Release note (Dec. 2, 2015)

Vigna angularis v1.0 (Vangularis_v1)

Overview

This is the release of Vangularis_v1, the first chromosome scale genome assembly of *Vigna angularis* var. *angularis* cultivar Shumari (azuki bean). This release was derived by assembling ~51x PacBio reads followed by correcting indel errors using Illumina short reads. Assembled contigs were scaffolded using Illumina mate-pair reads and then ordered and oriented using genetic map derived by genotyping the F2 population of the azuki bean and its wild relative, *Vigna nepalensis*. Gene loci were predicted by a combination of RNA-Seq based and *ab initio* predictions. Genome sequencing and annotation were carried out with the collaboration of National Institute of Agrobiological Sciences, Japan and Okinawa Institute of Advanced Sciences, Japan. Genome assembly and annotation have been deposited in DNA Data Bank of Japan (DDBJ) under accessions AP015034 -AP017294.

Assembly statistics

Genome

Estimated genome size: 540Mbp

Genome coverage (pseudomolecules/pseudomolecules + unanchored scaffolds): 87.3% / 96.8%

	No. of scaffolds	Total scaffolds length (bp)	Pseudomolecule length (bp)	Gap rate (%)	GC rate (%)
Anchored scaffolds					
Chr01	35	65,919,377	67,115,795	1.8	33.1
Chr02	28	45,096,540	45,515,668	1.0	33.6
Chr03	27	41,962,598	43,462,759	3.5	34.0
Chr04	25	53,470,133	54,073,736	1.2	33.4
Chr05	25	36,897,451	37,380,367	1.4	34.1
Chr06	23	38,789,640	38,860,970	0.3	34.1
Chr07	20	32,020,224	33,495,452	4.4	33.7
Chr08	32	46,426,616	46,951,433	1.2	33.3
Chr09	26	35,726,726	37,388,052	4.5	34.0
Chr10	17	28,785,286	28,886,040	0.4	33.7
Chr11	21	37,398,860	38,115,440	1.9	34.1
Total	279	462,493,451	471,245,712	1.9	33.7
Unanchored scaffolds	2,250	51,515,385	-	0.6	42.5

Loci (Vangularis_v1.a1)

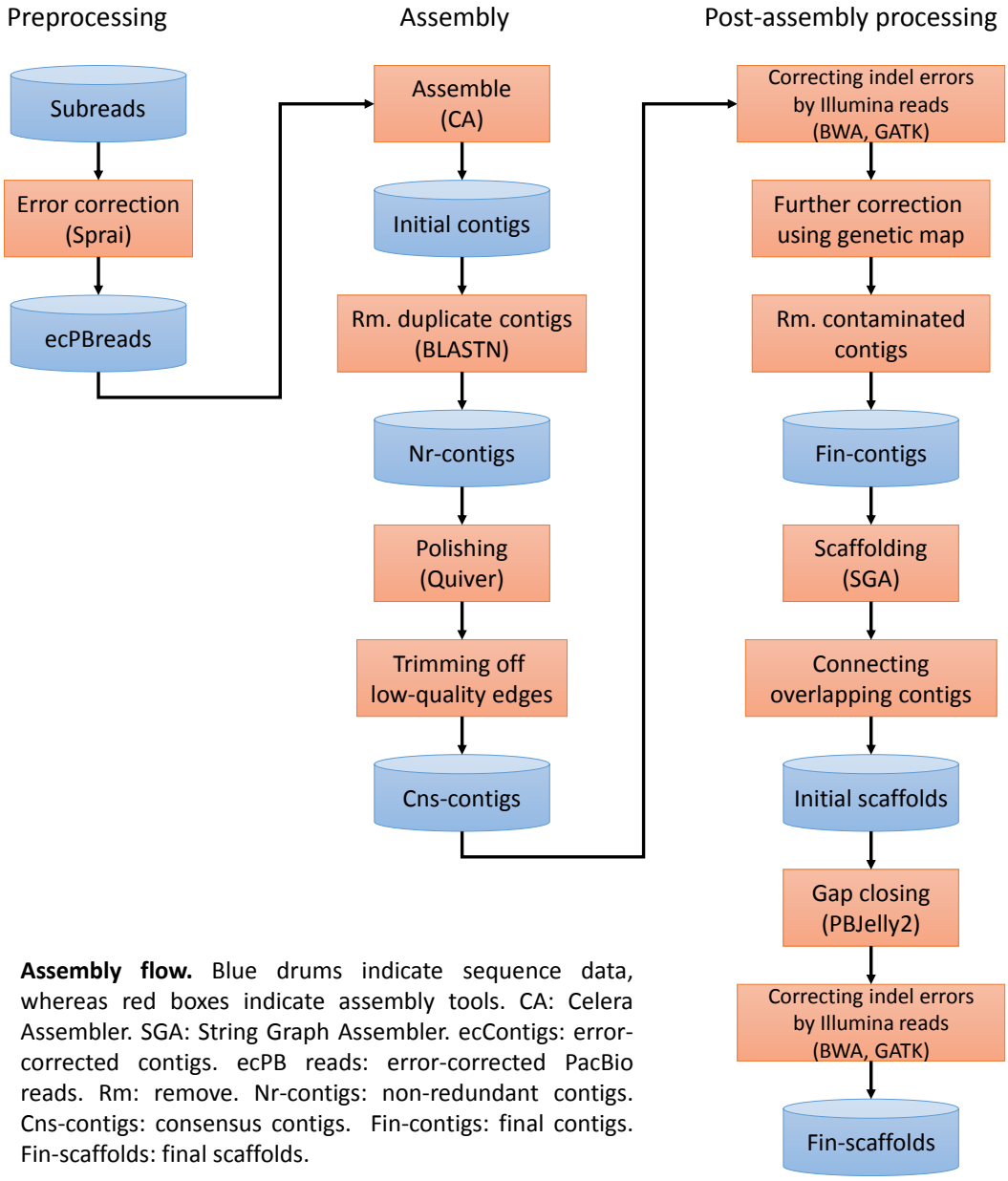
Chr.	No. of protein-coding genes	No. of non-protein coding genes	Total
Chr01	4,840	355	5,195
Chr02	3,002	215	3,217
Chr03	2,492	205	2,697
Chr04	3,785	290	4,075
Chr05	2,422	188	2,610
Chr06	2,282	194	2,476
Chr07	2,088	197	2,285
Chr08	3,229	260	3,489
Chr09	1,961	178	2,139
Chr10	2,189	164	2,353
Chr11	2,147	208	2,355
UM	804	40	844
Total	31,241	2,494	33,735

Sequence source

DNA/ RNA	Platform	Library	Insert size	Read length (mean)	# bases (bp)	Tissue
DNA	Illumina	Paired-end	270bp	151	45,537,536,288	Unexpanded leaves
		Mate-pair	3kb	30-100 (91.0)	16,345,960,620	Unexpanded leaves
			8kb	30-100 (90.9)	19,553,488,430	Unexpanded leaves
			20kb	30-100 (93.3)	12,237,150,417	Unexpanded leaves
			40kb	30-100 (92.8)	6,364,847,411	Unexpanded leaves
	PacBio	P4-C2	7kb	50-23,050 (2,761.8)	3,685,978,374	Unexpanded leaves
		P5-C3	7kb	35-39,432 (5,441.7)	27,641,829,986	Unexpanded leaves
RNA	Illumina	Paired-end	300bp	100	12,095,557,800	Cotyledon
			300bp	100	13,083,777,000	Axes
			300bp	100	8,753,977,000	Flower
			300bp	100	8,684,245,000	Leaf
			300bp	100	8,427,515,800	Nodule
			300bp	100	12,648,634,000	Pod

	300bp	100	12,382,545,800	Root
	300bp	100	10,367,533,600	Stem

Assembly method



Assembly flow. Blue drums indicate sequence data, whereas red boxes indicate assembly tools. CA: Celera Assembler. SGA: String Graph Assembler. ecContigs: error-corrected contigs. ecPB reads: error-corrected PacBio reads. Rm: remove. Nr-contigs: non-redundant contigs. Cns-contigs: consensus contigs. Fin-contigs: final contigs. Fin-scaffolds: final scaffolds.

Gene prediction

Gene structures were predicted by mapping RNA-Seq reads and *de novo* assembled contigs using TopHat, Cufflinks, Trinity, and PASA pipeline. Open reading frames (ORFs) were predicted using TransDecoder and Trinotate. A*b initio* gene prediction was also carried out using MEGANTE with the parameters of *V. unguiculata*. Genes with greater than 40% of their transcribed and exonic regions masked as repeat sequences were discarded. Gene structures predicted by TopHat and Cufflinks, and the PASA pipeline were clustered on the genome assembly, and the representative structure with the longest ORF was selected for each locus. Then, MEGANTE-predicted genes in unannotated regions were added. To screen transposon-related genes, BLASTP searches against UniProt and RefSeq were conducted and genes matching any transposon-related proteins with an E value

<1.0e-10 were discarded. Additionally, genes containing transposon-related functional domains were removed using InterProScan. The remaining genes were subjected to expression validation analysis with the RNA-Seq data. RNA-Seq reads from the eight libraries were mapped separately using TopHat. The expression level was estimated by Cuffquant for each locus and then normalized by Cuffnorm. A gene was discarded if its expression level was zero FPKM in all eight libraries and it did not have any homologs in the BLASTP result. The remaining genes were selected as the final gene set.

Reference

Sakai, H., Naito, K., et al., The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome., *Scientific Reports* 5, 16780; doi: 10.1038/srep16780 (2015)