRE para la regulación fina de ciertos genes. Sorprendentemente, con los bucles también se regula el ayuste de los genes para determinar qué exones e intrones se van a retener en el transcrito maduro.

Los datos que se conocen de los ratones en las situaciones que se hacen difíciles de estudiar en los humanos han demostrado lo valioso que resulta el estudio de los CRE en los animales. Por ejemplo, en los fetos murinos se ha observado que, a medida que avanza el desarrollo, se van desmetilando los CRE de la cromatina para instaurar nuevos modos de regulación génica rápidos y flexibles por la modificación de las histonas y por la accesibilidad de la cromatina. Así se espera conocer las bases moleculares de cCRE y genes que pueden ser responsables de los trastornos del desarrollo en los humanos, puesto que se ha visto que en los equivalentes humanos de los cCRE murinos de desarrollo se ubican muchas variantes asociadas a enfermedades relevantes.

Solo hay cinco tipos de células. En una de las publicaciones, esta vez en *Genome Research*, el grupo de investigación de Thomas Gingeras del Laboratorio Cold Spring Harbor (CHSL) de EE UU y el de Roderic Guigó del CRG de Barcelona concluyen que, en función del perfil de expresión de los genes (lo que denominamos el transcriptoma), en el cuerpo humano solo tenemos, contra todo pronóstico, cinco grupos de células diferentes, en lugar de presentar tantos perfiles como tejidos (que es lo que se esperaba). También comprobaron que la composición del transcriptoma tisular cambia con la edad, el sexo, y las enfermedades (esto, en cambio, no sorprendió nada). Está claro que se podrá conocer qué ocurre dentro de una célula normal y de una enferma, envejecida prematuramente, etc., y distinguir así entre los individuos enfermos y los saludables. También se han abordado estudios de perfiles de expresión de los genes en cada una de las células de un tejido. Cuando estos resultados por célula se integraron con los resultados por tejido, se consiguió predecir los CRE activos en cada uno de sus tipos de células.

Entonces llegó el consorcio T2T:

Seguro que muchos pensáis que tenemos un genoma completo y bien anotado en el que no caben muchas mejoras. Pero dijimos que tenía huecos irresolubles por culpa de la tecnología disponible. De ahí surge el Consorcio T2T (*telomere-to-telomere* → de telómero a telómero) —formado por más de 100 investigadores de todo el mundo, incluida España, y organizado entre la bióloga Karen Miga, profesora ayudante en la Universidad de California en Santa Cruz, y el bioinformático Adam Phillippy, investigador titular en el NHGRI (National Human Genome Research Institute → Instituto Nacional de Investigación del Genoma Humano), ambos en EE UU— con el objetivo de conseguir una secuencia completa y contigua de cada cromosoma, de telómero a telómero (T2T). Para ello había que cambiar de estrategia y echaron mano de las tecnologías de secuenciación más vanguardistas de Oxford Nanopore y de Pacific Biosciences capaces de secuenciar de un tirón moléculas muy largas. Lo primero que hicieron fue dedicarse a resolver lo más complicado: las secuencias teloméricas, subteloméricas y centroméricas. Así presentaron en 2018 el centrómero del cromosoma Y[16], luego en 2020 la secuencia completa del cromosoma X[24], y al año siguiente otro cromosoma entero, el cromosoma 8[19]. Otras regiones repetitivas, sobre todo la que codifica los miles de copias de los rRNA (RNA ribosómicos), que parecían imposibles de resolver[22] han dejado por fin de ser un obstáculo.

El consorcio T2T prepublicó en mayo de 2021 un nuevo hito: la secuenciación completa, de telómero a telómero, de todos los cromosomas del genoma humano de la línea celular uniformemente homocigota CHM13hTERT derivada de la línea CHM13 (*complete hydatidiform mole* → mola hidatiforme completa, que es un tumor derivado de un embrión humano que rechazó el ADN de su padre y duplicó el de su madre)[37] del hospital Magee-Womens de Pittsburgh EE UU)[24]. El único defecto es que falta el cromosoma Y (si fuera una línea masculina no sería totalmente homocigota, ya que los cromosomas X e Y no son idénticos)[26]. La primera versión, denominada T2T-CHM13 v1.0, contenía unos 200 millones de pares de bases nuevos con 115 genes codificantes que no estaban en el GRCh38.p13. La evolución de las versiones del T2T-CHM13, así como los artículos que los autores van publicando sobre él, van apareciendo en su portal de GitHub.

Casi un año después, este marzo de 2022, apareció publicado definitivamente el trabajo en la revista *Science*[27] con ciertas mejoras sobre la prepublicación tanto en la secuencia como en la identificación de genes nuevos, que pasa a ser de 1956, de los cuales tan solo 99 codifican proteínas. En la tabla 2 del artículo hacen una comparación exhaustiva con el genoma GRCH38. Los interesados en navegar por este genoma completo lo pueden hacer en el UCSC Genome Browser (el servidor de genomas de la Universidad de California en Santa Fe antes mencionado). En pocas palabras: se ha aumentado ligeramente el número de genes totales, el de genes codificantes, el de genes exclusivos, y el número de transcritos y proteínas que se pueden sintetizar.

Las mejoras introducidas por el T2T-CHM13 v2.0 se ven claramente en la **figura 1** del artículo de Nurk y colaboradores (2022)[27].

Aunque en el artículo de *Science* se contempla también el análisis de las regiones repetitivas, dado que la línea de investigación de Karen Miga se centra en el DNA satélite, es en el de Hoyt y colaboradores[14] donde las describen exhaustivamente:

- 43 repeticiones y variantes nuevas;
- 19 estructuras repetitivas complejas (muchas de ellas con genes en su interior);
- que se transcriben muchas menos regiones satélites que repeticiones transposónicas;

- que las repeticiones transposónicas sirven de frontera para la expansión de la metilación de los islotes CpG y los centrómeros;
- que las regiones repetitivas son hipervariables entre los humanos (aunque ya lo suponíamos).

Todo esto se ve completado con un estudio sobre las modificaciones epigenéticas de este genoma [10] y el estudio completo de todos los centrómeros [2], entre los que también se observa una enorme e inesperada variabilidad.

Para tener una visión general de lo que sabemos sobre las regiones repetitivas del genoma humano gracias al T2T-CM13, consulta la **figura resumen** del artículo de Hoyt y colaboradores (2022) [14].

¿Estábamos engañados con el artículo de 2004 [15]? Pues no: la reconstrucción del genoma es como resolver un puzzle de más de 3 200 millones de piezas muy parecidas. Los primeros genomas que se secuenciaron tenían huecos donde no sabíamos qué piezas colocar debido a la limitación de la longitud de las lecturas de secuenciación y de los algoritmos de ensamblaje. Pero sí que sabíamos dónde estaban esos huecos, su tamaño y por qué no conseguíamos rellenarlos. Así que podemos afirmar que teníamos lo mejor que podíamos construir con la tecnología disponible: un modelo bastante completo de lo que debía ser el genoma humano. Los avances en la tecnología de secuenciación han permitido que el consocio T2T se plantee secuenciar todos los cromosomas de telómero a telómero, sin huecos, ¡y que lo consiga!. Esto no invalida para nada todo lo que se sabía, sino que nos lo confirma y completa. Algunos investigadores confían en que en los huecos ahora rellenos resida la información necesaria para ciertas enfermedades que hasta ahora no se han logrado mapear en el genoma. Solo existe un problema: el T2T-CHM13 es distinto al GRCH38, por lo que la comunidad científica tendrá que plantearse seriamente si cambia de referencia o si, puestos a cambiar, mejor abandonamos el genoma unipersonal y evolucionamos hacia una estrategia pangenómica.

Próxima parada: el pangenoma:

No todo el mundo es consciente de que el genoma humano de referencia que llevamos usando (el GRCh38) es un mosaico de más de 20 genomas distintos (aunque hay un individuo que aporta cerca del 70 % de la secuencia). Por eso siguen existiendo errores y configuraciones estructurales que no se han detectado en ningún otro genoma humano secuenciado (lo de los huecos ya se ha resuelto con el T2T-CHM13). Pero el principal escollo de cualquier genoma de referencia es que no representa la amplia diversidad genómica de la población, ni humana [23] ni de ninguna otra especie. Para afrontarlo, en 2019 se empezó a trabajar en un pangenoma[a] humano que sirva de referencia y que contemple la máxima diversidad genómica humana conocida. La propuesta parte de las líneas celulares que se usaron en el proyecto de 1000 Genomas representativos de 26 poblaciones diferentes a las que se irán incluyendo muestras de más poblaciones humanas [34].

En la **figura 1** de la revisión de Miga de 2021 [23] encontrarás cómo ha progresado la secuenciación del genoma humano desde que apareció el primer borrador hasta que se inició tanto el proyecto del pangenoma como el consorcio T2T.

Esto no es un sueño, sino una realidad, dado que, como comentamos al principio, ya contamos con la resecuenciación de muchos miles de personas [1,30,31,20], de exomas andaluces [5], e incluso de nuestros ancestros [12,21]. En 2020 ya apareció el primer pangenoma humano a partir de 338 individuos muy bien ensamblados [36] gracias a que también han empezado a aparecer las herramientas bioinformáticas más aptas para estos análisis [3]. La aparición del primer genoma completo (T2T-CHM13) ha impulsado y facilitado de tal manera el proceso que el pasado 20 de abril de 2022 acaba de publicarse la primera versión de un pangenoma T2T con las 47 primeras secuencias completas (de telómero a telómero) de los cromosomas [34]. Toda la información al respecto se irá haciendo pública en el portal Human Pangenome.

El proteoma 3D sin cristalización:

Los artículos publicados por los consorcios ENCODE, T2T y del pangenoma humano demuestran lo importante que es generar datos a gran escala en la biología y ofrecerlos a la comunidad científica para que se puedan explorar con otros ojos. Lo mismo llevan haciendo otros consorcios de investigación, como The Cancer Genome Atlas (TCGA), el Human Cell

[a] Genoma surgido de la unión de todas las secuencias genómicas (codificantes y no codificantes) de los individuos secuenciados de una especie para que esté representada toda la diversidad genética de dicha especie

Atlas[28] y el 4D Nucleome Project[4]. Todos ellos dan sentido a la existencia de portales de datos de acceso libre a la comunidad científica (desde GeneBank a FigShare o Zenodo), así como revistas como *Scientific Data* o *Data in Brief*. Gracias a ello se ha realizado un último avance sin precedentes: **la estructura tridimensional de 992 316 proteínas**, entre las que están el proteoma humano, el de otras 47 especies más, así como muchas proteínas de la base de datos UniProtKB.

Hasta ahora, cuando teníamos los marcos abiertos de lectura de los genes resultaba trivial obtener la secuencia de la proteína, pero era muy difícil predecir con suficiente exactitud su estructura tridimensional. Desde 1994, las competiciones de los algoritmos de predicción (CASP: Critical Assessment of protein Structure Prediction → valoración crítica de la predicción de estructuras proteicas) ofrecían un acierto que rondaba del 10 al 30 %. Pero la exitosa explotación de la inteligencia artificial con AlphaFold, creada por DeepMind, compañía hermana de Google en Londres, en la CASP13 de 2018 empezó a cambiar las tornas, puesto que alcanzó una valoración de 80 cuando sus competidores más cercanos (aunque utilizaran inteligencia artificial) no alcanzaban el 40[29]. Tanto cambió, que en la CASP14 de 2020 presentaron una versión mejorada basada en los algoritmos de traducción automática de textos —AlphaFold2[17]— con la que han logrado predecir correctamente el 98,5 % de las proteínas humanas (quedaron excluidas las proteínas maleables que, por definición, no tienen una estructura fija) con una puntuación media de 92 sobre 100. Lo más sorprendente de estas predicciones es que el 58 % de los residuos están en una estructura predicha fiable, de los que el 36 % se consideran tan fiables como si hubieran sido obtenidos por cristalografía de rayos X[32]. Por fin empezamos a tener las primeras predicciones estructurales fiables después de décadas de tenues avances.

Vamos, que tras el genoma humano ha venido su anotación, el genoma completo, el pangenoma y la predicciones tridimensionales fiables. ¿Qué será lo próximo?

For the times they are a-changin'.

– Bob Dylan, *Official Audio*

Referencias

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. Nature, 526(7571):68–74, Oct 2015. doi: 10.1038/nature15393.

[2] Nicolas Altemose, Glennis A Logsdon, Andrey V Bzikadze, Pragya Sidhwani, Sasha A Langley, Gina V Caldas, Savannah J Hoyt, Lev Uralsky, Fedor D Ryabov, Colin J Shew, Michael E G Sauria, Matthew Borchers, Ariel Gershman, Alla Mikheenko, Valery A Shepelev, Tatiana Dvorkina, Olga Kunyavskaya, Mitchell R Vollger, Arang Rhie, Ann M McCartney, Mobin Asri, Ryan Lorig-Roach, Kishwar Shafin, Julian K Lucas, Sergey Aganezov, Daniel Olson, Leonardo Gomes de Lima, Tamara Potapova, Gabrielle A Hartley, Marina Haukness, Peter Kerpedjiev, Fedor Gusev, Kristof Tigyi, Shelise Brooks, Alice Young, Sergey Nurk, Sergey Koren, Sofie R Salama, Benedict Paten, Evgeny I Rogaev, Aaron Streets, Gary H Karpen, Abby F Dernburg, Beth A Sullivan, Aaron F Straight, Travis J Wheeler, Jennifer L Gerton, Evan E Eichler, Adam M Phillippy, Winston Timp, Megan Y Dennis, Rachel J O'Neill, Justin M Zook, Michael C Schatz, Pavel A Pevzner, Mark Diekhans, Charles H Langley, Ivan A Alexandrov, and Karen H Miga. Complete genomic and epigenetic maps of human centromeres. Science, 376(6588):eabl4178, Apr 2022. doi: 10.1126/science.abl4178.

[3] Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief Bioinform, 19 (1):118–135, 01 2018. doi: 10.1093/bib/bbw089.

[4] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'Shea, Peter J Park, Bing Ren, Joan C Ritland Politz, Jay Shendure, Sheng Zhong, and 4D Nucleome Network. The 4d nucleome project. Nature, 549(7671):219–226, 09 2017. doi: 10.1038/nature23884.

[5] Joaquín Dopazo, Alicia Amadoz, Marta Bleda, Luz Garcia-Alonso, Alejandro Alemán, Francisco García-García, Juan A Rodriguez, Josephine T Daub, Gerard Muntané, Antonio Rueda, Alicia Vela-Boza, Francisco J López-Domingo, Javier P Florido, Pablo Arce, Macarena Ruiz-Ferrer, Cristina Méndez-Vidal, Todd E Arnold, Olivia Spleiss, Miguel Alvarez-Tejado, Arcadi Navarro, Shomi S Bhattacharya, Salud Borrego, Javier Santoyo-López, and Guillermo Antiñolo. 267 spanish exomes reveal population-specific differences in disease-related genetic variation. Mol Biol Evol, 33(5):1205–18, 05 2016. doi: 10.1093/molbev/msw005.

[6] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. Nature, 489(7414):57–74, Sep 2012. doi: 10.1038/nature11247.

[7] ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K G Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter,

Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ulaş Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W James Kent, Eric A Stone, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, Alexander E Urban, Peter Kraus, Sara Van Calcar, Nate Heintzman, Tae Hoon Kim, Kun Wang, Chunxu Qu, Gary Hon, Rosa Luna, Christopher K Glass, M Geoff Rosenfeld, Shelley Force Aldred, Sara J Cooper, Anason Halees, Jane M Lin, Hennady P Shulha, Xiaoling Zhang, Mousheng Xu, Jaafar N S Haidar, Yong Yu, Yijun Ruan, Vishwanath R Iyer, Roland D Green, Claes Wadelius, Peggy J Farnham, Bing Ren, Rachel A Harte, Angie S Hinrichs, Heather Trumbower, Hiram Clawson, Jennifer Hillman-Jackson, Ann S Zweig, Kayla Smith, Archana Thakkapallayil, Galt Barber, Robert M Kuhn, Donna Karolchik, Lluis Armengol, Christine P Bird, Paul I W de Bakker, Andrew D Kern, Nuria Lopez-Bigas, Joel D Martin, Barbara E Stranger, Abigail Woodroffe, Eugene Davydov, Antigone Dimas, Eduardo Eyras, Ingileif B Hallgrímsdóttir, Julian Huppert, Michael C Zody, Gonçalo R Abecasis, Xavier Estivill, Gerard G Bouffard, Xiaobin Guan, Nancy F Hansen, Jacquelyn R Idol, Valerie V B Maduro, Baishali Maskeri, Jennifer C McDowell, Morgan Park, Pamela J Thomas, Alice C Young, Robert W Blakesley, Donna M Muzny, Erica Sodergren, David A Wheeler, Kim C Worley, Huaiyang Jiang, George M Weinstock, Richard A Gibbs, Tina Graves, Robert Fulton, Elaine R Mardis, Richard K Wilson, Michele Clamp, James Cuff, Sante Gnerre, David B Jaffe, Jean L Chang, Kerstin Lindblad-Toh, Eric S Lander, Maxim Koriabine, Mikhail Nefedov, Kazutoyo Osoegawa, Yuko Yoshinaga, Baoli Zhu, and Pieter J de Jong. Identification and analysis of functional elements in 1 % of the human genome by the encode pilot project. Nature, 447(7146): 799–816, Jun 2007. doi: 10.1038/nature05874.

[8] ENCODE Project Consortium, Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L Van Nostrand, Peter Freese, David U Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A Williams, Ali Mortazavi, Cheryl A Keller, Xiao-Ou Zhang, Shaimae I Elhajjajy, Jack Huey, Diane E Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B Chhetri, Jialing Zhang, Alec Victorsen, Kevin P White, Axel Visel, Gene W Yeo, Christopher B Burge, Eric Lécuyer, David M Gilbert, Job Dekker, John Rinn, Eric M Mendenhall, Joseph R Ecker, Manolis Kellis, Robert J Klein, William S Noble, Anshul Kundaje, Roderic Guigó, Peggy J Farnham, J Michael Cherry, Richard M Myers, Bing Ren, Brenton R Graveley, Mark B Gerstein, Len A Pennacchio, Michael P Snyder, Bradley E Bernstein, Barbara Wold, Ross C Hardison, Thomas R Gingeras, John A Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of dna elements in the human and mouse genomes. Nature, 583(7818):699–710, 07 2020. doi: 10.1038/s41586-020-2493-4.

[9] Nelson J R Fagundes, Rafael Bisso-Machado, Pedro I C C Figueiredo, Maikel Varal, and André L S Zani. What we talk about when we talk about "junk dna". Genome Biol Evol, 14 (5), 05 2022. doi: 10.1093/gbe/evac055.

[10] Ariel Gershman, Michael E G Sauria, Xavi Guitart, Mitchell R Vollger, Paul W Hook, Savannah J Hoyt, Miten Jain, Alaina Shumate, Roham Razaghi, Sergey Koren, Nicolas Altemose, Gina V Caldas, Glennis A Logsdon, Arang Rhie, Evan E Eichler, Michael C Schatz, Rachel J O'Neill, Adam M Phillippy, Karen H Miga, and Winston Timp. Epigenetic patterns in a complete human genome. Science, 376(6588):eabj5089, Apr 2022. doi: 10.1126/science.abj5089.

[11] Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B R Azevedo, Rebecca A Zufall, and Eran Elhaik. On the immortality of television sets: "functionïn the human genome according to the evolution-free gospel of encode. Genome Biol Evol, 5(3): 578–90, 2013. doi: 10.1093/gbe/evt028.

[12] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F Hansen, Eric Y Durand, Anna-Sapfo Malaspinas, Jeffrey D Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A Burbano, Jeffrey M Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B Doronichev, Liubov V Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W Schmitz, Philip L F Johnson, Evan E Eichler, Daniel Falush, Ewan Birney, James C Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo. A draft sequence of the neandertal genome. Science, 328(5979):710–722, May 2010. doi: 10.1126/science.1188021.

[13] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James G R Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E Antonarakis, and Roderic Guigo. Gencode: producing a reference annotation for encode. Genome Biol, 7 Suppl 1:S4.1–9, 2006. doi: 10.1186/gb-2006-7-s1-s4.

[14] Savannah J Hoyt, Jessica M Storer, Gabrielle A Hartley, Patrick G S Grady, Ariel Gershman, Leonardo G de Lima, Charles Limouse, Reza Halabian, Luke Wojenski, Matias Rodriguez, Nicolas Altemose, Arang Rhie, Leighton J Core, Jennifer L Gerton, Wojciech Makalowski, Daniel Olson, Jeb Rosen, Arian F A Smit, Aaron F Straight, Mitchell R Vollger, Travis J Wheeler, Michael C Schatz, Evan E Eichler, Adam M Phillippy, Winston Timp, Karen H Miga, and Rachel J O'Neill. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science, 376(6588):eabk3112, Apr 2022. doi: 10.1126/science.abk3112.

[15] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 431(7011):931–45, Oct 2004. doi: 10.1038/nature03001.

[16] Miten Jain, Hugh E Olsen, Daniel J Turner, David Stoddart, Kira V Bulazel, Benedict Paten, David Haussler, Huntington F Willard, Mark Akeson, and Karen H Miga. Linear assembly of a human centromere on the y chromosome. Nat Biotechnol, 36(4):321–323, 04 2018. doi: 10.1038/nbt.4109.

[17] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. Nature, 596(7873): 583–589, 08 2021. doi: 10.1038/s41586-021-03819-2.

[18] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 409 (6822):860–921, Feb 2001. doi: 10.1038/35057062.

[19] Glennis A Logsdon, Mitchell R Vollger, PingHsun Hsieh, Yafei Mao, Mikhail A Liskovykh, Sergey Koren, Sergey Nurk, Ludovica Mercuri, Philip C Dishuck, Arang Rhie, Leonardo G de Lima, Tatiana Dvorkina, David Porubsky, William T Harvey, Alla Mikheenko, Andrey V Bzikadze, Milinn Kremitzki, Tina A Graves-Lindsay, Chirag Jain, Kendra Hoekzema, Shwetha C Murali, Katherine M Munson, Carl Baker, Melanie Sorensen, Alexandra M Lewis, Urvashi Surti, Jennifer L Gerton, Vladimir Larionov, Mario Ventura, Karen H Miga, Adam M Phillippy, and Evan E Eichler. The structure, function and evolution of a complete human chromosome 8. Nature, 593(7857):101–107, 05 2021. doi: 10.1038/s41586-021-03420-7.

[20] Lasse Maretty, Jacob Malte Jensen, Bent Petersen, Jonas Andreas Sibbesen, Siyang Liu, Palle Villesen, Laurits Skov, Kirstine Belling, Christian Theil Have, Jose M G Izarzugaza, Marie Grosjean, Jette Bork-Jensen, Jakob Grove, Thomas D Als, Shujia Huang, Yuqi Chang, Ruiqi Xu, Weijian Ye, Junhua Rao, Xiaosen Guo, Jihua Sun, Hongzhi Cao, Chen Ye, Johan van Beusekom, Thomas Espeseth, Esben Flindt, Rune M Friborg, Anders E Halager, Stephanie Le Hellard, Christina M Hultman, Francesco Lescai, Shengting Li, Ole Lund, Peter Løngren, Thomas Mailund, Maria Luisa Matey-Hernandez, Ole Mors, Christian N S Pedersen, Thomas Sicheritz-Pontén, Patrick Sullivan, Ali Syed, David Westergaard, Rachita Yadav, Ning Li, Xun Xu, Torben Hansen, Anders Krogh, Lars Bolund, Thorkild I A Sørensen, Oluf Pedersen, Ramneek Gupta, Simon Rasmussen, Søren Besenbacher, Anders D Børglum, Jun Wang, Hans Eiberg, Karsten Kristiansen, Søren Brunak, and Mikkel Heide Schierup. Sequencing and de novo assembly of 150 genomes from denmark as a population reference. Nature, 548(7665): 87–91, 08 2017. doi: 10.1038/nature23264.

[21] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E Green, Katarzyna Bryc, Adrian W Briggs, Udo Stenzel, Jesse Dabney, Jay Shendure, Jacob Kitzman, Michael F Hammer, Michael V Shunkov, Anatoli P Derevianko, Nick Patterson, Aida M Andrés, Evan E Eichler, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. A high-coverage genome sequence from an archaic denisovan individual. Science, 338(6104):222–6, Oct 2012. doi: 10.1126/science.1224344.

[22] Karen H Miga. Breaking through the unknowns of the human reference genome. Nature, 590(7845):217–218, 02 2021. doi: 10.1038/d41586-021-00293-8.

[23] Karen H Miga and Ting Wang. The need for a human pangenome reference sequence. Annu Rev Genomics Hum Genet, 22: 81–102, 08 2021. doi: 10.1146/annurev-genom-120120-081921.

[24] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Amy B Wilfert, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy. Telomere-to-telomere assembly of a complete human x chromosome. Nature, 585(7823):79–84, 09 2020. doi: 10.1038/s41586-020-2547-7.

[25] M R Montminy and L M Bilezikjian. Binding of a nuclear protein to the cyclic-amp response element of the somatostatin gene. Nature, 328(6126):175–8, 1987. doi: 10.1038/328175a0.

[26] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina

Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. bioRxiv, 2021. doi: 10.1101/2021.05.26.445798. URL https://www.biorxiv.org/content/early/2021/05/27/2021.05.26.445798.

[27] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J Hoyt, Mark Diekhans, Glennis A Logsdon, Michael Alonge, Stylianos E Antonarakis, Matthew Borchers, Gerard G Bouffard, Shelise Y Brooks, Gina V Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G de Lima, Philip C Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T Fiddes, Giulio Formenti, Robert S Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G S Grady, Tina A Graves-Lindsay, Ira M Hall, Nancy F Hansen, Gabrielle A Hartley, Marina Haukness, Kerstin Howe, Michael W Hunkapiller, Chirag Jain, Miten Jain, Erich D Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V Maduro, Tobias Marschall, Ann M McCartney, Jennifer McDaniel, Danny E Miller, James C Mullikin, Eugene W Myers, Nathan D Olson, Benedict Paten, Paul Peluso, Pavel A Pevzner, David Porubsky, Tamara Potapova, Evgeny I Rogaev, Jeffrey A Rosenfeld, Steven L Salzberg, Valerie A Schneider, Fritz J Sedlazeck, Kishwar Shafin, Colin J Shew, Alaina Shumate, Ying Sims, Arian F A Smit, Daniela C Soto, Ivan Sović, Jessica M Storer, Aaron Streets, Beth A Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P Walenz, Aaron Wenger, Jonathan M D Wood, Chunlin Xiao, Stephanie M Yan, Alice C Young, Samantha Zarate, Urvashi Surti, Rajiv C McCoy, Megan Y Dennis, Ivan A Alexandrov, Jennifer L Gerton, Rachel J O'Neill, Winston Timp, Justin M Zook, Michael C Schatz, Evan E Eichler, Karen H Miga, and Adam M Phillippy. The complete sequence of a human genome. Science, 376(6588): 44–53, Apr 2022. doi: 10.1126/science.abj6987.

[28] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The human cell atlas. Elife, 6, 12 2017. doi: 10.7554/eLife.27041.

[29] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W R Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792):706–710, 01 2020. doi: 10.1038/s41586-019-1923-7.

[30] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K Konkel, Ankit Malhotra, Adrian M Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J P Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y K Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A Gibbs, Gabor Marth, Christopher E Mason, Androniki Menelaou, Donna M Muzny, Bradley J Nelson, Amina Noor, Nicholas F Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A Shabalin, Andreas Untergasser, Jerilyn A Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A Batzer, Steven A McCarroll, 1000 Genomes Project Consortium, Ryan E Mills, Mark B Gerstein, Ali Bashir, Oliver Stegle, Scott E Devine, Charles Lee, Evan E Eichler, and Jan O Korbel. An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571):75–81, Oct 2015. doi: 10.1038/nature15394.

[31] Amalio Telenti, Levi C T Pierce, William H Biggs, Julia di Iulio, Emily H M Wong, Martin M Fabani, Ewen F Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A Perkins, Franz J Och, Yaron Turpaz, and J Craig Venter. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A, 113(42):11901–11906, 10 2016. doi: 10.1073/pnas.1613365113.

[32] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. Nature, 596(7873):590–596, 08 2021. doi: 10.1038/s41586-021-03828-1.

[33] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport,

R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. Science, 291(5507):1304–51, Feb 2001. doi: 10.1126/science.1058040.

[34] Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A. Lawson, Julian K. Lucas, Adam M. Phillippy, Alice B. Popejoy, Mobin Asri, Caryn Carson, Mark J. P. Chaisson, Xian Chang, Robert Cook-Deegan, Adam L. Felsenfeld, Robert S. Fulton, Erik P. Garrison, Nanibaa'A. Garrison, Tina A. Graves-Lindsay, Hanlee Ji, Eimear E. Kenny, Barbara A. Koenig, Daofeng Li, Tobias Marschall, Joshua F. McMichael, Adam M. Novak, Deepak Purushotham, Valerie A. Schneider, Baergen I. Schultz, Michael W. Smith, Heidi J. Sofia, Tsachy Weissman, Paul Flicek, Heng Li, Karen H. Miga, Benedict Paten, Erich D. Jarvis, Ira M. Hall, Evan E. Eichler, David Haussler, and the Human Pangenome Reference Consortium. The human pangenome project: a global resource to map genomic diversity. Nature, 604 (7906):437–446, 2022. doi: 10.1038/s41586-022-04601-8. URL https://doi.org/10.1038/s41586-022-04601-8.

[35] J D WATSON and F H CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171 (4356):737–8, Apr 1953. doi: 10.1038/171737a0.

[36] Karen H Y Wong, Walfred Ma, Chun-Yu Wei, Erh-Chan Yeh, Wan-Jia Lin, Elin H F Wang, Jen-Ping Su, Feng-Jen Hsieh, Hsiao-Jung Kao, Hsiao-Huei Chen, Stephen K Chow, Eleanor Young, Catherine Chu, Annie Poon, Chi-Fan Yang, Dar-Shong Lin, Yu-Feng Hu, Jer-Yuarn Wu, Ni-Chung Lee, Wuh-Liang Hwu, Dario Boffelli, David Martin, Ming Xiao, and Pui-Yan Kwok. Towards a reference genome that captures global genetic diversity. Nat Commun, 11(1):5482, 10 2020. doi: 10.1038/s41467-020-19311-w.

[37] Eiko Yamamoto, Kaoru Niimi, Tohru Kiyono, Toshimichi Yamamoto, Kimihiro Nishino, Kenichi Nakamura, Tomomi Kotani, Hiroaki Kajiyama, Kiyosumi Shibata, and Fumitaka Kikkawa. Establishment and characterization of cell lines derived from complete hydatidiform mole. Int J Mol Med, 40(3):614–622, Sep 2017. doi: 10.3892/ijmm.2017.3067.

[38] Feng Yue, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, Zhihai Ma, Carrie Davis, Benjamin D Pope, Yin Shen, Dmitri D Pervouchine, Sarah Djebali, Robert E Thurman, Rajinder Kaul, Eric Rynes, Anthony Kirilusha, Georgi K Marinov, Brian A Williams, Diane Trout, Henry Amrhein, Katherine Fisher-Aylor, Igor Antoshechkin, Gilberto DeSalvo, Lei-Hoon See, Meagan Fastuca, Jorg Drenkow, Chris Zaleski, Alex Dobin, Pablo Prieto, Julien Lagarde, Giovanni Bussotti, Andrea Tanzer, Olgert Denas, Kanwei Li, M A Bender, Miaohua Zhang, Rachel Byron, Mark T Groudine, David McCleary, Long Pham, Zhen Ye, Samantha Kuan, Lee Edsall, Yi-Chieh Wu, Matthew D Rasmussen, Mukul S Bansal, Manolis Kellis, Cheryl A Keller, Christapher S Morrissey, Tejaswini Mishra, Deepti Jain, Nergiz Dogan, Robert S Harris, Philip Cayting, Trupti Kawli, Alan P Boyle, Ghia Euskirchen, Anshul Kundaje, Shin Lin, Yiing Lin, Camden Jansen, Venkat S Malladi, Melissa S Cline, Drew T Erickson, Vanessa M Kirkup, Katrina Learned, Cricket A Sloan, Kate R Rosenbloom, Beatriz Lacerda de Sousa, Kathryn Beal, Miguel Pignatelli, Paul Flicek, Jin Lian, Tamer Kahveci, Dongwon Lee, W James Kent, Miguel Ramalho Santos, Javier Herrero, Cedric Notredame, Audra Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Theresa Canfield, Peter J Sabo, Matthew S Wilken, Thomas A Reh, Erika Giste, Anthony Shafer, Tanya Kutyavin, Eric Haugen, Douglas Dunn, Alex P Reynolds, Shane Neph, Richard Humbert, R Scott Hansen, Marella De Bruijn, Licia Selleri, Alexander Rudensky, Steven Josefowicz, Robert Samstein, Evan E Eichler, Stuart H Orkin, Dana Levasseur, Thalia Papayannopoulou, Kai-Hsin Chang, Arthur Skoultchi, Srikanta Gosh, Christine Disteche, Piper Treuting, Yanli Wang, Mitchell J Weiss, Gerd A Blobel, Xiaoyi Cao, Sheng Zhong, Ting Wang, Peter J Good, Rebecca F Lowdon, Leslie B Adams, Xiao-Qiao Zhou, Michael J Pazin, Elise A Feingold, Barbara Wold, James Taylor, Ali Mortazavi, Sherman M Weissman, John A Stamatoyannopoulos, Michael P Snyder, Roderic Guigo, Thomas R Gingeras, David M Gilbert, Ross C Hardison, Michael A Beer, Bing Ren, and Mouse ENCODE Consortium. A comparative encyclopedia of dna elements in the mouse genome. Nature, 515(7527):355–64, Nov 2014. doi: 10.1038/nature13992.

Borfitz, D (2022) Scientists Finally Finish The Quest For A Gapless Human Genome. BioIT World. [consulta 20-IV-22]