The Potato Genome Sequencing Initiative

The Potato Genome Sequencing Consortium

Sanjeev Kumar Sharma, Programme of Genetics, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom; sanjeev.sharma@scri.ac.uk

Introduction

Potato is the world's most important vegetable crop, the 3rd largest global food crop and a unique biological system belonging to Solanaceae. In order to decipher the structure and function of its genes, the 840 Mb genome of potato (*Solanum tuberosum* L.) consisting of 12 chromosomes has been sequenced by the global Potato Genome Sequencing Consortium (PGSC*). The PGSC was initiated through Wageningen University and Research Centre and currently comprise member institutions from 15 different countries.

Rationale

Potato is a highly heterozygous tetraploid that suffers severe inbreeding depression upon self-pollination. Despite its importance as a food crop throughout the world, the genetics of many potato traits is poorly understood and is complicated by its polyploid genome. Many important qualitative and quantitative agronomic traits are poorly understood, genes affecting these traits remain largely undiscovered and QTL locations are often imprecise. The sequencing of the potato genome will provide a major boost to gaining a better understanding of potato trait biology and will underpin future breeding efforts.

Initial Sequencing Strategy (2005 onwards)

Sequencing started using a heterozygous diploid potato clone (RH89-039-16) and adopting a chromosome by Chromosome and BAC by BAC Sanger sequencing strategy. RH was chosen because it is the parent of the UHD mapping population with a very extensive genetic map. Sequencing progressed with anchored RH seed BACs, involved 6x coverage and ~ 800 - 1000 BACs per chromosome. The strategy employed RH physical map to choose tiling path across each chromosome and individual PGSC partners were assigned different chromosomes. This led to significant resource and capability development for potato genome sequencing but also had following drawbacks:

- Sanger based BAC by BAC approach was slow
- Heterozygosity of RH limited the progress of physical mapping and complicated the assembly of the genome (Figure 1a)
- Large gaps were present in physical map reducing number of seed BACs
- Only 30-40% of genome covered by the map and average contig tile path was only 2.5 BAC clones
- Disparity in chromosome sequencing progress

Revised Strategy (2008 onwards)

With the advent of Next Generation Sequencing (NGS) technologies, Whole Genome Shotgun (WGS) sequencing has become more feasible and economical (data/$). PGSC reviewed RH sequencing related issues and adopted a revised strategy which mainly involved:

- Additional use of highly homozygous genotype (Figure 1b and 2) to get around heterozygosity and assembly problems of RH (Figure 1a).
- Use of NGS technologies (in addition to Sanger sequencing) to generate WGS sequence of potato
- Improving the RH physical map using WGP™ (Keygene)
- Delegation of tasks according to capability and available resource, rather than a chromosome by chromosome approach

Figure 1: (A) Depiction of heterozygosity and sequencing issues with diploid genotype RH. Each chromosome has two versions (= 'phases') '0' and '1'; WGS and BACs sequence data come from two chromosome versions '0' and '1' and, consequently, RH genome assembly is complicated and requires two separate tiling paths; (B) Homozygous doubled monoploid genome. Each chromosome has same version (only 1 phase and no phase issues). WGS and BACs sequence data come from same chromosome versions and, consequently, resolves DM genome assembly process

Figure 2: The homozygous genotype introduced for sequencing in the revised strategy. Doubled monoploid (DM) homozygous potato (*S. tuberosum* Phureja Group) clone DM 1-3 516 R44 (CIP 801092). The DM phenotype (A) and tubers (B) are shown above. DM flowers well and can be used as a female parent in crosses with most diploid potato germplasm [Paz MM, Veilleux RE (1997) Genetic diversity based on randomly amplified polymorphic DNA (RAPD) and its relationship with the performance of diploid potato hybrids. J. Am. Soc. Hort. Sci. 122: 740-747]

Genome Assembly and Annotation

- A high quality draft sequence assembly (version 3.0) of DM based on Illumina & Roche 454 short reads and Sanger sequenced BAC & Fosmid -ends (Table 1) has been generated by using the short reads assembly software - SOAPdenovo (version - 1014) developed by BGI (Figures 3 and 4, Table 2)
- Assembly of RH is progressing fast using NGS, WGP™ and Sanger data
- Integration of the two genome assemblies will generate three virtual molecules corresponding to the three haplotypes (Figure 5)
- Three gene-prediction methods (Figure 6) applied to annotate protein-coding genes
- Consensus gene set (Table 3) built by merging all genetic resources and prediction approaches
- Validation by deep transcriptome profiling and RNAseq analysis from 16 RH and 29 DM libraries

Sequenced Clone	In Progress	Sanger Sequencing	Illumina Runs	Roche/454 Runs
DM	WGS + 500 bp to 20 kb libraries			10x coverage
	WGS + 200 bp to 10 kb libraries		122 x coverage	
	Fosmid library (~35 kb)	190K Fosmid -end sequences		
	BAC library (>100 kb)	160K BAC -end sequences		

Table 1: Sequencing efforts for DM line. Sequencing methods being employed are listed alongwith estimated coverage of the ~840 Mb potato genome

	Contig Size (Kb)	Contig No.	Scaffold Size (Kb)	Scaffold No.	Super - Sca ffold Size (Kb)	Super - Scaffold No.
N90	06.9	23,392	92.0	1,935	253.8	622
N80	13.1	16,371	168.5	1,366	510.8	423
N70	18.9	12,046	240.2	1,003	784.7	307
N60	24.8	8,893	308.0	735	1068 .6	228
N50	31.4	6,446	386.6	524	1318.5	167
Total Size	682,695	-	727,233	-	727,424	-

Table 2: Statistics of DM v3 assembly

Transcripts/Genes with:	Transcripts		Genes	
	Number	Percentage	Number	Percentage
Protein support	56,770	81.7	33,936	71.7
EST and/or RNA -Seq support	46,777	67.3	34,392	72.6
Protein & EST and/or RNA -Seq support	41,589	59.9	23,220	49.0
Protein or EST and/or RNA -Seq support	60,956	87.8	40,206	84.9
Protein support only	15,181	21.9	10,716	22.6
EST and/or RNA -Seq support only	5,188	7.5	2,715	5.7
Ab initio	8,500	12.2	8,260	17.4

Table 3: Prediction classification at the transcript and gene levels

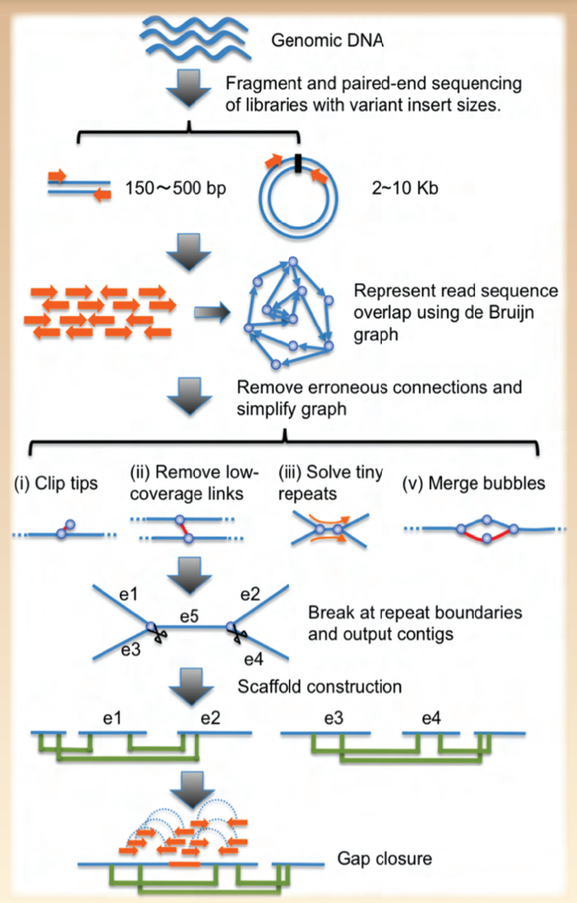

Figure 3: Schematic overview of the assembly algorithm developed by BGI

Mapping/Anchoring

In order to augment the genetic and physical anchoring of the sequenced DM genome, a segregating backcross population (Figure 7) between the DM clone and a heterozygous diploid *S. goniocalyx* clone (CIP No. 703825) as the recurrent parent was established.

The polymorphism across 169 progeny clones (Figure 8) was assessed using a total of 4836 STS (sequence tagged sites) markers including 2174 DArT™, 378 SSR alleles and 2304 SNP marker types. SSR and SNP markers were designed directly to scaffolds, whereas polymorphic DArT marker sequences were searched against the super-scaffolds for high quality unique matches.

The marker data was analysed using JoinMap®4 and a DM genetic map containing the expected 12 potato linkage groups was developed *de novo*. The unique position and prior sequence information of the mapped STS markers facilitated their direct anchoring to the DM superscaffolds. This in turn assisted in physical anchoring of DM superscaffolds on to the DM/DI//DI linkage map. Overall, using other available resources, we are able to genetically anchor 623 Mb (85.7%) of the assembled 727 Mb genome arranged in 651 superscaffolds to an approximate location onto one of the twelve potato linkage groups.

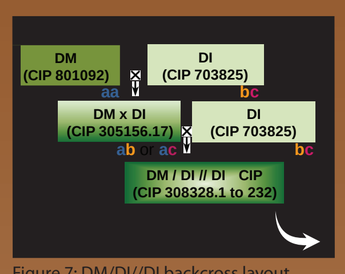

Figure 7: DM/DI//DI backcross layout

Figure 8: DM/DI//DI backcross progeny tubers (DM mapping population) showing phenotypic (genotypic) diversity

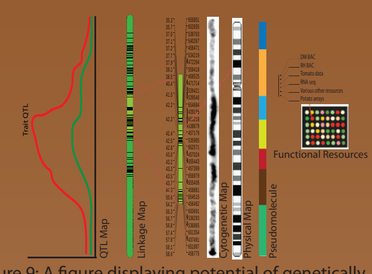

Figure 9: A figure displaying potential of genetically anchored physical map in accelerating move from mapped trait to candidate gene sequences

Benefits of the anchored genome sequence

- Radical effects on efficiency of potato breeding
- Overcome many negative aspects of potato as a genetic system
- Enhance our ability to identify the desirable allelic variants of genes underlying important quantitative traits in potato
- Facilitate gene isolation and allow molecular geneticists to accelerate trait gene discovery
- Shorten the time taken to breed new varieties as well as reducing the cost
- Integrated sequence and genetic reference map will form an important resource for linking to all future genetic mapping efforts by the potato community
- Shift towards sequence based markers
- Will virtually replace centimorgan (cM) position by sequence co-ordinates
- Greatly increase the information output and accuracy of mapping procedures (Figure 9)

The PGSC had made the potato genome assembly (DM v3) available for the public via a genome browser (Figure 10)

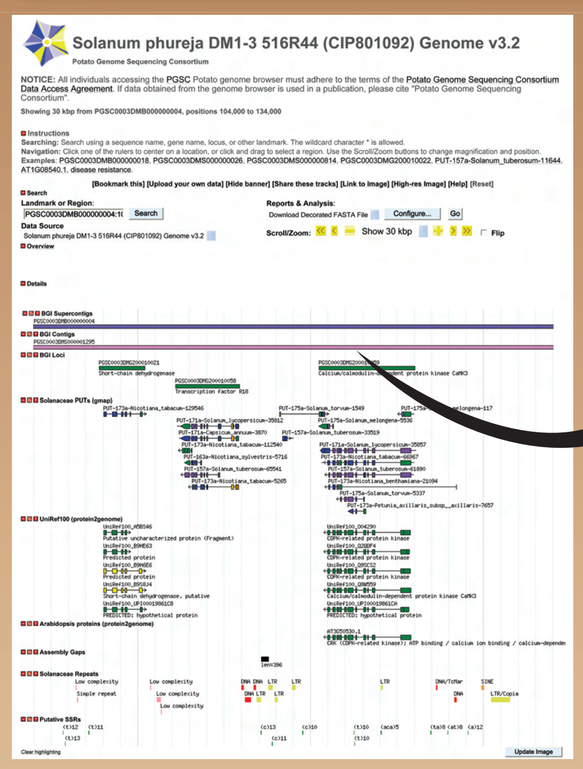

Figure 10. A view of the Public Genome Browser available at "http://www.potatogenome.net/index.php/Main_Page" for searching various features of the assembled and anchored potato genome

Data dissemination

The consortium is committed to open access. All the data produced by the sequencing effort will be released (under a public data access agreement) immediately after assembly and quality control to the wider public. Periodic updates will be made over the next six months as additional data is generated. For more information, visit http://www.potatogenome.net or send an email to potato-genome-i@googlegroups.com.

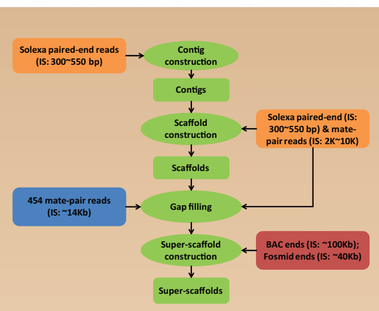

Figure 4: Flowchart of DM genome assembly

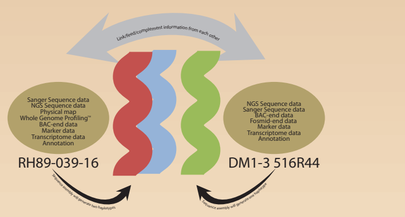

Figure 5: An overview of the DM and RH resource sharing and genome assembling strategy. Integration of the two sequencing strategies will yield three comparable haplotypes

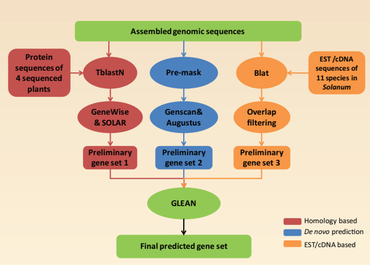

Figure 6: Flowchart of DM gene annotation

Acknowledgements

*

- Aalborg University, Denmark (Mads Sønderkær and Kåre Lehmann Nielsen)
- BGI-Shenzhen, China (Sanwen Huang, Ruiqiang Li, Xun Xu, Wei Fan, Peixiang Ni, Hongmei Zhu, Desheng Mu, Bicheng Yang, Jian Wang and Jun Wang)
- Center Bioengineering RAS, Russia (Boris Kuznetsov)
- Central Potato Research Institute, India (Swarup Chakrabarti, V.U. Patil, Shashi Rawat and S.K. Pandey)
- Chinese Academy of Agricultural Sciences, China (Sanwen Huang, Zhonghua Zhang and Dongyu Qu)
- University of Dundee, United Kingdom (Dan Bolser and David Martin)
- ENEA, Italian National Agency for New Technologies, Energy and the Environment, Italy (Giovanni Giuliano and Gaetano Perrotta)
- Imperial College London, United Kingdom (Gerard Bishop)
- International Potato Center (CIP), Peru (Merideth Bonierbale, Marc Ghislain and Reinhard Simon)
- Institute of Biochemistry and Biophysics (PAS), Poland (Wlodzimierz Zagorski, Jacek Hennig, Pawel Szczesny, Piotr Zielenkiewicz and Robert Gromadka)
- Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina (Gabriela Massa, Leandro Barreiro and Sergio Feingold)
- Instituto de Investigaciones Agropecuarias (INIA), Chile (Boris Sagredo, Alex Di Genova and Nilo Mejía)
- Michigan State University, USA (Robin Buell, Steven Lundback and Brett Whitty)
- New Zealand Institute for Plant & Food Research, New Zealand (Jeanne Jacobs, Mark Fiers and Susan Thomson)
- Scottish Crop Research Institute, United Kingdom (Glenn Bryan, David Marshall, Robbie Waugh and Sanjeev Kumar Sharma)
- Teagasc Agriculture and Food Development Authority, Ireland (Dan Milbourne, Istvan Nagy and Marialaura Destefanis)
- Universidad Peruana Cayetano Heredia, Peru (Gisella Orjeda, Frank Guzman, Michael Torres, Tomas Miranda, German de la Cruz, Roberto Lozano and Olga Ponce)
- Virginia Polytechnic Institute & State University, USA (Richard E. Veilleux)
- Wageningen University, The Netherlands (Bas te Lintel Hekkert, Christian Bachem, Erwin Datema, Jan de Boer, Richard Visser, Roeland van Ham, Theo Borm and Xiaomin Tang)