Development of genetic and genomic resources for breeding improved sweetpotato

Roland Schafleitner, Luz Tincopa, Genoveva Rossel, Jorge Espinoza, Julio Solis, Omar Palomino, Carlos Rivera, Cinthia Quispe, Jack Hou, Ji Young Kim, Wolfgang Gruneberg, Reinhard Simon, Anthony Collins, David Tay and Merideth Bonierbale

International Potato Center – CIP, Avenida La Molina 1895, Apartado 1558, Lima 12, Peru.

*DArT/PL, 1 Wilf Crane Crescent, Yarralumla, ACT 2600, Australia

E-mail: r.schafleitner@cgiar.org

Production of sweetpotato (*Ipomoea batatas* L. (Lam)), an important staple food in Sub-Saharan Africa, is limited by a number of constraints, such as low adaptability of available varieties and landraces, virus diseases, insect pests and drought. Consequently, yields achieved by resource-poor farmers in SSA are low. Improved and well adapted sweetpotato varieties with increased tolerance to biotic and abiotic stresses could significantly contribute to augment productivity and, once available, would have a large positive impact on food and income security in Sub-Saharan Africa. However, breeding efforts are limited by the crop's genetic complexity, lack of information about its genetic resources and access to genomics tools for this crop for modern breeding. To mobilize allelic diversity and to facilitate introgression of desirable alleles into breeding populations, we have established genetic and genomics tools including a well defined Composite Genotype Set and a gene index. Furthermore we have designed and tested more than 200 new microsatellite markers and identified 200 SNP markers in stress response genes. A sweetpotato DArT marker system is under development. For establishing the gene index, we applied Next Generation sequencing technologies to characterize the sweetpotato transcriptome. The index comprises 31.165 contigs and 29.080 singletons and was annotated based on sequence comparisons with known proteins. The Composite Genotype Set and the genomics tools will support trait capture efforts on molecular level, will improve the understanding of the sweetpotato gene pools and finally will enhance access to allelic diversity for breeding improved varieties.

Keywords: Sweetpotato, genetic resources, transcriptome, molecular marker.

Introduction

Sweetpotato, (*Ipomoea batatas* L. (Lam)) the fifth-most important food crop in developing countries, is a rustic crop with generally high levels of stress tolerance. Nevertheless yields are limited by a number of constraints, such as virus diseases, insect pests and drought. Improved and well adapted sweetpotato varieties with increased tolerance to biotic and abiotic stresses could significantly contribute to increasing productivity, which would have a large positive impact on food and income security of resource-poor farmers in Sub-Saharan Africa. However, breeding efforts are limited by the genetic complexity of this hexaploid and highly heterozygous crop and by the lack of genomic resources for this plant.

CIP holds in trust 5,961 sweetpotato accessions including breeding lines, improved varieties, landraces, and wild accessions from 58 countries, which contributes to global conservation efforts. The identification and characterization of a subset of this collection that represents the available diversity for agronomical and resistance traits, nutritional quality and breeding efforts, referred to as a 'composite genotype set' (CGS) will facilitate improving the knowledge of genetic diversity of sweetpotato and contribute to the development of robust molecular tools for exploration of sweetpotato genetic resources. Prior to this work, sequence information for sweetpotato was limited to some 20.000 expressed sequence tags (ESTs) and about 1500 gene sequences deposited in public databases. At PlantGDB an assembly of all genebank-deposited *Ipomoea batatas* sequences is available (//www.plantgdb.org/download/download.php?dir=/Sequence/ESTcontig/Ipomoea_batatas/current_version). This database contains 12.464 sequences, partly contigs and partly singletons. Genomic tools were restricted to a medium density cDNA microarray, medium-density maps and some 100 published SSR markers (Hu et al. 2004, Arizio et al. 2008). Additional genomic

resources such as gene sequences and markers are urgently required to mobilize sweetpotato biodiversity for breeding programs.

Materials and methods

Assembly of the CGS

CIP curators and breeders used CIP databases and evaluation data to establish a sweetpotato CGS. From over 5000 accessions, sweetpotato biodiversity was sampled based on geographical origin and available SSR data. The set was extended with clones with high levels of resistances to abiotic and biotic stresses and high nutritional value. Morphological information and molecular fingerprints (SSR genotypes) are available to facilitate the identification and tracking of the clones of the CGS.

Developing the gene index

For the development of the sweetpotato gene index, shoots from field-grown plants of the sweetpotato *I. batatas* variety "Tanzania" (CIP number 440166) For the development of the sweetpotato gene index, shoots from field-grown plants of the sweetpotato *I. batatas* variety "Tanzania" (CIP accession number 440166) were acclimated to the greenhouse and cultivated in pots for one month. Then drought was imposed by not watering the plants for eight weeks. After that time, leaf and stem tissue was sampled separately and total RNA was produced using the Trizol reagent according to the instructions of the supplier (Invitrogen). Complementary DNA was synthesized from the two RNA batches by Evrogen (www.evrogen.com) and normalized according to Shagin et al., (2002). 7 μg cDNA of each normalized library was submitted to a quarter 454 sequencing run at the School of Biological Sciences, University of Liverpool. The cDNA library of leaves was sequenced with the 454 FLX technology and for sequencing the stem library the 454 FLX TITANIUM system was available.

Sequence cleaning was performed on the CIP High Performance Computer (http://hpc.cip.cgiar.org/). Adaptor primer, oligonucleotide sequences derived from cDNA library construction as well as low complexity regions present in the 454 raw reads were masked using the open source software RepeatMasker 3.2.7 (http://www.repeatmasker.org/RMDownload.html). 20.094 publically available sweetpotato EST sequences were downloaded from http://www.ncbi.nlm.nih.gov/sites/entrez and cleaned from vector sequences using SeqClean (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/download.pl?ftp_dir=software_&file_dir=seqclean/seqclean.tar.gz). The 454 reads were assembled together with the Genbank ESTs with the NGen software (Lasergene) on a 64-bit desktop computer. For optimization purposes, assembly was tried at 85, 80 and 75% minimal match percentage (high, medium and moderate stringency). The final assembly was done using the following parameters: matchsize: 25, gap penalty: 7, mismatch penalty:12, match score: 10, minimal match percentage: 75, match spacing: 40. The quality of the assembly was assessed manually for 400 representative contigs by checking the correct alignment of the fragments.

Assembly quality regarding redundancy and gene representation was assessed by self-blast using a cut-off value of e^{-30} and by blastn comparison of the assemblies with mRNA sequences of *Solanales* (NCBI Genebank). Final annotation of the gene index was done by blastx against the 3.A_thaliana-p protein database (TAIR, www.arabidopsis.org). All blast analyses were run with BLAST 2.2.20 on the CIP High Performance Computer (http://hpc.cip.cgiar.org). Gene ontology attribution was performed using Blast2GO according to Conesa er at. (2005).

Marker development

Microsatellite (SSR) sequences were identified in the sequence assembly with the SSRlocator (http://minerva.ufpel.edu.br/~lmaia.faem/ssr1.html) limiting the hits to motives that consisted of at least 10 dimers, 7 trimers or 5 tetramers. Primers for SSR loci were designed with Primer3 with 100 – 200 bp amplicon size. SNPs were searched in 200 manually selected contigs that represent stress-sensitive genes using Seqman (Lasergene Inc.). Diversity Array Technology (DArT) marker development is in progress at DArT)/PL, as described by Wenzl et al. (2004).

Results and Discussion

A sweetpotato CGS consisting of 480 accessions was established. This set contains a broad range of diversity for agronomical and resistance traits, nutritional quality and breeding including clones with high beta-carotene, starch, iron and zinc contents, nematode, virus and drought resistance. All clones of the CGS are available as pathogen-free in vitro plants ready for international distribution. A complete list of the CGS is available at http://gcpcr.grinfo.net/files/cr_files/gcpcr_file832.xls.

To increase the available gene sequence information for functional genomics approaches on sweetpotato, we have produced two normalized cDNA libraries, one derived from leaves and one from stems of the sweetpotato variety Tanzania (CIP accession no. 440166). To increase the representation of stress-related genes, we have submitted the plants to drought stress before sampling. Both cDNA libraries were submitted to 454 pyrosequencing. Pyrosequencing has become a popular method for high throughput sequencing, as it provides a huge amount of sequence information at relative low error frequency and cost. It is widely used for genome re-sequencing (Bentley 2006), de novo sequencing of small genomes (e.g. Thomson et al., 2008), SNP-detection (Bundock et al. 2009), or transcriptome sequencing (Vera et al., 2008).

In total, we obtained 87.307 raw reads comprising 21.292.096 bases with a 454 FLX quarter run of a normalized cDNA library of leaves and further 436.817 raw reads consisting of 136.844.411 bases from another quarter run of 454 FLX TITANIUM. The average length per read amounted to 213.9 bp for the 454 FLX run and to 313.3 for the 454 FLX TITANIUM run. Short reads (<40 bp) and low quality sequences were eliminated and the remaining 523.914 reads were assembled together with 22.094 ESTs from the Genebank. Assembly was optimized by variation of the assembly parameters, until obtaining an assembly with maximal representativeness and minimal redundancy (Table. 1).

Parameter combination	Very stringent	Stringent	Moderately stringent
Total sequences	77629	65685	60245
Contigs	37593	33079	31165
Singletons	40036	32606	29080
Redundacy between contigs (self-blast matches)	13339	8804	6983
Unique blastn hits in *Solanales* database	3801	4337	4463

Table. 1 Result of hybrid assemblies of ~500.000 454 raw reads with ~22.000 ESTs derived from the genebank at different stringency levels. Lowering the assembly stringency decreased the redundancy and increased the representativeness of the assembly. The assembly parameters at different stringency levels are given in materials and methods.

The final assembly was performed with moderately stringent parameters as suggested by the relatively low redundancy and good gene representation. The assembly and annotation strategy is shown in Fig. 1. From in total 546.013 reads, 424.833 were assembled to 33.165 contigs, while 121.180 sequences remained unassembled. 29 080 unassembled sequences that were larger than 100 bp were considered as singletons. The mean length of the contigs was 787.6 bp and 8011 contigs comprised more than 1000 bp. The average coverage per contig was 14-fold and for the 8011 contigs larger than 1000 bp the mean coverage was 29.6-fold.

The sequence assembly was annotated by blast-x comparison with proteins of the manually reviewed UNIPROT *A. thaliana* protein database. 24.994 contigs and singletons had significant matches to protein sequences of the database, including 14.145 unique hits, while 37.251 remained without significant match. In the next step we have tried to attribute gene ontologies (GO, The Gene Ontology Consortium 2008), and could attribute cellular compartment, biological processes and molecular function to about half of the sequences represented in the assembly. The full information of the assembly is available at http://gcpcr.grinfo.net/index.php?app=datasets&inc=dataset_details&dataset_ id=712.

The relative high number of un-annotated sequences might be due to different causes. First, some of the un-annotated sequences could represent new genes that were not yet identified in other species. Second, these sequences could represent non-protein coding RNAs and therefore we failed to find similar sequences in protein

databases. A part of the un-annotated singletons however could result from contamination of the cDNA with genomic or chloroplast DNA. This issue is currently under investigation. A high number of sequences without significant hit in databases was found also in other 454 transcriptome sequencing projects (e.g. Vera et al., 2008). In spite of the lack of annotation for a significant part of the gene index, the large number of new gene sequences and the high rate of unique genes will strongly facilitate functional genomics approaches in sweetpotato.

The resulting sequence information has been used to design primers for new microsatellite (SSR) markers. Those SSR markers that were successfully amplified and yielded polymorphic bands in a test panel of eight sweetpotato accessions can be found at http://gcpcr.grinfo.net/index.php?app=datasets&inc=dataset_details&dataset_id=712. High heterozygosity of sweetpotato allowed for the mining of the sequence assembly for SNPs, although the majority of the sequences are derived from one sweetpotato clone only. We have identified SNPs in 200 sequences that correspond to stress response genes. These SNPs can be used as gene-based markers to tackle stress gene alleles in germplasm screening and crossing efforts. The SNP data also are available at http://gcpcr.grinfo.net/files/cr_files/gcpcr_file867.xls. DArT markers as well as a diploid reference map for a near relative of sweetpotato, *I. trifida*, are under development and will be available in 2010.

Acknowledgement

The work was supported by the Generation Challenge Program (//www.generationcp.org).

References

Bentley D.R. 2006. Whole-genome re-sequencing. Current Opinion in Genetics & Development 2006, 16:545–552.

Bundock P.C., Eliott F.G., Ablett G., Benson A.D., Casu R.E., Aitken K.S., Henry R.J. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploidy plant species using 454 sequencing. Plant Biotechnol J. 7, 347-54.

Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M., Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21,3674-3676.

Shagin D.A., Rebrikov D.V., Kozhemyako V.B., Altshuler I.M., Shcheglov A.S., Zhulidov P.A., Bogdanova E.A., Staroverov D.B., Rasskazov V.A., Lukyanov S. 2002. A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Res. 12,1935-1942.

The Gene Ontology Consortium. 2008. The Gene Ontology project in 2008. Nuc. acids res. 36, D440–4.

Thomson N.R., Holden M.T.G., Carder C., Lennard N., Lockey S.J., Marsh P., Skipp P., O'Connor C.D., Goodhead I., Norbertzcak H., Harris B., Ormond D., Rance R., Quail M.A., Parkhill J., Stephens R.S., Clarke I.N. *Chlamydia trachomatis*: Genome sequence analysis of lymphogranuloma venereum isolates. Genome Res. 2008. 18: 161-171.

Vera J.C., Wheat C.W., Fescemyer H.W., Frilander M.J., Crawford D.L., Hanski I., Marden J.H. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. 17, 1636–1647.

Wenzl, P., Carling J., Kudrna D., Jaccoud D., Huttner D., Kleinhofs A., Kilian A. 2004. Diversity arrays technology (DArT) for whole-genome profiling of barley. PNAS 101, 9915 9920.

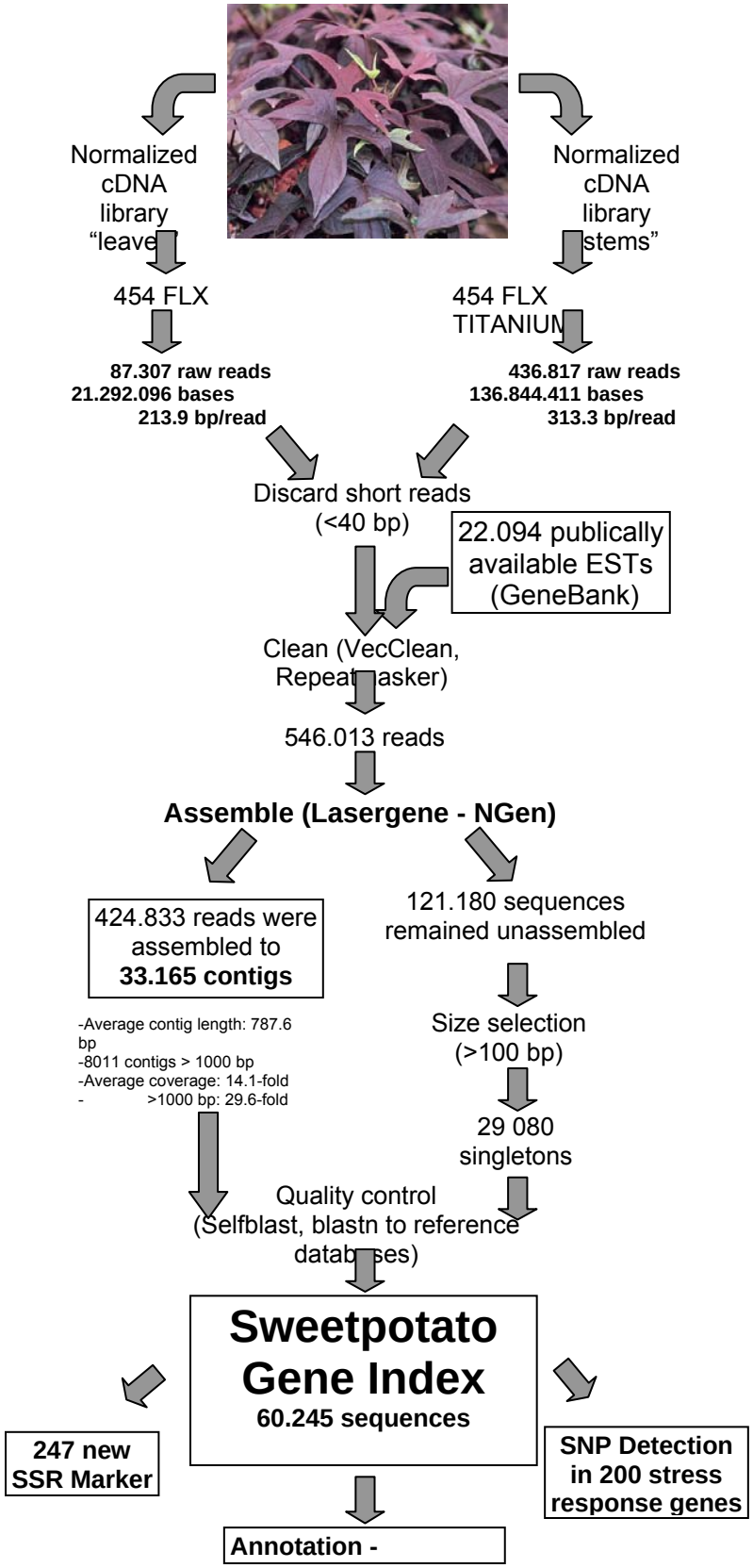

Fig.1: Sweetpotato transcriptome sequencing strategy. Details are given in the text.