GIẢI MÃ BỘ GEN CỦA SẮN
(CASSAVA GENOME)

BÙI CHÍ BỬU

SẮN hay KHOAI MÌ (*Manihot esculenta* Crantz) được trồng ở vùng nhiệt đới thuộc Châu Phi, Châu Á, và Châu Mỹ. Sắn là cây trồng giàu tinh bột tích trữ trong củ và lá sắn cũng có thể ăn được. Sắn là lương thực thực phẩm của hơn 800 triệu người trên toàn cầu. Phần lớn vùng trồng sắn đều chịu ảnh hưởng ít hoặc nhiều của khô hạn, đặc biệt là đầu tư cho canh tác sắn rất thấp (Ceballos và ctv. 2010). Hàm lượng tinh bột cao (20-40 %) làm cho cây sắn trở thành một nguồn năng lượng mong muốn của nhân loại, kể cả mục tiêu lương thực và nhiên liệu sinh học (Balat và Balat 2009; FAO 2008; Schmitz và Kavallari 2009). Tuy nhiên, giá trị dinh dưỡng của nó rất hạn chế vì củ sắn có hàm lượng protein thấp, hàm lượng vi chất dinh dưỡng thấp, hàm lượng cyanogenic cao. Cây sắn rất dễ nhiễm bệnh vi khuẩn (Boher và Verdier 1994), dễ bị sâu hại tấn công truyền bệnh virus (Hillocks và Jennings 2003; Patil và Fauquet 2009). Trong khi đó, rễ củ có thể bị bỏ lại trong đất nhiều tháng liền, trước khi thu hoạch. Sự hư hỏng sau thu hoạch diễn ra nhanh, làm trở ngại cho phát triển kinh tế của nông dân trồng sắn (Reilly và ctv. 2007). Sắn là loài thụ phấn chéo, dị hợp, nên việc nhân giống chủ yếu theo phương pháp vô tính (trồng bằng hom thân được cắt ngắn). Những đặc điểm như vậy đã làm chậm lại tiến trình cải tiếng giống sắn, tính kháng sâu bệnh hại và cải tiến hàm lượng dinh dưỡng trên cơ sở chọn tạo giống truyền thống.

Việc giải mã bộ gen cây sắn (genome sequence) đã được tiến hành. Một bản thảo bộ gen sắn được hình thành với sự hợp tác khoa học có tính chất quốc tế từ một mẫu giống sắn trong ngân hàng gen. Một catalog của những biến dị di truyền phổ biến (common genetic variants) cũng được in ấn tại CIAT, phục vụ cho cả hai mục đích "dinh dưỡng" và "nhiên liệu sinh học". Sắn (2n=36) thuộc họ **Euphorbiaceae**, và **Fabid** superfamily (còn được gọi là **aseurosids I**), chúng còn bao gồm nhiều loài cây trồng có quan hệ khá xa thí dụ như rosids, cây họ đậu và cây poplars (cây dương).

Hình 1: cây rosid (trái) và cây poplar (phải)

Dự án giải mã bộ gen sắn được bắt đầu từ năm **2003** với sự chủ trì của **GGP21** (Global Cassava Partnership), đồng chủ trì bởi Dr. Claude Fauquet, Giám Đốc Phòng thí nghiệm quốc tế ILTAB, Donald Danforth Plant Science Center (DDPSC), và Dr. Joe

Tohme, CIAT. Công việc được thực hiện theo sáng kiến 2006 của Fauquet, Tohme và 12 nhà khoa học quốc tế khác, cộng với chương trình giải mã của JGJ của Hoa Kỳ (Energy Joint Genome Institute).

Kiến nghị của dự án đã được hội đồng khoa học tuyển chọn thành một dự án điểm (pilot project). Sau đó vài năm, người ta thực hiện được 700.000 reads (với ~0.8x random) theo pp "shotgun". Một nửa genome này được lấy mẫu, nhưng chỉ có các đoạn trình tự ngắn (700 bp) cho kết quả tốt.

Tại hội nghị quốc tế về genome thực vật và động vật ở San Diego, Hoa Kỳ; Steve Rounsley thuộc ĐH Arizona chấp nhận một thỏa thuận nhằm hoàn thiện bộ gen sắn của tổ chức **454 Life Sciences** và **JGI**, với sự khuyến khích của Bill & Melinda Gates Foundation (BMGF). Tổ chức 454 Life Sciences và JGI đã chọn lựa công cụ **454's Genome Sequencer FLX Titanium** nhanh chóng tạo ra được các dữ liệu chuỗi trình tự DNA. "454 Life Sciences" sau đó cho kết quả những chuỗi trình tự dài hơn với kỹ thuật "extra long read technology".

Như vậy cơ sở dữ liệu về chuỗi trình tự ban đầu này (raw sequence data) có thể nói đã **hoàn thành vào mùa xuân 2009**, sớm hơn 90 ngày theo kế hoạch đã thỏa thuận tại San Diego.

Trình tự genome ban đầu ấy được thực hiện bằng pp **WGS** (whole genome shotgun).

Nội dung tổng hợp và sắp xếp lại (assembling) bộ gen cây sắn vẫn đang trong giai đoạn thách thức với 2 lý do: (1) các chuỗi trình tự có tính chất lập lại thường có quan hệ với những **transposon**, nằm rải rác giữa các gen chen kín trong một vùng; (2) sắn là loài thụ phấn chéo, có biến dị alen, kể cả chỉ thị SNPs và những thể "đa hình theo kiến trúc" như InDel.

Độ lớn của genome cây sắn ước khoảng **770 Mb** (Awoleye và ctv. 1994) với **18 cặp nhiễm sắc thể**.

Tổng số trình tự trong cơ sở dữ liệu thô là: 22,4 tỷ cặp base (bp), đủ để phủ hết toàn bộ genome khoảng 29 lần.

Cho dù "genome assembly" chỉ có gần **13.000 pieces** (mẫu), nhưng một nửa được bắt giữ trong **487 scaffolds**, mỗi sacaffold có kích thước lớn hơn **258 kbp** và chứa nhiều hơn 49 gen.

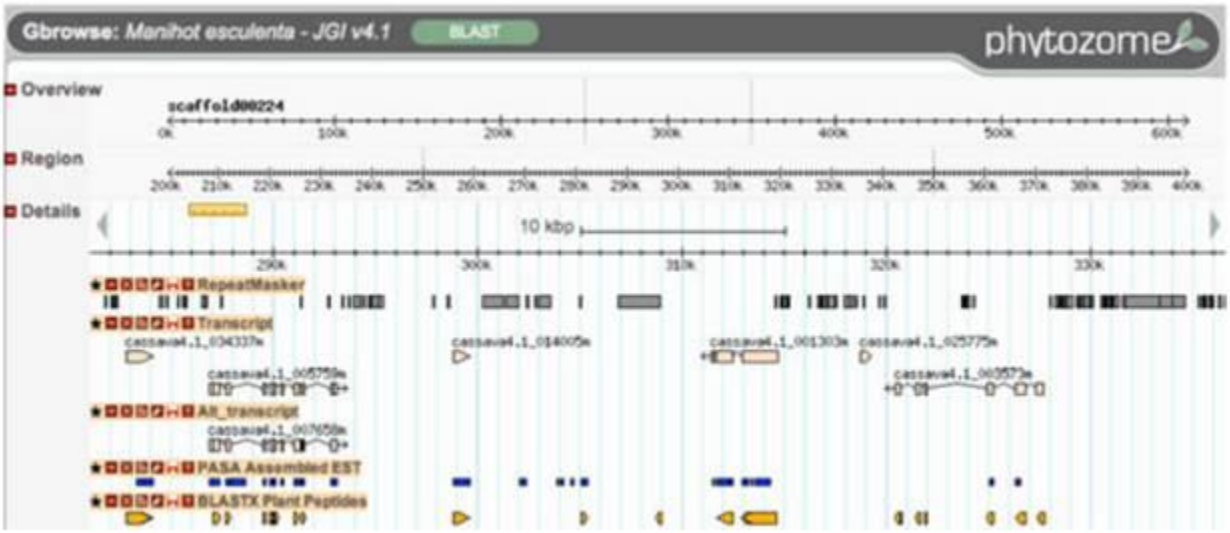
Trước đây, chỉ có 8 transposon của sắn được mô tả trên các cơ sở dữ liệu đại chúng (Kapitonov và Jurka 2008). Theo kết quả mới, một cơ sở dữ liệu định tính khá hoàn chỉnh các nội dung có tính chất lập lại, nhờ kỹ thuật scanning các trình tự nhiều lần.

Bảng 1: Chú thích số lượng các gen mã hóa protein (annotation) trong bộ gen sắn được so sánh với thầu dầu, Arabidopsis , đậu nành và lúa [Simon Prochnik và ctv. 2012].

Đặc điểm	Sắn v. 4.1	Thầu dầu (TIGR v. 0.1)	Arabidopsis thaliana Columbia (TAIR10)	Đậu nành (JGI v. 1)	Lúa (MSU v. 6.0)
Số loci các gen mã hóa protein	30.666	31.221	27.416	46.367	40.838
Phân tử transcript có PFAM domain (KOG orthology)	20.641 (12.307)	16.720 (9.321)	19.419 (12.184)	34.065 (20.601)	20.766 (9.933)

KẾT QUẢ TÓM TẮT

Chiều dài của scaffold	532.5 Mb
Độ dài tổng của trình tự contig	419.5 Mb (21% gaps)
Tổng số scaffolds	12,977
Một nửa assembly là scaffolds dài hơn 258.1 kbp (487 scaffolds)	
Phần được chú thích rồi	37.5% genome
Trình tự Gypsy	140 Mb (10M reads)
rDNA sequence	54 Mb (3.9 M reads, 6.000 bản sao)
Trình tự gen polygalacturonase	12 Mb (850 k reads, 3.000 bản sao)



Hình 2: Tổng quát về "**phytozome genome**" của sắn

TÀI LIỆU THAM KHẢO

Awoleye F, Duren M, Dolezel J et al (1994) Nuclear DNA content and in vitro induced somatic polyploidization cassava (Manihot esculenta Crantz) breeding. Euphytica 76: 19 5–202

Balat M, Balat H (2009) Recent trends in gl obal production and utilization of bio-ethanol fuel. Appl Energ 86:2273 –2282

Boher B, Verdier V (1994) Cassava bacterial blight in Africa: the state of knowledge and implications for designing control strategies. Afr Crop Sci J 2:505– 509

Ceballos H, Okogbenin E, Pérez JC et al (2010) Cassava. In: Bradshaw JE (ed) Root and tuber crops, handbook of plant breeding, vol 7.Springer, New York, pp 53 – 96

FAO (2008) Cassava for food and energy security. FAO Newsroom. http://www.fao.org/newsroom/en/news/2008/1000899/index.html. Cited 19 Nov 2011

Hillocks RJ, Jenning s DL (2003) Cassava brown streak disease: a review of present knowledge and research needs. Int J Pest Manag 49:225 –234

Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9:411– 412, author reply 414Monger WA, Alicai T, Ndunguru J, Kinyua ZM, Potts M, Reeder RH, Miano DW, Adams IP, Boonham N, Glover RH, Smith J. (2010).

The complete genome sequence of the Tanzanian strain of Cassava brown streak virus and comparison with the Ugandan strain sequence. Acta Virol 155(3):429-33

Oteng-Frimpong R, Levy Y, Torkpo SK, Danquah EY, Offei SK, Gafni Y. (2012). Complete genome sequencing of two causative viruses of cassava mosaic disease in Ghana. Acta Virol 56(4):305-314

Patil BL, Fauquet CM (2009) Cassava mosaic gemini viruses: actual knowledge and perspectives. Mol Pl ant Pathol 10:685 – 701

Rebecca Bart et al. (2012). High-throughput genomic sequencing of cassava bacterial blight strains identifies conserved effectors to target for durable resistance. PNAS on line E1972–E1979

Reilly K, Bernal D, Cortes DF et al (2007) Towards identifying the full set of genes expressed during cassava post-harvest physiological deterioration. Plant Mol Biol 64:187 –203

Schmitz PM, Kavallari A (2009) Crop plants versus energy plants-on the international food crisis. Bioorg Med Chem 17:4020– 4021

Simon Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S. (2012). The Cassava Genome: Current Progress, Future Directions. Tropical Plant Biol 5:88 – 94

DIỄN GIẢI THUẬT NGỮ

Shotgun sequencing hay **shotgun cloning** là phương pháp giải trình tự các đoạn phân tử DNA dài, thông qua kỹ thuật dò dẫm trên nhiễm sắc thể (chromosome walking) bằng những contig (BAC clones).
Một **scaffold** là một phần của genome sequence được tái tạo từ các clones trong kỹ thuật shotgun trên toàn bộ genome tính từ đầu đến cuối dây DNA.
Scaffolds bao gồm tất cả những contigs và những đoạn hở (gaps). Gaps xảy ra trong quá trình đọc trình tự của máy từ hai đầu trình tự (two sequenced ends) của ít nhất một lần trùng lập (overlap) với các kết quả đọc khác giữa hai contigs khác nhau.