The Potato Genome Sequencing Initiative

The Potato Genome Sequencing Consortium

Glenn J Bryan (on behalf of PGSC), Programme of Genetics, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom; glenn.bryan@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 is currently being sequenced by the 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-polination. 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.

General Goals

- Sequence the complete genome of potato by early 2010
- Build capacity in countries with less developed plant genomics infrastructure
- Form the basis of a research network for the scientific exploitation of the sequence data in the post-genomics era

Specific Goals

- More than 95 % of genes plus regulatory regions
- More than 95 % of ESTs (>250 bp, 10 Ns)
- More than 50 % of the genome anchored to chromosome
- Complete set of annotated genes
- N50 contig size > 15 kb
- N50 scaffold size > 0.5 Mb

Sequencing Strategy

Initial Strategy

Started in 2005/6 taking 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 started with anchored RH seed BACs, involved 6x coverage and ~ 800 - 1000 BACs per chromosome

Employed RH physical map to choose tiling path across each chromosome and individual PGSC partners were assigned different chromosomes

Problems With Initial Sequencing Strategy

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

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
- 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]

Mapping/Anchoring

- Aim to anchor >50 % of the genome assembly to a genetic map, this is supplemented by an improved physical map of RH using Whole Genome Profiling™ and the development of an anchored genetic reference map based on DM
- Backcross between DM and heterozygous DI (CIP No. 703825), a heterozygous diploid *S. goniocalyx* clone, comprise ~200 progeny clones, generated by International Potato Center, Peru
- Scaffolds are being anchored to a genetic map with different types of sequenced markers - SSRs, DArT, SNPs

Additional resources: 148 Sequence Tagged Markers (STM), known to map to regions spanning all 12 chromosomes, ~60 Ste markers, currently being mapped in an SH x RH population

SNP markers: 1920 SNPs (5 Illumina Goldengate OPAs) designed for Illumina BeadXpress platform, uniformaly cover (every ~150 kb) the entire DM genome, uniquely selected from ~75000 SNPs designed using potato EST data aligned to DM genome assembly (courtesy - Robin Buell, MSU, USA SolCAP project)

SSRs: 550 SSR markers designed directly to DM scaffolds

DArT data: Discovery arrays with over 30k probes, discovered 7500 candidate markers, DArT markers have been sequenced and these will provide direct anchoring to scaffolds, DM DArT map (~500 – 700 unique markers) constructed (Figure 3)

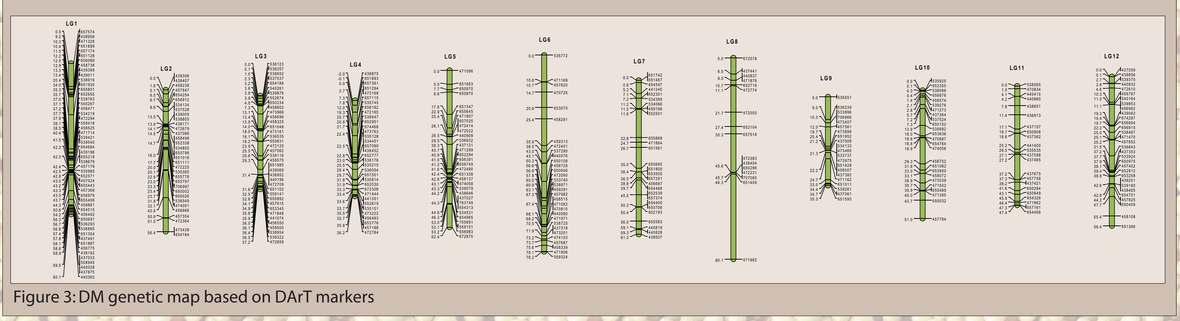

Figure 3: DM genetic map based on DArT markers

Genome Assembly and Annotation

Genome Assembly

- First draft assembly of DM based on Illumina short reads and Sanger sequenced BAC-ends and Fosmid-ends (Table 1) has been generated by using the short reads assembly software - SOAPdenovo (version - 1014) developed by BGI (Figures 4 and 5, Table 2)
- Assembly of RH is progressing using NGS, WGP and Sanger data (Table 3)
- Integration of the two genome assemblies will generate three virtual molecules corresponding to the three haplotypes (Figure 6)

Structural and Functional Annotation

- Three gene-prediction methods (Figure 7) applied to annotate protein-coding genes
- Consensus gene set (Table 4) built by merging all genetic resources and prediction approaches
- Validation by deep transcriptome profiling and RNAseq analysis

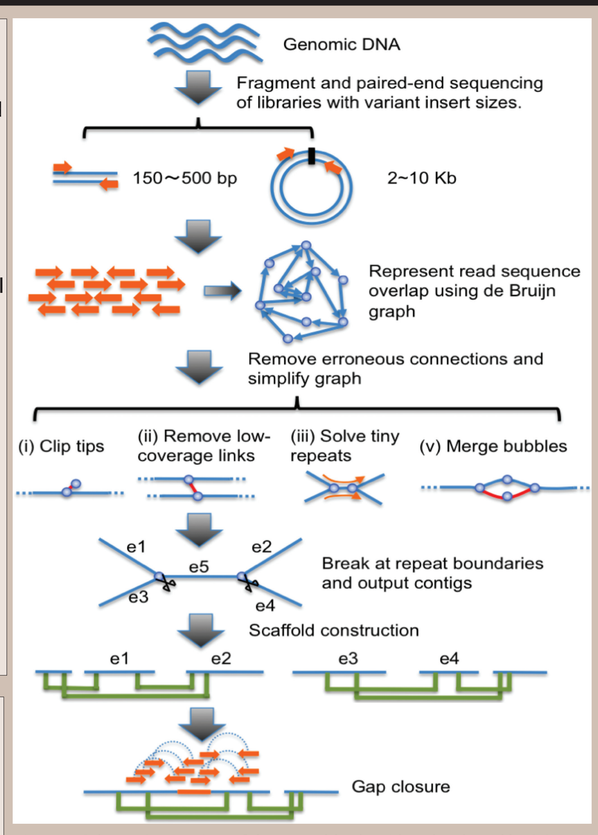

Figure 4: Schematic overview of the assembly algorithm developed by BGI

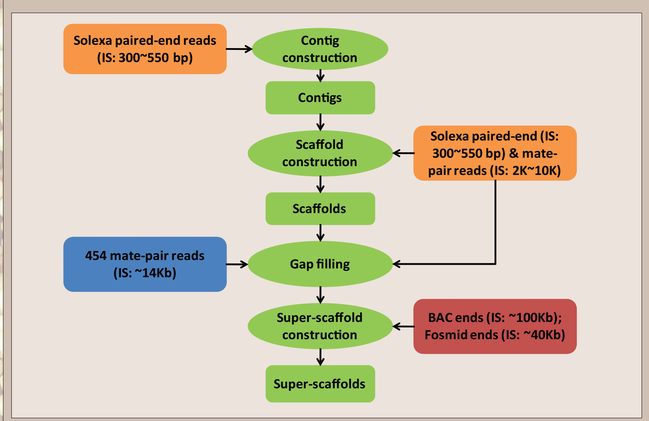

Figure 5: Flowchart of DM genome assembly

Sequenced Clone	In Progress	Sanger Sequencing	Illumina Runs	Roche/454 Runs
DM	WGS + 500 bp to 20 kb libraries			15x coverage
	WGS + 200 bp to 10 kb libraries		65x 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 - Scaf fold Size (Kb)	Super - Scaffold No.
N90	06.9	23,392	092.0	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	08,893	307.9	735	1068.6	228
N50	31.4	06,446	386.6	524	1318.5	167
Total Size	682,695	-	727,233	-	727,424	-

Table 2: Statistics of DM assembly

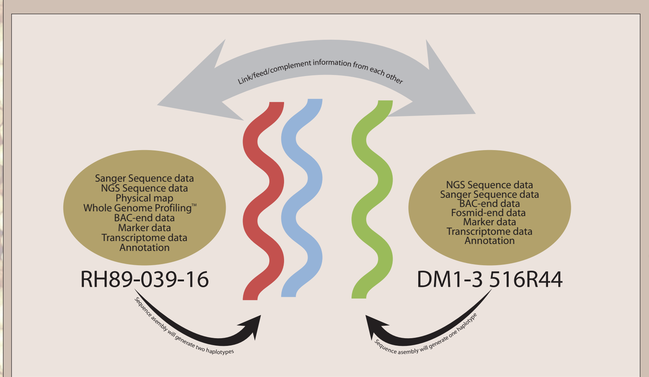

Figure 6: An overview of the DM and RH resource sharing and genome assembling strategy. Integration of the two sequencing strategies will yield three comparable haplotypes

Sequenced Clone	In Progress	Sanger Sequencing	Illumina Runs	Roche/454 Runs
RH	WGS + Long Jump libraries			10x coverage
	WGS		120 x coverage	
	BAC library	150K BAC -end sequences + 2K BAC cl ones		
	Random sheared BAC library (~100 kb)	120K BAC -end sequences		
	Whole Genome Profiling™	54K BACs with 34 tags on average merged into a ~3400 contig physical map		

Table 3: Sequencing efforts for RH line. Sequencing methods being employed are listed alongwith estimated coverage of the ~840 Mb potato genome

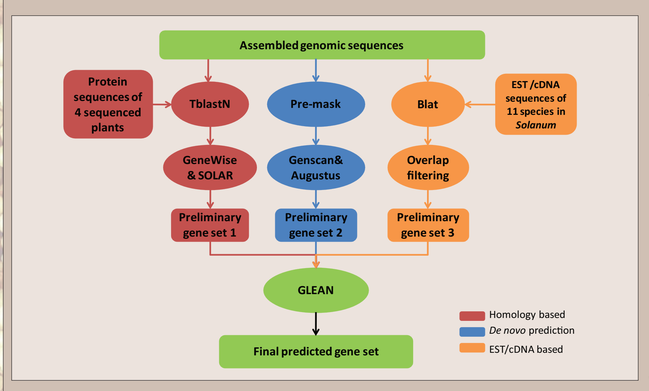

Figure 7: Flowchart of DM gene annotation

Type	Number	Average length (bp)	Total length (bp)
gene	40,842	3420.0	136,952,139
miRNA	329	114.4	37,623
rRNA	375	189.6	71,092
snRNA	546	128.0	69,888
tRNA	881	74.9	65,989

Table 4: Gene annotation from DM genome

Ongoing and Future Steps

- Increase scaffold size and generate hybrid assembly using Solexa, Roche 454 and Sanger data
- Quality assessment of the DM assembly by Sanger-sequenced DM BACs
- Anchor genome assembly to a genetic map
- Develop informatics tools to integrate resources (physical map, genome sequence, marker/gene data)
- Complete potato genome sequence by early 2010

Benefits

- 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 use candidate gene approaches for trait gene discovery
- Shorten the time taken to breed new varieties as well as reducing the cost

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

Acknowledgements

- BGI-Shenzhen, China (Samwen 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 Barrieto and Sergio Feingold)
- Instituto de Investigaciones Agropecuarias (INIA), Chile (Boris Sagredo, Alex Di Genova and Nilo Mejia)
- 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)