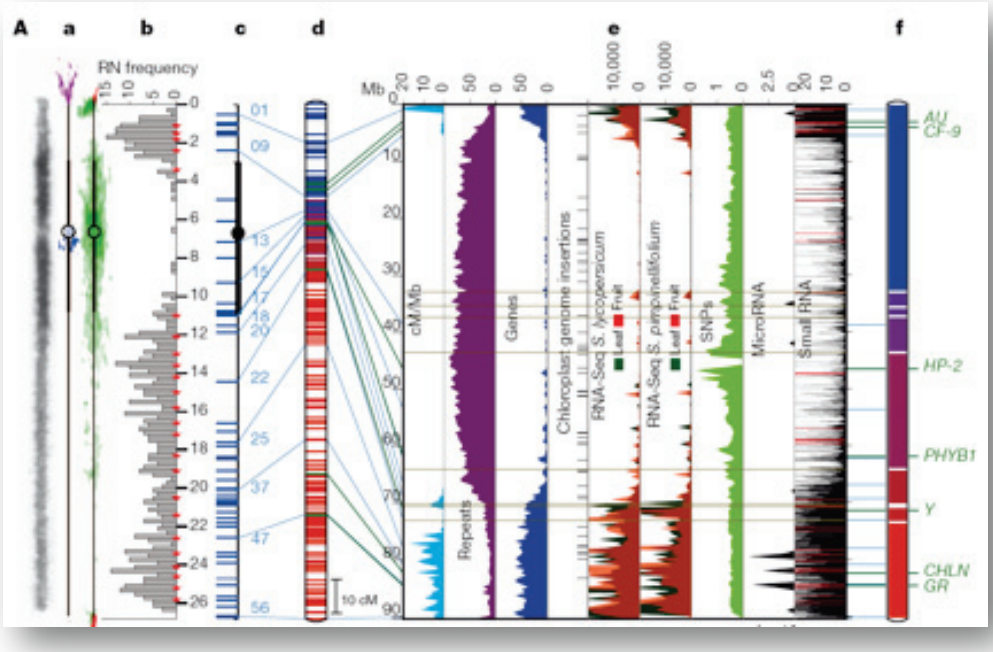
La participación española en el proyecto de secuenciación del tomate y de su pariente más próximo *S. pimpinellifolium*

Antonio Granell Richart
Instituto de Biología Molecular y Celular de Plantas de Valencia,
Universidad Politécnica de Valencia
agranell@ibmcp.upv.es

Este creo que ha sido el primer genoma de planta de cultivo en cuya secuenciación nuestro país ha participado significativamente; ha sido un camino largo y tortuoso, pero lo importante es que hemos obtenido un recurso importante para la comunidad científica y además fue portada en la revista "*Nature*". Nos comprometimos a participar en el proyecto internacional de secuenciación del genoma del tomate en el 2003. Siendo España a la sazón el principal exportador europeo de tomate en fresco era natural pesar en su inclusión. Y también influyó claro, el que relacionado con su importancia estratégica existieran en nuestro país grupos de investigación reconocidos que utilizaban el tomate como sistema modelo. El proyecto de secuenciación nació tras una reunión que tuvo lugar en Washington ese mismo año y en la que científicos representantes de una serie de países consideramos de interés el obtener una secuencia de calidad del genoma del tomate que sirviera como referencia para ésta y otras Solanáceas entre las que se incluyen plantas de claro interés agrícola como el pimiento, la berenjena, la petunia, etc. Lograrlo se consideraba como fundamental para entender la diversidad y adaptación de las Solanáceas a ambientes muy diversos que van desde los desiertos como el Atacama a la pluviselva y del litoral hacia alturas superiores a los 4000 m, y también para entender las bases genéticas de los caracteres de interés agronómico. Todo ello permitiría mejorar la calidad y tolerancia de los cultivos de Solanáceas en unas condiciones de clima cambiantes. La secuenciación pues se planteó desde el principio como un proyecto de alto interés científico que podría tener importante repercusiones económicas y que permitiría acelerar y dirigir los programas de mejora. A pesar de las reticencias iniciales por parte de algunos socios, se acordó que no solo la secuenciación tenía que ser una tarea consorciada, sino que además la información tenía que ser liberada al dominio público tan pronto como su calidad fuera contrastada. La idea era generar un recurso que estuviera disponible a toda la comunidad: ese recurso tenía que ser fácilmente accesible, sin restricciones y de alta calidad para ser realmente útil. Para ello se habilitó un portal web donde se iban depositando las secuencias, y que se iba actualizando y refinando con anotaciones de forma continuada. [http://solgenomics.net/].

La secuenciación se realizó inicialmente utilizando tecnología de Sanger que permitía lecturas largas de alta calidad sobre colecciones de BACs que representaban el genoma del tomate de la variedad Heinz en fragmentos de unas 130kb. Esta estrategia respondía a que las regiones ricas en genes se localizan, en el caso del tomate, en regiones eucromáticas distales de los cromosomas, las cuales están bastante delimitadas de las heterocromáticas pericentrométricas, por lo que se propuso secuenciar solo los BACs que mapeaban en las regiones eucromáticas ricas en genes. El método consistía pues en identificar para cada cromosoma de los 12 del tomate una serie de BACs distribuidos a lo largo de las regiones eucromáticas e ir irradiando a partir de esos BACs semillas a los siguientes, basándose en una serie de técnicas como *fingerprints*, secuencias de los extremos de BACs, etc, para elegir el BAC contiguo con mínimo solapamiento. Esa forma de hacer las cosas era la clásica en el momento inicial del proyecto, producía secuencias de alta calidad pera era muy costosa y lenta. Nuestro grupo fue encargado de la secuenciación del cromosoma 9 y nos pusimos a seleccionar, confirmar y secuenciar los BACs correspondientes a este cromosoma con la participación de una empresa española de secuenciación. Con la incorporación de las nuevas tecnologías de secuenciación masiva 454, *Illumina*, etc., que suponen un abaratamiento en la secuenciación y un incremento en el cantidad de secuencias, propusimos tomar un cambio de estrategia.

En 2009 un pequeño consorcio formado por grupos de Holanda, Italia, Francia, EEUU y España decidimos secuenciar no solo las regiones eucromáticas sino todo el genoma utilizando las nuevas tecnología de secuenciación de nueva generación. Nos dividimos las tareas y cada grupo construyó librerías de DNA de tamaño diferente, desde *shotgun* de pequeño tamaño hasta librerías de tamaños grandes para cada una de las tecnología disponibles, 454, *Illumina* y SOLID. Nuestro grupo se encargó de generar genotecas de 6kb para secuenciación SOLID. Cada una de las genotecas obtenidas por el consorcio se secuenció hasta que no generaban información nueva y se generaron lecturas con cada una de esas tecnologías que en global representaban 250 veces el genoma del tomate. Lo importante no era disponer de tal profundidad en la secuencia (en promedio cualquier posición debería de haberse leído pues 250 veces aunque eso no es estrictamente así), sino que las lecturas de cada tipo de librería llevaba información que facilita el ensamblado, como por ejemplo en algunos casos se obtenían lecturas apareadas correspondientes a cada uno de los extremos de cada fragmento de DNA del cual teníamos además información de la distancia entre ellas por el tamaño de la librería. Esa información es relevante ya que permite ensamblar trozos distantes cuyos huecos se irán llenando posteriormente con más lecturas. La estrategia consistió en unir primero las lecturas largas generadas por 454 y utilizar las lecturas más cortas de *Illumina* y SOLID para confirmar y verificar la secuencias y su posicionamiento. Dos grupos fundamentalmente en Holanda y Francia llevaron a cabo el ensamblaje global de las secuencias generadas por los diferentes participantes utilizando los ensambladores *Newbler* y *Cabog* respectivamente. El 90 % de los 950 millones de pares de bases (una tercera parte del tamaño del genoma humano) está representado en la versión actual en aproximadamente 80 *contigs* de secuencia continua (el ensamblaje optimo sería 24 seudo moléculas correspondientes a los 12 cromosomas con dos brazos cada uno). La tasa de error es inferior a 1 en 10000 y la mayor parte de la secuencia no ensamblada, como ocurre con otros genomas incluido el humano, el de *Arabidopsis* o el del arroz corresponde a secuencias altamente repetidas con pocos o ningún gen y por lo tanto en principio no muy interesante. Para que la secuencia del genoma sea útil, es necesario alinearlas con los mapas físico y genéticos y nuestro grupo del IBMCP en colaboración con el de la ISHM y otros grupos en EEUU y Italia contribuimos a ello desde nuestra experiencia con poblaciones de líneas consanguíneas y de introgresión que teníamos ampliamente genotipadas. Esto junto con localización i*n situ* mediante FISH permitió confirmar y orientar algunos ensamblajes dudosos. La secuencia del genoma es poco útil sin una anotación adecuada que indique donde se localizan los genes, si hay indicaciones que se expresen y donde, etc. y eso fue la tarea de entre otros grupos españoles especializados localizados en el CRG y en los centros de análisis genómico y de supercomputación de Barcelona.



A grandes rasgos la secuencia del tomate y su comparación con otros genomas secuenciados incluido uva ha permitido averiguar que el tomate ha sufrido varios eventos de triplicación y que durante el tiempo evolutivo algunos de los genes han sido seleccionados y adaptados para dotar al fruto de algunas de sus características como maduración controlada por etileno, su regulación por la luz y la acumulación de carotenoides en el fruto. De forma similar la comparación del tomate con su pariente silvestre más próximo, ha permitido averiguar que hay varios centenares de genes que producirían proteínas truncadas, seguramente no viables y que seguramente son responsables de parte de los cambios fenotípicos que se observan entre el tomate cultivado y los de *S pimpinellifolium*. La información de la secuencia y la anotación así como un "buscador" dotado de diferentes herramientas de búsqueda están disponibles de forma libre y sin ataduras para la comunidad científica internacional a través del portal antes indicado.

Un aspecto a destacar es que fruto de este consorcio formado alrededor de la secuenciación se organizan anualmente reuniones de la comunidad de solanceas (SOL *meetings*), se han implementado plataformas de genotipado y analisis de expresión génica y se ha facilitado la formulación de proyectos de secuenciación de otras especies relacionadas de interés, incluida la secuenciación de numerosas otras solanaceas SOL 100, o 100 variedades para 100 cultivos. Todos estos proyectos y otros que se están planteando nos permitirán reconstruir el proceso de domesticación e identificar de forma más eficiente las bases genéticas de los caracteres de interés. En ese sentido la obtención de la secuencia ha facilitado en este ultimo año la identificación de una serie de mutaciones que afectan la calidad como por ejemplo la mutación "y" que produce mutantes de fruto rosa muy apreciados en el mercado asiático y del "*uniform fruit ripening*" o "u" que afecta la calidad del fruto.

Estos y otros avances recientes están demostrando que el esfuerzo conjunto realizado es útil a la comunidad científica y es de esperar que aumente la precisión y acelere el ritmo de producción de nuevas variedades de mayor calidad y mejor adaptadas a las condiciones de cultivo y que satisfagan las necesidades de consumidores y productores.