WUR Plant Breeding
Attn. Rene Klein Lankhorst
Droevendaalsesteeg 1
6708 PB Wageningen

Date	Our ref.	Your ref.
August 13, 2009	ndr.WUR09.02R	

Subject: Project report 5.5H.LE.AC.0

Dear Rene,

Please find enclosed the report for project 5.5H.LE.AC.0: '10 X whole genome profiling of the tomato Heinz1706 genome'. This project has been executed as part of the 'Material transfer agreement among CBSG2012 partners' signed on 27-01-09 and 10-02-09 by Keygne and PRI respectively (Keygene reference L1174b). The current project report covers the projects 5.5H.LE.AC.0, 5.34.LE.AB.0 and 5.34.LE.AB.1, who are all part of this MTA.

If there are any questions, please do not hesitate to contact me.

Best regards,

Keygene Applied Research

Marco van Schriek

Enclosures: Project report
CD ROM

PROJECT REPORT

Date	:	August 13, 2009
Project code	:	5.5H.LE.AC.0
Species	:	*Solanum lycopersicum*
Goal	:	10 X Whole genome profiling of the tomato Heinz1706 genome
Number of BAC clones	:	92,160
Number of WGP tags	:	261,913

10 X WHOLE GENOME PROFILING OF THE TOMATO HEINZ1706 GENOME

1. INTRODUCTION

The objective of this project is to generate a whole genome profile map of the tomato Heinz1706 genome with a scope of 10x coverage (125 Kb BAC insert sizes assumed).

The resulting WGP map can serve as sequence-based scaffold for assembly of the tomato genome sequence by CBSG and partners.

2. MATERIALS

2.1 Material at project initiation

Upon initiation of the project, the material to be used was decided as follows;

2X Heinz1706 *Hind*III library (~15000 BAC clones at 125Kb average insert size)
2X Heinz1706 *Mbo*I library (~15000 BAC clones at 125Kb average insert size)
2X Heinz1706 *Eco*RI library (~15000 BAC clones at 125Kb average insert size)
4X Random Shear library (~30000 BAC clones at 125Kb average insert size)

The three enzyme libraries were directly available upon initation of the project. The random shear library would first be generated by Lucigen and subsequently used in the project (see also minutes of the meeting dd. Nov. 28, 2008).

2.2 Material alterations during project execution

During the course of the project the random shear library was evaluated and two modifications were executed to the original material plan.

1) It was decided by the WUR not to pool 4X equivalents of the random sheared library, but to extend this with 2X to a total of 6X coverage (assuming 125Kb average insert sizes). This additional work is executed in project 5.34.LE.AB.0
2) After quality control of the Random Shear library it was clear that the average BAC insert size is not in the 125 Kb range but rather in the 100 Kb range, reducing WGP efficiency and therefore additional BACs were added and analysed in the project. To be exact, the additional BACs described in topic 1 (previous paragraph) were also analysed using the WGP procedure. This additional work is executed in project 5.34.LE.AB.1 to reach the original aim of 10X coverage. The data available for the current project results in a BAC setup as described below

~2X Heinz1706 *Hind*III library (15360 BAC clones at 125Kb average insert size), named H_Hind_001_A01 for plate 1 well A1, etc.
~2X Heinz1706 *Mbo*I library (15360 BAC clones at 125Kb average insert size) named H_Mbo_001_A01 for plate 1 well A1, etc.
~2X Heinz1706 *Eco*RI library (15360 BAC clones at 125Kb average insert size) named H_Eco_001_A01 for plate 1 well A1, etc.
~4X Random Shear library (46080 BAC clones at 100Kb average insert size) named H_R-S_001_A01 for plate 1 well A1, etc.

Results reported here are all based on the full dataset as was the original scope of the project.

2.3 Material quality control

To ensure the identical material and material nomenclature as is presently used in the public domain a set of ~100 BACs were randomly selected and used for BAC-end sequencing. These BAC-end sequences were then used to validate identical BAC-end sequences and BAC nomenclature versus the public domain. This was executed for the BACs provided in Table 1 resulting in a confirmation of the BAC material and BAC nomenclature.

Table 1: BACs validated for material quality control

Eco_25_B02 F	Hind_03_D04_R	Hind_21_B02_F	Mbo_23_B02_F
Eco_26_D04_F	Hind_04_B02_R	Hind_21_B02_R	Mbo_23_D04_F
Eco_27_D04 R	Hind_04_D04_R	Hind_21_D04_F	Mbo_24_B02_F
Eco_28_B02 R	Hind_05_B02_R	Hind_21_D04_R	Mbo_25_D04_F
Eco_28_D04 R	Hind_08_B02_F	Hind_23_B02_F	Mbo_26_D04_F
Eco_28_D04_F	Hind_09_B02_R	Hind_23_B02_R	Mbo_27_D04_F
Eco_29_B02 R	Hind_09_D04_F	Mbo_13_D04_F	Mbo_28_D04_F
Eco_29_D04 R	Hind_09_D04_R	Mbo_14_B02_F	Mbo_30_D04_F
Eco_29_D04_F	Hind_10_D04_F	Mbo_14_D04_F	Mbo_32_D04_F
Eco_30_B02 F	Hind_11_B02_R	Mbo_15_B02_F	Mbo_32_D04_F
Eco_30_D04_F	Hind_13_B02_F	Mbo_15_D04_F	Mbo_33_D04_F
Hind_01_B02_F	Hind_13_B02_R	Mbo_16_B02_F	Mbo_33_D04_F
Hind_01_B02_R	Hind_13_D04_F	Mbo_17_D04_R	Mbo_35_D04_F
Hind_01_D04_R	Hind_15_B02_F	Mbo_18_D04_R	Mbo_36_D04_F
Hind_02_D04_F	Hind_15_B02_R	Mbo_20_B02_F	Mbo_36_D04_F
Hind_02_D04_R	Hind_17_B02_R	Mbo_21_D04_F	
Hind_03_D04_F	Hind_20_B02_R	Mbo_22_D04_F	

3. WGP PROCEDURE & RESULTS

3.1 WGP data production

The WGP data production process encompassed the following steps:

1) Pooling individual BAC clones in a 2-dimensional format
2) Isolation of high concentration, low *E. coli* level, pooled BAC DNA.
3) Digestion with restriction enzymes *Eco*RI/*Mse*I, ligation of Genome Analyzer (GA) adaptor sequences containing sample identification tags (a.k.a. "sample barcodes") and PCR amplification
4) Pooling PCR products
5) Cluster amplification
6) Sequencing using the GA II with 36 nt read length.

GA II sequencing resulted in a total of ~327 million high-quality sequence reads which were available for further processing.

3.2 WGP data processing

The high-quality GA II reads were used for WGP data processing, which included the following steps;

1) Deconvolution, i.e. assigning sequence reads to individual BACs in the 2-D pools;
2) Assigning WGP tags to BACs based on deconvoluted reads;
3) Filtering of WGP tags using various quality control measures to reduce noise to an absolute minimum
4) FPC contig map creation using the filtered data as input (paragraph 4)

Table 2: Overview general WGP input parameters and GA II sequence data Processing

WGP parameter	**Tomato**	Enzyme	Random Sheared
Genome size	950 Mbp		
WGP tag length incl. restriction site	26 nt		
# BACs tested	92,160	46,080	46,080
Genome equivalents BACs tested	10.9	6.1	4.8
Enzyme combination WGP fragments	*Eco*RI/*Mse*I		
# high-quality GA reads produced (M)	326.9		
# deconvolutable GA reads (M)	136.7		
% deconvolutable GA reads	42		
# unique WGP tags (FPC ready)	261,913		
# tagged BACs (FPC ready)	66,084	37,912	28,172
% tagged BACs (FPC ready)	72%		
average # WGP tags/ BAC	33.2	35.0	30.7
average # reads/ tag	50	46	54

3.3 FPC map assembly

Sequence-based physical BAC maps were assembled using an improved version of FPC software (Keygene N.V.), capable of processing sequence-based BAC fingerprint (WGP) data instead of fragment mobility information as used in the original FPC (C. Soderlund, I. Longden and R. Mott, 1997: FPC: a system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. 13: 523-535).

The filtered WGP data described in paragraph 3.2 were used as input in the FPC map assembly. The assembly was performed using a E^{-30} stringency in a "high stringency" map. A second map was created using the contigs derived from the high stringency map and allowing singletons to be added at a lower stringency of E^{-15} which is referred to as a "reduced stringency" map. The results of these two assemblies in terms of the numbers of BAC clones incorporated in contigs, broken down by BAC library, are shown in Table 3. Summary FPC results for the high- and reduced stringency WGP map assemblies are shown in Table 4 and further discussed in the summary & discussion paragraph.

Table 3: FPC map assembly results: high and reduced stringency WGP maps

Library	# BACs input FPC	# BACs in contigs high stringency map	% BACs in contigs high stringency map	# BACs in contigs reduced stringency map	% BACs in contigs reduced stringency map
Enzyme	37,912	30,464	80.4	33,891	89.4
Random sheared	28,172	22,153	78.6	25,167	89.4
Total	66,084	52,617	79.6	59,058	89.4

Table 4: FPC results for contig building of WGP tomato

	High stringency[1] WGP map assembly	Reduced stringency[1] WGP map assembly
total # of BACs in FPC	66,084	66,084
# of contigs	2,521	2,424
# BACs in contigs	52,617	59,058
# Singleton BACs	13,467	7,026
Coverage (Mbp)	953**	947**
Average contig size (BACs)[2]	21	24
N50 contig size (BACs)[3]	26	54
Average contig size (Mbp)[4]	0.378**	0.391**
N50 contig size (Mbp)[5]	0.563**	0.601**

1 Parameter setting: tolerance = 0; cut-off = 1.0E-30 for high stringency map and cut-off is 1.0E-15 for the reduced stringency map.
2 This is the mean number of BACs per contig.
3 This number indicates that more than 50% of the contig coverage comprises contigs with at least this number of BACs
4 This number is the mean contig size in million basepairs
5 This number indicates that more than 50% of the contig coverage comprises at least this number of million basepairs
** *It is noted that these figures are based upon the multiplication of the FPC band units and the average distance between two tags. This multiplication, executed for all contigs, results in a large spread. Therefore these figures should be handled with caution.*

3.4 QC FPC maps

The quality of the resulting FPC map was verified using the current dataset and public domain data. Previously, a tomato Heinz FPC map was generated (mainly) using the *Hind*III library as input which was the basis of BAC-by-BAC sequencing approaches.
To verify the FPC output, the following checks have been performed:

1. BLAST hit overlap between BACs in a FPC contig.
2. Correlation between public domain (SGN) contigs *vs.* Keygene physical map contigs.

Ad1)
For the BLAST check, sequence tags were BLASTed against public data such as ESTs, Full BAC clones and Unigenes. This data was linked to individual BAC clones and FPC clusters respectively. Overlap between multiple BACs within a contig was found confirming the correct placement of these BACs in a contig.

Ad2)
The correlation with the SGN physical map was investigated by counting the number of shared BACs per correlated contig pair. Table 5 summarizes the results obtained for the "high stringency" and "reduced stringency" maps.

Table 5: Comparison of the "high stringency" and "reduced stringency" FPC Physical maps against SGN Physical map.

	High stringency WGP map assembly	Reduced stringency WGP map assembly
Correlated contigs [a.]	82 %	82 %
Correlated BACs in contigs [b.]	42 %	40 %
1:1 Correspondence [c.]	16 %	14 %
Corresponding SGN contigs [d.]	2.6	2.8

Where:

a. Correlated contigs: how many (in %) of the FPC contigs have a given BAC overlap with SGN contigs.
b. Correlated BACs in contigs: how many (in %) of the BACs are identical in the overlapped contigs.
c. 1:1 Correspondence: how many (in %) of the FPC contigs correspond to one single SGN contig.
d. Corresponding SGN contigs: how many (weighted average) SGN contigs corresponds to one FPC contig.

4. SUMMARY & DISCUSSION

Using four different BAC libraries, all derived from the Heinz1706 line, 92,160 BACs (~11X) were used for WGP analysis. The WGP analysis resulted in a high stringency map in a total of 2521 contigs with an N50 contig size larger than 563 Kbp. The high stringency map places 52,617 BACs in contigs resulting in an ~6X effective coverage.

Both the high stringency and reduced stringency have a calculated genome coverage about the estimated genome size (950 Mbp). This figure indicates that the assembly is executed at a high stringency resulting in braking up a single contigs into two or more contigs and thereby

effectively increasing the map size. We feel the high stringency map is a good starting point when having genome wide assemblies in mind.
Next also a reduced stringency assembly has been calculated allowing singletons to be added to the high stringency contigs. This to provide the maximum number of BACs for a given region.
Lastly specific sets can be created using the FPC tool delivered with this project report.

5. DELIVERABLES

The following deliverables are included with this report on CD:

1) sequences of all filtered WGP tags in FASTA format (file named: *55HLEAC0 - tomato.fasta*)
2) filtered WGP tags linked to BAC identifiers in Access database format (file named *55HLEAC0 - tomato.mdb*).
3)FPC input files named: *55HLEAC0 - tomato.bands*
4) High- and Reduced stringency WGP physical maps in FPC output format (files named *55HLEAC0 – Tomato High stringency map.fpc* and *55HLEAC0 – Tomato Reduced stringency map.fpc*)
5) WGP compatible FPC version in LINUX executable (binary) format named (*fpc_bin_20090804.tgz* and corresponding source code named *fpc_src_20090804.tgz*)
6) This project report in .pdf format.

Marco van Schriek