EDITORIAL

MÁS ALLÁ DE LA SECUENCIACIÓN

El modelo de la estructura tridimensional del DNA propuesto por Watson y Crick en 1953, permitió concebir por un lado cómo es que se transmite la información de generación en generación y por el otro cómo es que esta información se expresa en cada individuo. Ambos aspectos dependen directamente de la secuencia de nucleótidos en los genomas. Por esta razón, era claro que el poder determinar la secuencia de los ácidos nucleicos era un objetivo lógico. Sin embargo la tecnología de la época y durante más de dos décadas tuvo muchas limitantes. De hecho, dado que no era factible secuenciar grandes cantidades de DNA, mucha de la información inicial se obtuvo de "segunda mano" usando los métodos de Fred Sanger para determinar la secuencia de residuos de aminoácidos de las proteínas. A pesar de esto, a cuentagotas se fueron obteniendo secuencias de DNA que, aunque cortas, eran notablemente relevantes. Por ejemplo, las secuencias de algunos promotores bacterianos y del operador Lac, permitieron comenzar a intuir cómo es que ocurren las interacciones DNA-proteína, fundamentales en la regulación de la expresión génica. Sin embargo, la búsqueda del grial, obtener la secuencia de nucleótidos de genes y genomas continuaba.

Finalmente con la llegada de las metodologías de secuenciación desarrolladas por Fred Sanger y su grupo, la meta comenzó a vislumbrarse: Ya era posible secuenciar segmentos relativamente largos de DNA de manera confiable. Aunque estos métodos tenían algunas limitaciones, se puede decir que con el método de Sanger llegamos no sólo a la etapa del gen sino a la etapa genómica, ya que de hecho a finales de los años setenta del siglo pasado se secuenció el primer genoma completo (el de ϕX174 / 5,375 pares de bases). Aunque el avance fue lento, finalmente en 1995 el grupo de Craig Venter publicó el primer genoma de una bacteria de vida libre, *Hemophilus influenzae*, de 1.8 Mb (mega pares de bases = un millón de pares de bases). A este primer genoma se unieron muchos otros y en 2001 se publicaron los primeros borradores del genoma humano. Los esfuerzos por alcanzar este logro habían comenzado muchos años antes y fueron largos y tortuosos. Por ejemplo, el grupo Celera utilizó 200 secuenciadores "automáticos", que trabajaron las 24 horas durante tres años y completaron la secuencia de cada uno de los cromosomas humanos. Esto fue más del doble del tiempo de lo que le tomó al consorcio público concluir su trabajo.

Los avances continuaron y llegaron los protocolos llamados de segunda generación (NGS, Next Generation Sequencing), que permitieron obtener millones de nucleótidos secuenciados en cuestión de horas y además los costos se disminuyeron de manera muy importante. Por supuesto, estos métodos aumentaron exponencialmente el número de secuencias reportadas, por lo que también fue necesario que la bioinformática avanzara en paralelo. Y surgió la metagenómica, que pretende identificar organismos no cultivables en el laboratorio que son la mayor parte de la población presente en cualquier ecosistema. A pesar de esto, en realidad la cantidad de genomas secuenciados sigue siendo una pequeña fracción del número de especies, sobre todo unicelulares, que se estima que existen. Para la segunda década de este siglo llegó la secuenciación de tercera generación, que permite secuenciar de manera corrida hasta 100, 000 pares de bases. Dicho de otro modo, la secuenciación ya no es una limitante.

¿Y ahora qué sigue?

Actualmente hemos progresado de tan solo leer a comenzar a escribir el DNA. La síntesis de DNA no es nueva, la síntesis química de oligonucleótidos lleva bastante tiempo en escena. Sin embargo, en la actualidad ya somos capaces de sintetizar DNA en escala de genomas. El mismo Venter, cuyo desarrollo de la secuenciación por fragmentos (shotgun) y el consecuente ensamblado genómico de fragmentos aceleró vertiginosamente la posibilidad de secuenciar genomas, ahora habla de que el visualiza a los genomas como un software o más bien como el sistema operativo de una célula y se ha lanzado a las primeras incursiones del cómo hacer un sistema operativo biológico mediante la síntesis de un genoma mínimo. Si bien tomó como base a los ya reducidos genomas bacterianos del género *Mycoplasma*, fue capaz de tener una estrategia sistematizada para el diseño, síntesis química, ensamblado y trasplante del genoma.

La idea del genoma mínimo es atractiva desde múltiples aristas. Desde la genómica comparativa, la idea de un genoma mínimo lleva a preguntas

fundamentales, por ejemplo: ¿Qué genes son los conservados en un grupo biológico? ¿Qué mantiene la coherencia de un género, de una especie, a nivel molecular? Para este tipo de preguntas las respuestas tentativas provienen de múltiples disciplinas previas a las ómicas, tales como la genética clásica, la bioquímica, la biología molecular y en eras recientes, la genómica funcional. Éstas nos han proporcionado las herramientas para comenzar a entender la función de los genes.

Resulta curioso que hemos superado una limitante clásica de nuestro proceso de comprensión y ya es rutinario secuenciar genomas bajo la menor provocación y que podemos hacerlo en casi cualquier laboratorio. Sin embargo, la velocidad a la que se hacen los experimentos gen por gen no avanza a la velocidad de la secuenciación. Desde que Jacques Monod, François Jacob y Andre Lwoff, galardonados en 1965 con el premio Nobel por su trabajo de regulación genética, descubrieron los operones en la bacteria *Escherichia coli*, describieron principios aplicables a toda la vida; se esperaría que en 2017 ya fuéramos capaces de integrar un modelo donde a partir del genoma de *E. coli* pudiéramos predecir su fisiología, metabolismo, función y regulación de sus genes. Intentos hay varios, pero estamos limitados al universo de los genes cuya función ya está descrita y aún nos falta trayecto por recorrer.

Se calcula que en el genoma de E. coli se codifican alrededor de 4,140 proteínas. De estas, aún quedan 147 proteínas predichas que no están bien caracterizadas. Un segundo modelo bacteriano bien estudiado es, sin duda, *Bacillus subtilis*. En este genoma se ubican alrededor de 4,260 genes que codifican para proteínas, de las cuales 847 son hipotéticas. Hipotéticas es el eufemismo para disimular el hecho de que no tenemos mucha idea de para qué sirven. Otras 231 son hipotéticas sin ninguna secuencia homóloga en las bases de datos. Así, no tenemos idea de un 19.88% de las funciones de los genes de uno de los organismos mejor estudiados en la historia. El panorama no mejora cuando nos movemos fuera de los organismos modelo. Con las herramientas de genética y genómica disponibles, no es difícil encontrar un 30% de genes sin parecido a nada en las bases de datos de genes y proteínas por cada genoma bacteriano nuevo secuenciado. Inclusive en el genoma "sintético" del *Mycoplasma de Venter*, con tan solo 473 genes, se incluyen 79 genes que codifican proteínas (16.70% del genoma) de los cuales no se tienen evidencia de sus funciones moleculares fuera de que son esenciales para la célula y mejor tenerlos incluidos en el genoma sintético.

La secuenciación va dejando de ser un problema y ya hay múltiples tecnologías maduras para secuenciar miles de genomas de cualquier tamaño de forma casi rutinaria y asequible. El futuro después de la secuenciación de próxima generación y de tercera generación será poner a punto las estrategias de genómica funcional automatizadas y de alto desempeño (high throughput); sintetizar de forma automática los genes que deseamos estudiar; lograr generar mutantes de pérdida y ganancia de función mediante sistemas de edición genómica (i.e. CRISPR-Cas9); evaluar fenotipos con sistemas de monitoreo automático; analizar transcriptomas en múltiples condiciones; mejorar sistemas de expresión de proteínas, interacciones DNA-proteína, proteína-proteína, etc. Se han hecho avances significativos en temas de automatización, pero todavía seguimos dependiendo de técnicas y habilidades clásicas de genética, biología molecular y bioquímica para entender las funciones moleculares codificadas en los genomas. Quizá de la mano de la automatización de procesos de tareas del laboratorio se podría pensar en la siguiente generación y quizá hablar del next generation phenotype. La tarea pendiente es enorme, pero nos permitirá aventurarnos a tratar de recorrer el camino desde el genoma hasta el fenotipo y seguir intentando modelar de forma global el funcionamiento de células, poblaciones, comunidades y ecosistemas. Desde un punto de vista muy general, un genoma es un conjunto de genes. La genómica evolutiva permite relacionar genomas donde el repertorio es parecido pero no idéntico (además de que el orden puede cambiar entre genomas a veces muy relacionados). Las pérdidas y ganancias de regiones genómicas son sumamente comunes y seguramente juegan un papel importante en la diferenciación filogenética.

Vamos, hay mucha vida y trabajo por delante del post next generation sequencing. Es una época emocionante para trabajar donde la genética, la biología molecular y la bioquímica tradicionales siguen siendo el fundamento y donde no debemos de perder a las preguntas biológicas fundamentales como los actores principales, auxiliadas por todo el desarrollo tecnológico asociado.

Luis David Alcaraz Peraza
Claudia Segal Kischinevzky
Viviana Escobar Sánchez
Luisa Alba Lois
Víctor Valdés López
Laboratorio de Biología Molecular y Genómica.
Departamento de Biología Celular.
Facultad de Ciencias.
Universidad Nacional Autónoma de México.