Genomas y rompecabezas: una visión sobre el ensamblaje de genomas

Aureliano Bombarely Gómez
Investigador Asociado, Department of Plant Biology, Cornell University, Ithaca, Estados Unidos
aubombarely@gmail.com

Resumen

El desarrollo nuevas tecnologías de secuenciación ha revolucionado el análisis de genomas. Los grandes proyectos de secuenciación se han ido sustituyendo por aproximaciones más modestas, tanto en personal como en costes. Actualmente es posible secuenciar, ensamblar y analizar un genoma vegetal de tamaño medio con una cantidad limitada de recursos, si bien todavía estamos lejos de poder ensamblar cualquier genoma. Genomas de gran tamaño, con un gran contenido en repeticiones, poliploides, o genomas con una elevada heterocigosidad pueden ser un problema de difícil solución.

Desde el primer virus hasta la primera planta poliploide

La secuenciación de genomas nace en 1976 con el genoma del Bacteriófago MS2 [1]. No más grande que muchos tránscritos eucariotas (3,568 nucleótidos), fue secuenciado mediante nucleótidos marcados con 32P, digeridos con la ribonucleasa T1 y separados en un gel de poliacrilamida, en una época donde todavía no se había desarrollado la reacción en cadena de la polimerasa [2] y donde los ordenadores portátiles pesaban 25 kilos y tenían 64 Kb de memoria RAM (como el modelo IBM 5100).

El desarrollo de los secuenciadores automáticos de ADN mediante el método de Sanger por parte de Applied Biosystems en 1987 supuso un salto cuantitativo. El primer genoma bacteriano, *Haemophilus influenzae* con un tamaño de 1.83 Mb, fue secuenciado en 1995 por Craig Venter mediante la metodología de Whole Genome Shotgun (WGS) usando 14 secuenciadores AB373 durante 3 meses. Las lecturas se ensamblaron con TIGR ASSEMBLER [3], un programa desarrollado por The Institute for Genomic Research (TIGR) que se basaba en el solapamiento de secuencias alineadas mediante una versión modificada del algoritmo Smith-Waterman. Se tardó 30 horas en un equipo con un solo procesador y 512 Mb de memoria RAM [4]. Le seguirían la secuenciación del primer genoma eucariota (con un tamaño de 12.1 Mb), *Saccharomyces cerevisiae* [5]; el del primer animal multicelular (con un tamaño de 100 Mb), *Caenorhabditis elegans* [6]; el de la primera planta (con un tamaño de 157 Mb), *Arabidopsis thaliana* [7]; el borrador del genoma del ser humano (con un tamaño de 3.2 Gb) [8,9]; el genoma del primer vertebrado tras el ser humano (con un tamaño de 390 Mb), F*ugu rubripes* [10]y el del primer mamífero tras el ser humano (con un tamaño de 2.5 Gb), *Mus musculus* [11]. Para todos estos proyectos se utilizó la misma tecnología de secuenciación, y generalmente eran el fruto del esfuerzo de muchos grupos de investigación durante varios años. Exceptuando en los proyectos donde se utilizó la metodología de *BAC-by-BAC* como en *Arabidopsis*, los programas utilizados para el ensamblaje de (*Arachne* [12], *WGA assembler* [13]) no diferían mucho en su planteamiento del desarrollado por TIGR años atrás basado en el solapamiento de secuencias, si bien eran programas mejor estructurados (con diferentes fases durante el ensamblaje) y que aprovechaban ciertos recursos computacionales como el uso de grupo de servidores (*server farm*) para usar cientos de procesadores. Por ejemplo en el proyecto del genoma humano dirigido por Craig Venter se utilizó un sistema con 40 servidores (con 4 procesadores y 4 Gb de memoria RAM cada uno) trabajando en paralelo durante 5 días.

El ritmo de secuenciación de genomas eucariotas durante los primeros años del nuevo milenio ha sido de aproximadamente dos o tres genomas por año en el mejor de los casos pero actualmente dicho ritmo se ha disparado. En el año 2012, solo en el área de plantas se han publicado más de una docena de genomas[14-27]. Este cambio se debe a dos factores: El desarrollo de las nuevas tecnologías de secuenciación (*Next Generation Sequencing*, NGS) que han permitido el abaratamiento del coste de la secuenciación y el uso de nuevos programas de ensamblaje más rápidos y eficientes. En el año 2005 se publicó la primera de las nuevas metodologías de secuenciación basada en una reacción de pirólisis sobre una matriz sólida con millones de puntos, cada uno representado un secuencia [28]. *454* (que más tarde sería comprada por *Roche*) sacó al mercado un nuevo sistema de secuenciación capaz de producir millones de lecturas por proceso. Al pirosecuenciamiento de 454 le han seguido la secuenciación basada en terminadores de química reversible de *Illumina* (2005) [29], la secuenciación por ligamiento de *SOLiD* (2007) [30], la secuenciación por iones semiconductores de *Ion Torrent Biosystems* (2011) [31] y la secuenciación de una sola molécula en tiempo real de *Pacific*

Biosciences (2012) [32] entre otros, aunque sin duda alguna, la más popular es *Illumina*. Actualmente puede producir 600 Gb de lecturas por proceso, 200 veces el tamaño del genoma humano en tan solo 11 días haciendo posible la secuenciación de genomas de tamaño similares en cortos periodos de tiempo.

Diseñando un proyecto de secuenciación

Antes de embarcarse en un proyecto de secuenciación de un genoma es crítico tener un diseño experimental adecuado. La aproximación puede ser totalmente distinta incluso entre individuos de la misma especie dependiendo de algunas diferencias genéticas como el grado de heterocigosidad del individuo en cuestión. Cada uno de ellos tienen distintas características que pueden hacer totalmente inadecuadas algunas metodologías. Por otro lado el presupuesto y tiempo disponible pueden limitar el uso de algunas metodologías como el uso de cromosomas artificiales bacterianos (BAC), que si bien son más seguras y exactas, pueden disparar el coste de procesamiento y secuenciación varios órdenes de magnitud.

1- Conoce a tu enemigo: Que secuenciar.

Unos de los primeros pasos para enfrentarse a un proyecto de secuenciación es el de conocer algunas características del genoma a secuenciar tales como:

- Tamaño de genoma. Existen algunas bases de datos que pueden dar una información orientativa del tamaño del genoma a través de estudios citogenéticos. Un buen ejemplo de ello es la base de datos de *Plant DNA C-values* [33] donde pueden encontrarse medidas de tamaños para más de 1,200 genomas vegetales.
- Poliploidía. Al igual que en el caso anterior, estudios citogenéticos previos pueden facilitar este tipo de información. El ensamblaje de una especie poliploide tiene el gran problema de que una gran parte de las regiones homoeólogas (provenientes de uno o varios progenitores, en el caso de auto- y alo-poliploides respectivamente) van a colapsar en una misma secuencia consenso [23]. Existen distintas opciones para minimizar este problema como el uso de la información de pares para crear una fase para las regiones homoeólogas aunque por el momento la aproximación más usada es la de la secuenciación de uno de los progenitores diploides para su uso como referencia [27].
- Heterocigozidad. De forma parecida a la autopoliploidía, una baja heterocigosidad de la muestra puede conducir al colapso de dos alelos, aunque en este caso es un efecto deseable. Por otro lado una elevada heterocigosidad puede traducirse en la creación de una secuencia consenso para cada alelo en las regiones con más variabilidad, lo que se traduce en burbujas de ensamblaje y que a menudo suelen ser difícil de incluir en la reconstrucción del genoma. Lo que es más, la cobertura efectiva (cuantas veces está representado el genoma en el set de secuencias) disminuye dificultando el ensamblaje.
- Eventos de duplicación. Una gran mayoría de organismos eucariotas presentan algún evento de duplicación genómica *(Whole Genome Duplication*, WGD) en su historia. Este fenómeno es especialmente representativo en plantas donde los eventos de duplicaciones son relativamente comunes. Por ejemplo existen dos eventos de duplicación (datados alrededor de 319 y 192 Ma respectivamente) comunes para todas las angioespermas [34]. En el caso de *Arabidopsis* existen otros 3 eventos de duplicación que han dado forma al genoma que se conoce hoy día (el primero producido tras la separación de monocotiledóneas y dicotiledóneas, y el segundo y tercero durante de la formación de las brasicáceas [35,36]) aunque no han influenciado de forma notable el ensamblaje de su genoma. Distinto es el caso de la soja (*Glycine max*) con dos eventos de duplicación, el más reciente con una antigüedad de 13 Ma y con un gran elevado contenido de repeticiones [37].
- Repeticiones. Genomas con un elevado contenido en repeticiones pueden ser difíciles de ensamblar. A fin de facilitar el proceso de ensamblaje, estos programas filtran las lecturas extremadamente representadas pudiendo producir genomas muy fragmentados en caso de que las repeticiones se encuentren uniformemente distribuidas a lo largo de todo el genoma. Es conveniente realizar estudios preliminares sobre el contenido en repeticiones usando una secuenciación de baja cobertura [38]y/o análisis citogenéticos mediante FISH (*Fluorescent In-Situ Hybridization*) [39]. Si los resultados preliminares revelan un alto contenido en elementos repetitivos uniformemente distribuidos en la eucromatina puede que el uso de la metodología de WGS sin el apoyo de secuenciación de BACs sea inviable.

De esta manera es importante que en la medida de lo posible se simplifique el proyecto de secuenciación seleccionando variedades con una baja heterocigosidad, dobles haploides si la especie es de forma natural un autopoliploide o secuenciando los progenitores diploides si la especie es un alopoliploide y generando estudios preliminares sobre el contenido en repeticiones antes de comenzar a secuenciar.

2- Estima la cantidad de datos necesaria: Cómo y cuánto secuenciar.

El siguiente paso es decidir que tecnología usar y cuanto secuenciar. Para secuenciaciones mediante WGS (Whole Genome Shotgun) es esencial, independientemente de la tecnología utilizada, el uso de pares (Pair Ends y Mate Pairs)[40,41]. Estos sirven para relacionar las secuencias consenso (contigs) entre sí y crear así estructuras de mayor tamaño (scaffolds) con una estimación aproximada de la distancia entre contigs. Es importante el uso de combinaciones de librerías de pares con insertos de varios tamaños, por ejemplo una o dos librerías con insertos de tamaños entre 170 y 800 pb y dos o tres librerías con insertos entre 2 y 20 Kb. La longitud de las lecturas utilizadas dependerá de la tecnología de secuenciación seleccionada, pero es aconsejable secuenciar con la máxima longitud disponible siempre y cuando no tenga un efecto drástico en la calidad de las secuencias (actualmente 500 pb para 454 y 150 pb para *Illumina HiSeq* y 250 pb para *Illumina MiSeq*). Finalmente queda decidir cuanto ha de secuenciarse, y de nuevo dependerá de la tecnología de secuenciación usada. Para 454 es recomendable usar al menos 10X (es decir 10 veces el tamaño del genoma a secuenciar), si bien se obtienen buenos resultados a partir de 30X. Para *Illumina* la comunidad se ha puesto de acuerdo para recomendar coberturas en torno a 100X. De esta manera significa que, por ejemplo, si se quisiera secuenciar una especie como el olivo (*Olea europaea*) con un tamaño estimado de 1.9 Gb se necesitarían al menos 57 Gb o 190 Gb de lecturas procesadas, producidas por *454* o *Illumina* respectivamente (sin contar posibles problemas inherentes a esta especie como su elevada heterocigosidad).

3- Se consciente de que no sólo es un problema de secuencias: ¿qué recursos computacionales y genéticos son necesarios?

Otro factor a tener en cuenta son los recursos computacionales disponibles ya que los ensambladores utilizan una gran cantidad de memoria RAM. Por ejemplo, el genoma de *Nicotiana benthamiana* (con un tamaño estimado de 3 Gb, secuenciado con una cobertura de 63X, 229.2 Gb de lecturas [23]) no pudo ensamblarse en un servidor de 512 Gb de memoria RAM, y tuvo que crearse un *subset* de datos con 2/3 del set original para ajustarse a los recursos disponibles a partir del cual se creó el ensamblaje base. Los *gaps* fueron completados con el set completo de datos en una operación que necesitó menos recursos computaciones. Respecto al software utilizado para el ensamblaje, dependerá en gran medida de la tecnología de secuenciación y la metodología utilizada, pero los más populares son *AllPath_LG* [42] y *SOAPdenovo* [43]. Una vez el genoma está ensamblado es conveniente mapear las lecturas y llamar SNPs para valorar la heterocigosidad del genoma, o en caso de poliploides, estimar el porcentaje de colapso entre regiones homoeólogas.

Una vez se ha conseguido un ensamblaje, el siguiente paso es asignar los *contigs* y *scaffolds* producidos a diferentes cromosomas usando mapas genéticos y marcadores moleculares. A este proceso se le denomina anclaje de secuencias en pseudomoléculas. Para este proceso es importante tener mapas genéticos de alta densidad con un gran número de marcadores. El número necesario de marcadores para anclar un ensamblaje dependerá de la calidad de este. Por ejemplo, el ensamblaje de *N. benthamiana* posee un valor de N90=30,261 (es decir que el 90% del ensamblaje está representado por 30,261 secuencias) de manera que el anclaje del 90% del ensamblaje requeriría al menos de 60,522 marcadores (dos marcadores por secuencia para poder orientarlas). Si bien el uso de NGS puede generar cientos de miles de marcadores, su uso para la creación de un mapa dependerá del tamaño de la población utilizada y del numero de eventos de recombinación producidos al generar la población de mapeo. El uso de *Genotyping-By-Sequencing* (GBS) [44] y *microarrays* de genotipado [45]ha permitido impulsar la creación de mapas varios órdenes de magnitud hasta miles de marcadores. En el caso de *N. benthamiana* se necesitaría mejorar el ensamblaje al menos un orden de magnitud (N50 ~ 6,000) antes de abordar un anclaje con un mapa de alta densidad. Ensamblajes con *scaffolds* de mayor tamaño disminuyen el número de marcadores necesarios para anclar el ensamblaje. Por ejemplo para la versión 2.40 de *S. lycopersicum* donde el 95% del ensamblaje está contenido en 72 *scaffolds* de al menos 1.96 Mb se usaron dos mapas físicos y un mapa genético. En total se ancló un 97% del ensamblaje [16].

4- Buscando el ensamblaje útil: cómo anotar un ensamblaje

Independientemente de que se haya o no anclado una gran parte del genoma, un paso determinante en un proyecto de secuenciación de un genoma es la anotación estructural del mismo. Existen dos anotaciones estructurales que comúnmente se utilizan sobre cualquier ensamblaje: Repeticiones y genes:

- Para la anotación de repeticiones se analiza el número de ocurrencias de distintos fragmentos del genoma y se compara con bases de datos de repeticiones como *RepBase* [46]. La herramienta más utilizada es *RepeatModeler* como *pipeline* que integra *RepeatScout* [47].
- La anotación de genes es algo más compleja. Se combinan dos tipos de metodologías: Predicciones de-novo y creación de modelos de genes basados en alineamientos con tránscritos. En el primer caso un programa analiza la secuencia producida en el ensamblaje en busca de marcos de lectura (ORF). Los programas más usados son *Augustus* [48], *SNAP* [49] o *GeneMark* [50]. En el segundo caso se necesita una buena representación del transcriptoma lo que implica el uso de diferentes librerías de ESTs (*Expressed Sequence Tags*) ensambladas en unigenes o diferentes *sets* de datos de *RNAseq*. Por ejemplo en la anotación del genoma de *N. benthamiana* se usó librerías de *RNAseq* de hoja, raíz, flores y distintos estreses bióticos y abióticos a fin de capturar la

máxima diversidad transcriptómica. Como programas se usan *Exonerate* para unigenes [51]o *Tophat* para *RNAseq* [52]. Otra alternativa es el uso de la secuencia de la proteína tal y como hace *GeneWise* [53]. Todos estos programas suelen usarse en una *pipeline* de análisis que integra los resultados de las predicciones *de novo* y de las predicciones basadas en alineamientos con tránscritos. La más popular es *Maker* [54]. Al igual que los ensamblajes, que requieren de un buen poder computacional, la anotación de genomas requiere del uso de sistemas multinúcleo o granjas de computadores con sistemas *MPI* o *Sun Grid Engine* a fin de realizar la anotación en una cantidad razonable de tiempo. Por ejemplo, la anotación del genoma de *N. benthamiana* se realizó en un servidor con 64 núcleos y 512 Gb de memoria RAM (aunque en este caso no llegó a usarse más de 32 Gb) durante aproximadamente 20 días.

Una vez se ha generado una anotación estructural es conveniente visualizar los resultados usando un navegador genómico (*Genome Browser*). Si bien los más populares son UCSC Genome Browser [55]y Gbrowse [56], son navegadores difíciles de instalar generalmente orientados a ser usados por una base de datos. Es más adecuado el uso de programas orientados a una instalación local. Un buen ejemplo es IGV (Integrative Genome Viewer) que permite además cargar otro tipo de datos (como mapas de secuencias) sobre la anotación estructural [57].

La anotación estructural no asigna posibles funciones a cada uno de los genes producidos durante el proceso de anotación. Para ello es necesario efectuar una anotación funcional de los mismos comparando las secuencias de CDS o proteínas predecidas con las bases de datos existentes. Generalmente se utilizan dos aproximaciones complementarias: La primera, la búsqueda de genes homólogos a través de alineamientos de estas secuencias con las secuencias de distintas bases de datos usando el algoritmo de Smith-Waterman. La herramienta más utilizada es *Blast* [58]y las bases de datos más comunes son GenBank [59], SwissProt y TrEmbl [60]. También se suelen comparar con los modelos de las especies más conocidas dentro del clase de estudio, por ejemplo, en plantas suele utilizarse arabidopsis (*Arabidopsis thaliana*) y arroz (*Oryza sativa*) como modelo para dicotiledóneas y monocotiledóneas respectivamente. La segunda es la búsqueda de homología de los dominios funcionales de las proteínas predecidas. La herramienta usada para ello es *InterProScan* y la base de datos usada InterPro, compuesta a su vez por diferentes bases de datos de dominios como Pfam o Panther [61]. La anotación funcional de dominios lleva asociada la asignación de términos basados en categorías de vocabulario controlado procedentes de las ontologías de genes (Gene Ontology, GO terms) [62].

Genómica comparativa. Más allá de un genoma

El proceso de ensamblaje y anotación puede resumirse en varios ficheros fasta con las secuencias producidas durante el ensamblaje y la anotación (*contigs, scaffolds,* pseudomoléculas, genes, ARNm, secuencias codificantes, proteínas y repeticiones) y *.gff3* con la información de mapeado (cómo se integran los *contigs* en los *scaffolds* o cómo se integran éstos en las pseudomoléculas, o cómo son las relaciones estructurales entre los distintos elementos de la anotación como genes y exones)[63]. Pero es justo en este momento cuando comienza el verdadero análisis, cuando se puede buscar sentido a los datos obtenidos generalmente a través de la comparación con otras especies secuenciadas anteriormente. Los análisis más comunes son:

- Análisis de familias de genes. Los genes producidos se agrupan con genes de otras especies basados en porcentajes de homología entre proteínas. El programa más utilizado es Ortho-MCL [64]. Este análisis también permite filtrar aquellos genes provenientes de múltiples repeticiones de transposones ya que generalmente se agrupan en familias con una enorme cantidad de genes de la misma especie.
- Análisis de enriquecimiento de términos GO. También es común la agrupación de genes por categorías funcionales y su comparación con otras especies. Para ello se aplican tests estadísticos como el análisis de enriquecimiento [65]a fin de discernir si existen categorías funcionales sobrerrepresentadas en la especie analizada.
- Análisis de sintenia con otras especies. Consiste en comparar el orden lineal de los genes entre especies distintas a fin de descubrir el grado de conservación de dicho orden. En especies más cercanas desde un punto de vista filogenético cabe esperar una mayor conservación en el orden de los genes. Existen varias herramientas diseñadas para el estudio de sintenia entre especies. Destacan SyMap [66]por su uso sencillo y MCScanX [67]por su capacidad de computar los valores de Ks (ratio de sustituciones sinónimas) a fin de estimar la edad de divergencia entre bloques de sintenia. Este tipo de análisis pueden utilizarse para el estudio de WGD (*Whole Genome Duplications*) a través de los bloques sintenia internos del genoma en cuestión.

Perspectivas para un Futuro Inmediato

El desarrollo de nuevas metodologías y aplicaciones en el campo de la secuenciación es un proceso continuo y rápido. Desde que se diseña un proyecto de secuenciación hasta que finalmente llega la financiación ocurren cambios en tecnologías conocidas o aparecen nuevas aplicaciones que pueden modificar parte del plan original. Son comunes la disminución en los precios de secuenciación o la generación de lecturas de mayor longitud por el mismo precio. Otros cambios interesan-

tes son el desarrollo de nuevos protocolos y reactivos para crear librerías de pares con insertos de mayor tamaño como *Nextera®* [68] o N*xSeq® 40 Kb* librerías [69]. También está por ver si el desarrollo de métodos como el de *Moleculo®* para secuenciar fragmentos únicos de 10 Kb [70] o el uso de la tecnología de fragmentos largos (*Long Fragment Read,* LFR)[71]solventará parte de los problemas derivados de genomas poliploides o de una elevada heterocigosidad. Desde un punto de vista computacional existe un continuo avance en el desarrollo de nuevos procesadores y el abaratamiento de la memoria RAM que hace más accesible la compra de grandes equipos por parte de pequeños grupos de investigación. A largo plazo existe la posibilidad de que cambie de forma radical los sistemas de computación por ejemplo mediante el desarrollo de una nueva generación de transistores de grafeno de alto rendimiento [72,73]con frecuencias mucho mayores que los tradicionales transistores de silicio.

Conclusión

La secuenciación de genomas, lejos de convertirse en un proceso rutinario, es un proceso accesible a cualquier grupo de investigación siempre y cuando posea los medios adecuados y el genoma no sea excepcionalmente grande o polimórfico. El número de genomas eucariotas de tamaño medio se ha multiplicado exponencialmente en los últimos años no solo abriendo la puerta a interesantes estudios evolutivos, sino también generando una importante fuente de recursos para el estudio de enfermedades, la generación de herramientas para la mejora de animales y plantas o la caracterización de la biodiversidad poblacional de cientos de individuos de la misma especie. El desarrollo de las tecnologías de la secuenciación y de análisis de la información están acelerando la generación de conocimiento a límites impensables a principios de siglo. Queda por ver que depara el futuro de la genómica y la bioinformática y cual será el papel que juegue en la ciencia, y a más largo plazo en la historia del ser humano.

AGRADECIMIENTO: El autor quiere agradecer al Dr. Noe Fernández la ayuda prestada en la edición del artículo.

Bibliografía citada:

1.Fiers, W., Contreras, R., Duerinck, F. & Haegeman, G. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature (1976).
2.Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Meth. Enzymol. 155, 335–350 (1987).
3.Sutton, G. G., WHITE, O., Adams, M. D. & KERLAVAGE, A. R. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Science and Technology 1, 9–19 (1995).
4.Fleischmann, R. D., Adams, M. D., White, O. & Clayton, R. A. Whole-genome random sequencing and assembly of Haemophilus. Science 269, 496–512 (1995).
5.Goffeau, A. et al. Life with 6000 genes. Science 274, 546–563–7 (1996).
6.C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282, 2012–2018 (1998).
7.Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000).
8.Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
9.Venter, J. C. et al. The Sequence of the Human Genome. Science Signaling 291, 1304 (2001).
10.Aparicio, S. Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. Science 297, 1301–1310 (2002).
11.Chinwalla, A. T. et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).
12.Batzoglou, S. ARACHNE: A Whole-Genome Shotgun Assembler. Genome Res 12, 177–189 (2002).
13.Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M. & Fasulo, D. P. A Whole-Genome Assembly of Drosophila. Science (2000).
14.Zhang, Q. et al. The genome of Prunus mume. Nat Commun 3, 1318 (2012).
15.Naim, F. et al. Advanced Engineering of Lipid Metabolism in Nicotiana benthamiana Using a Draft Genome and the V2 Viral Silencing-Suppressor Protein. PLoS ONE 7, e52717 (2012).
16.Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641 (2012).
17.Zhang, G. et al. Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nat Biotechnol 30, 549–554 (2012).
18.Bennetzen, J. L. et al. Reference genome sequence of the model plant Setaria. Nat Biotechnol 30, 555–561 (2012).
19.Garcia-Mas, J. et al. The genome of melon (Cucumis melo L.). P Natl Acad Sci Usa (2012). doi:10.1073/pnas.1205415109
20.Wang, Z. et al. The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. Plant J (2012). doi: 10.1111/j.1365-313X.2012.05093.x
21.Wu, H.-J. et al. Insights into salt tolerance from the genome of Thellungiella salsuginea. P Natl Acad Sci Usa (2012). doi:10.1073/pnas.1209954109
22.D'Hont, A. et al. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature (2012). doi:10.1038/nature11241
23.Bombarely, A. et al. A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. Mol. Plant Microbe Interact. (2012). doi:10.1094/MPMI-06-12-0148-TA
24.Wang, K. et al. The draft genome of a diploid cotton Gossypium raimondii. Nat Genet 44, 1098–1103 (2012).
25.Xu, Q. et al. The draft genome of sweet orange (Citrus sinensis). Nat Genet 45, 59–66 (2012).

26.Wu, J. et al. The genome of the pear (Pyrus bretschneideri Rehd.). Genome Res (2012). doi:10.1101/gr.144311.112
27.Paterson, A. H. et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492, 423–427 (2012).
28.Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380 (2005).
29.Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).
30.Valouev, A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18, 1051–1063 (2008).
31.Rothberg, J. M. et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352 (2011).
32.Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138 (2009).
33.Bennett, M. D. & Leitch, I. J. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Ann Bot-London (2011).
34.Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–U113 (2011).
35.Bowers, J., Chapman, B., Rong, J. & Paterson, A. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422, 433–438 (2003).
36.Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467 (2007).
37.Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183 (2010).
38.Novak, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. Bmc Bioinformatics 11, 378 (2010).
39.Lim, K., Matyasek, R., Lichtenstein, C. & Leitch, A. Molecular cytogenetic analyses and phylogenetic studies in the Nicotiana section Tomentosae. Chromosoma 109, 245–258 (2000).
40.Ng, P. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic Acids Res 34, e84–e84 (2006).
41.Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res 19, 521–532 (2009).
42.Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. P Natl Acad Sci Usa 108, 1513–1518 (2011).
43.Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18 (2012).
44.Lu, F. et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9, e1003215 (2013).
45.Desjardins, C. A. et al. Fine-scale mapping of the Nasonia genome to chromosomes using a high-density genotyping microarray. G3 (Bethesda) 3, 205–215 (2013).
46.Kapitonov, V. V. & Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9, 411–2– author reply 414 (2008).
47.Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1, i351–8 (2005).
48.Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34, W435–W439 (2006).
49.Korf, I. Gene finding in novel genomes. Bmc Bioinformatics 5, 59 (2004).
50.Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res (1998).
51.Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. Bmc Bioinformatics 6, 31 (2005).
52.Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
53.Birney, E. Using GeneWise in the Drosophila Annotation Experiment. Genome Res 10, 547–548 (2000).
54.Cantarel, B. L. et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188–196 (2007).
55.Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. Brief. Bioinformatics 14, 144–161 (2013).
56.Stein, L. D. Using GBrowse 2.0 to visualize and share next-generation sequence data. Brief. Bioinformatics 14, 162–171 (2013).
57.Robinson, J. T. et al. Integrative genomics viewer. Nat Biotechnol 29, 24–26 (2011).
58.McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res (2004).
59.Benson, D. A. et al. GenBank. Nucleic Acids Res 41, D36–42 (2013).
60.Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, bar009 (2011).
61.Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol. Biol. 396, 59–70 (2007).
62.Gene Ontology Consortium. Gene Ontology annotations and resources. Nucleic Acids Res 41, D530–5 (2013).
63.Moore, B., Fan, G. & Eilbeck, K. SOBA: sequence ontology bioinformatics analysis. Nucleic Acids Res 38, W161–4 (2010).
64.Li, L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res 13, 2178–2189 (2003).
65.Subramanian, A. et al. Application of a priori established gene sets to discover biologically important differential expression in microarray data. P Natl Acad Sci Usa 102, 15278–15279 (2005).
66.Soderlund, C., Bomhoff, M. & Nelson, W. M. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res 39, e68 (2011).
67.Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40, e49 (2012).
68.Kaper, F. et al. Whole-genome haplotyping by dilution, amplification, and sequencing. P Natl Acad Sci Usa (2013). doi:10.1073/pnas.1218696110
69.Wu, C. C., Ye, R., Jasinovica, S., Wagner, M. & Godiska, R. Long-span, mate-pair scaffolding and other methods for faster next-generation sequencing library creation. Nat Meth (2012).
70.Waldbieser, G. Production Of Long (1.5kb – 15.0kb), Accurate, DNA Sequencing Reads Using An Illumina HiSeq2000 To Support De Novo Assembly Of The Blue Catfish Genome. Plant and Animal Genome XXI Conference (2013).
71.Peters, B. A. et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487, 190–195 (2012).
72.Wu, Y. et al. High-frequency, scaled graphene transistors on diamond-like carbon. Nature 472, 74–78 (2011).
73.Nakaharai, S. et al. Electrostatically-reversible polarity of dual-gated graphene transistors with He ion irradiated channel: Toward reconfigurable CMOS applications. in 2012 IEEE International Electron Devices Meeting (IEDM) 4.2.1–4.2.4 (IEEE, 2012). doi:10.1109/IEDM.2012.6478976