Pyrosequencing/Sanger Plant Genome Assembly (Limitations, Problems And Solutions) - On The Way To Cucumber (Cucumis sativus L. cv. Borszczagowski) Draft Genome Sequence Publishing

Rafal Woycicki*, Zbigniew Przybecki

Dept. of Plant Genetics, Breeding & Biotechnology, Faculty of Horticulture and Landscape Architecture,
Warsaw University of Life Sciences - SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland

Motivation.
Cucumber with the genome size of 367 Mbp is an edible fruit and the plant belongs to the family Cucurbitaceae. The genus *Cucumis* contains more than 30 species, of which two are economically important crops, namely cucumber (*C. sativus* L) and melon (*C. melo* L). As a vegetable crop cucumber has big agriculture value, however the economic importance varies with the region It is a model plant in studying sex determination and a lot of molecular marker maps with the help of BAC libraries and STCs should work fine to achieve this aim by positional cloning (1).
Genome sequence give the opportunity to make all the range and better quality of omics studies. New sequencing technologies make it possible to achieve the sequencing reads fast and cheap. Since the assembly step of such next generation reads is still not well standardized it is the most cumbersome part of sequencing projects.

Methods.
More then 15 millions sequences used in the hybrid assembly came from genome coverage of 8x unpaired and 4x paired pyrosequenced 454 XLR Titanium reads, together with 64022 BAC library ends fragments (2,3), both sequenced at Agencourt Bioscience Corporation. The fragment reads were charqacterized using sffinfo script coming from Newbler, and sff_extract 0.2.3 (4) script The hybrid assembly were done using different approaches including Newbler ver. 2.0.00.20, Celera Assembler v 5.4, Arachne v. 3.0, but the final result were done using the last two programs. Conting assembled by Celera were cutted into overlapping by 100 nt fragments of 800 nt length and introduced to Arachne together with BAC ends. Additionaly information about Titanium supercontings were added with artificially prepaired paired ends (foreword and reverse). The assembly were done on 64bit workstation with 8 SSP and 32 GB RAM and took maxiumum 36h.

Table 2. Summary of assembly steps.

Program used	Newbler	Newbler + Arachne	Celera	Celera	Newbler	Newbler + Arachne	Celera	Celera	Celera + Arachne
Reads type	8x UN Titanium	8x UN Titanium + BAC ends	8x Titanium	8x UN Titanium + BAC ends	8x UN + 4x PA Titanium	8x UN, 4x PA + BAC ends	8x UN + 4x PA Titanium	8x UN, 4x PA + BAC ends	8x UN, 4x PA + BAC ends
Titanium reads used	7423671	NA	6418446	6418446	13404638	NA	11125634	11125634	NA
BAC ends reads used	NA	62750	NA	62750	NA	62750	NA	62750	64022
Fake reads used	NA	215663	NA	NA	NA	176305	NA	NA	285037
Fake mate paired read used	NA	NA	NA	NA	NA	NA	NA	NA	82496
No. Contigs	57783	40294	35613	34753	225309	46487	15514	15667	16403
Contigs length sum (Mbp)	190	177	183	189	192	110	196	197	193
% genome size in contigs	51.77%	48.23%	49.86%	51.39%	52.32%	29.97%	53.30%	53.79%	52.59%
No. Contigs >1000 nt	31768	37102	35612	34753	62846	43055	15222	15196	15686
Mean contig length	5656	4708	5155	5428	2107	2496	12839	12972	12264
N50 contig length	8361	6206	7635	8281	2288	2601	24714	27086	23280
No. Supercontigs	NA	39071	NA	33495	12087	46256	4976	4173	12944
Supercontigs span sum	NA	278	NA	300	126	126	198	224	321
% genome size in supercontigs	NA	75.61%	NA	81.74%	34.33%	34.33%	53.84%	61.14%	87.47%
No. Supercontigs >1000 nt	NA	35894	NA	33495	12087	42839	4967	4157	12310
Mean supercontig span	NA	7664	NA	8948	10473	2873	39799	54070	26046
N50 supercontig span	NA	11992	NA	19122	16537	2853	158310	2324038	372416

Results
Average length of 7'970'914 unpaired reads was 374 nt. Reads coming from sequencing of paired Titanium library consisted of 3'204'606 paired reads (containing linker) with the length of 171 nt, 892328 unpaired reads (multi or partial linker) with the length of 195 nt and 3106927 unpaired reads (no linker) with the length of 227 nt. Lengths of pyrosequenced fragments are showed on Fig. 1 i 2 and Table 1. Average length of 64022 BAC end sequences was 737 nt. Summary of assembly results is shown in Table 2. Below you can find detailed information about finally chosen method. Titanium 8x unpaired & 4x paired reads were assembled using Celera. From 11125634 (73.31% of all reads) 8506533 were assembler into 15514 contigs of 195.6 Mbp length with the average length of 12839 nt and N50 of 24714 nt. Number of supercontigs was 4976 of coverage 197. 6. Average coverage of supercontigs was 39799 nt and N50 was 158310 nt. Number of reads recognized to be repeated sequences and not uniqally assembled was 2338980. From 2823642 paired Titainium reads recognized by Celera (88.11% of recognized by sff_extract), 1610390 (57.03%) were used in the assembly and only 1298323 (45.98%) were assembled into contigs. Beside resulting contigs and supercontigs, Celera had made also 160714 degenerate contigs of the total length of 76.5 Mbp which consisted of partially assembled repeated sequences which could not be uniqally assembled into contigs. The degenerate contigs were not used in the next step of hybrid assembly. Hybrid assembly were finished using Arachne assembler. The assembler used 285037 unpaired pseudo-reads generated from 15514 Celera contigs together with 64022 BAC ends and 82496 artificial paired reads containing information about Titanium Celera supercontigs. Number of generated contigs were 16547 of the length 193.2 Mbp, average length was 12214 nt and N50 23280 nt. Number of supercontigs were 13129 with the coverage of 323 Mbp. The average coverage of supercontigs was 25865.9 nt and N50 was 323092 nt. Corectness of the assembly were propen after mapping 95,56% of 63035 cucumber unigenes (5) with the homology higher then 95%.
As shown in Table 2, assembly of Titanium and BAC end reads with Celera resulted in increasing of supercontigs coverage only by 27 Mbp from 197 Mbp to 224 Mbp in compare to 323 Mbp achieved by Arachne. This could be the result of the fact that on the sequencing contigs Celera mapped only 42774 of used BAC ends (66.54%), the rest of them were in degenerate contigs. On the contrary number of BAC ends mapped on sequencing contigs by Arachne was 52524 (82.04%).

Summary
The above results show that its now possibile to sequence and correctly denovo assembler practically every genome. Advanced draft assembly of highly repetitive plant genome were achieved after hybrid approach with the use of pyrosequencing 454 XLR Titanium 12x coverage paired and unpaired reads and BAC end sequences. The coverage used came to be optimall for such a genome.

Table 1. Titanium reads statistics

	4x unpaired partial/multi linker	4x paired full linker	4x unpaired no linker	8x UN	Summary
All reads No.	892 328	3 204 606	3 106 927	7 970 914	15 174 775
<100 nt	28.05%	33.67%	25.62%	6.49%	
100-200 nt	26.89%	28.46%	25.03%	9.45%	
200-300 nt	22.12%	21.09%	18.03%	10.59%	
300-400 nt	16.90%	13.01%	14.32%	18.80%	
400-500 nt	5.85%	3.65%	13.87%	36.05%	
>500 nt	0.19%	0.12%	3.04%	18.25%	
Av. length	195.45	171.53	227.07	374.00	290.66
Total length	174 409 126	549 690 047	705 481 264	2 981 159 897	4 410 740 334
Genome coverage	0.48	1.50	1.92	8.12	12.02

Figure 1. Titanium reads lengths (with linker)

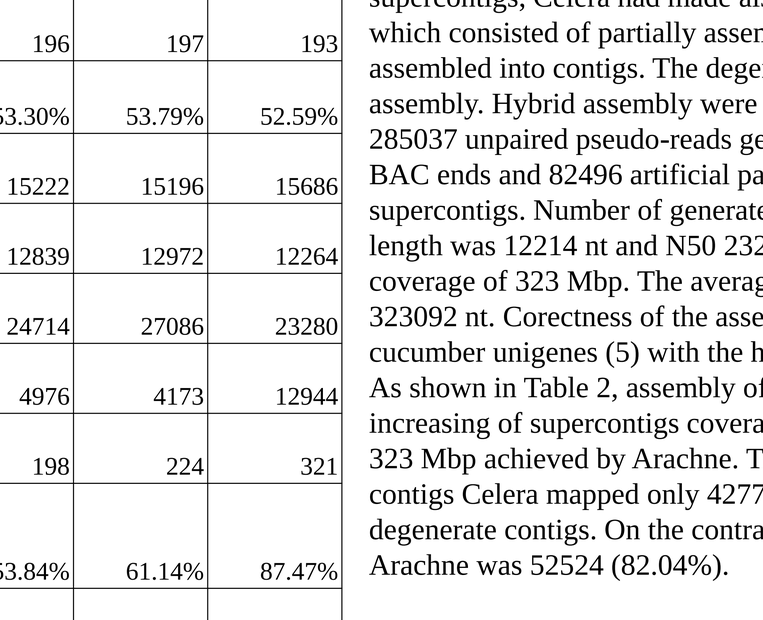
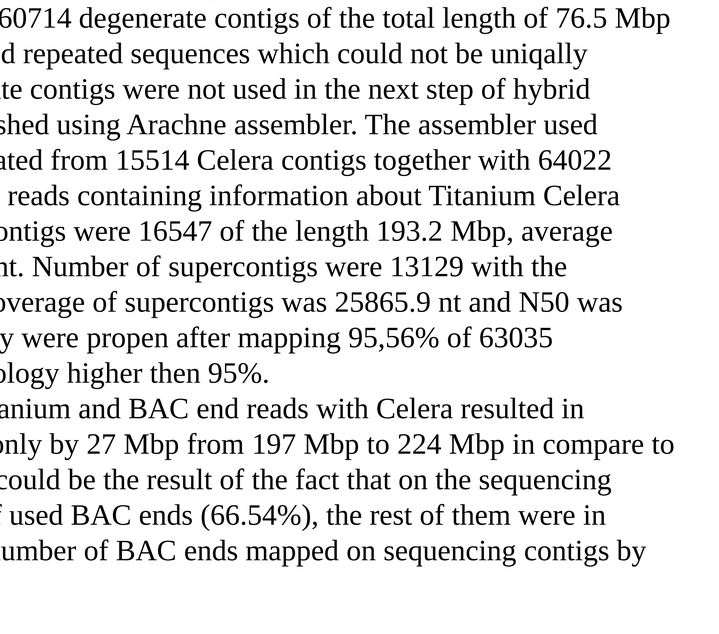


Figure 2. Titanium reads lengths after linker extraction



Literature:
1. Malepszy S., K. Niemirowicz-Szczytt, 1991. Sex determination in cucumber (*Cucumis sativus* L.) as a model system for molecular biology. Plant Science 80: 39-47
2. Gutman W, Pawełkowicz M, Woycicki R, Piszczek E, Przybecki Z., 2008, The construction and characteristics of a BAC library for Cucumis sativus L. 'B10'. Cell Mol Biol Lett 13/1: 74-91
3. Amplicon Express, Pullman, WA, USA
4. Blanca J., Chevreux B, http://bioinf.comav.upv.es/sff_extract/index.html
5. ICuGI, http://www.icugi.org/cgi-bin/ICuGI/EST/home.cgi?organism=cucumber

*email: rafal_woycicki@sggw.pl