Common Bean Genome Project

Scott Jackson, UGA
Jeremy Schmutz, JGI & HA
Perry Cregan, USDA-Beltsville
Phil McClean, NDSU
Dan Rokhsar, JGI

United States Department of Agriculture

National Institute of Food and Agriculture

This project was supported by the Agriculture and Food Research Initiative Competitive from the USDA National Institute of Food and Agriculture.

Goals

- Produce a reference quality genome for non-repetitive portion of *Phaseolus* using primarily 454 based sequence
- Scaffold scale contiguity with BAC end sequences and fosmid end sequences
- Construct chromosome scale contiguity with dense genetic map
- High quality annotation using RNA-seq to bolster small amount of transcript sequences

Sequencing for the V0.9 Genome

- 5 linear 454 libraries (281.6 bp ave. HQ) and 10 paired 454 recombination libraries (123.1 bp ave. HQ)

Library	Coverage (x)	Average Insert Size
Linear	17.04	NA
GPNB	0.14	4,806
GGAS	0.54	6,505
GXSF	0.11	7,733
HYFB	0.23	7,962
HYFA	0.34	7,995
HYFC	0.31	8,085
HXTI	0.32	11,847
GXNX	0.22	12,950
HXWF	0.17	16,786
HXWH	0.10	17,087
TOTAL	**19.50x**	

Results of the V0.9 Genome

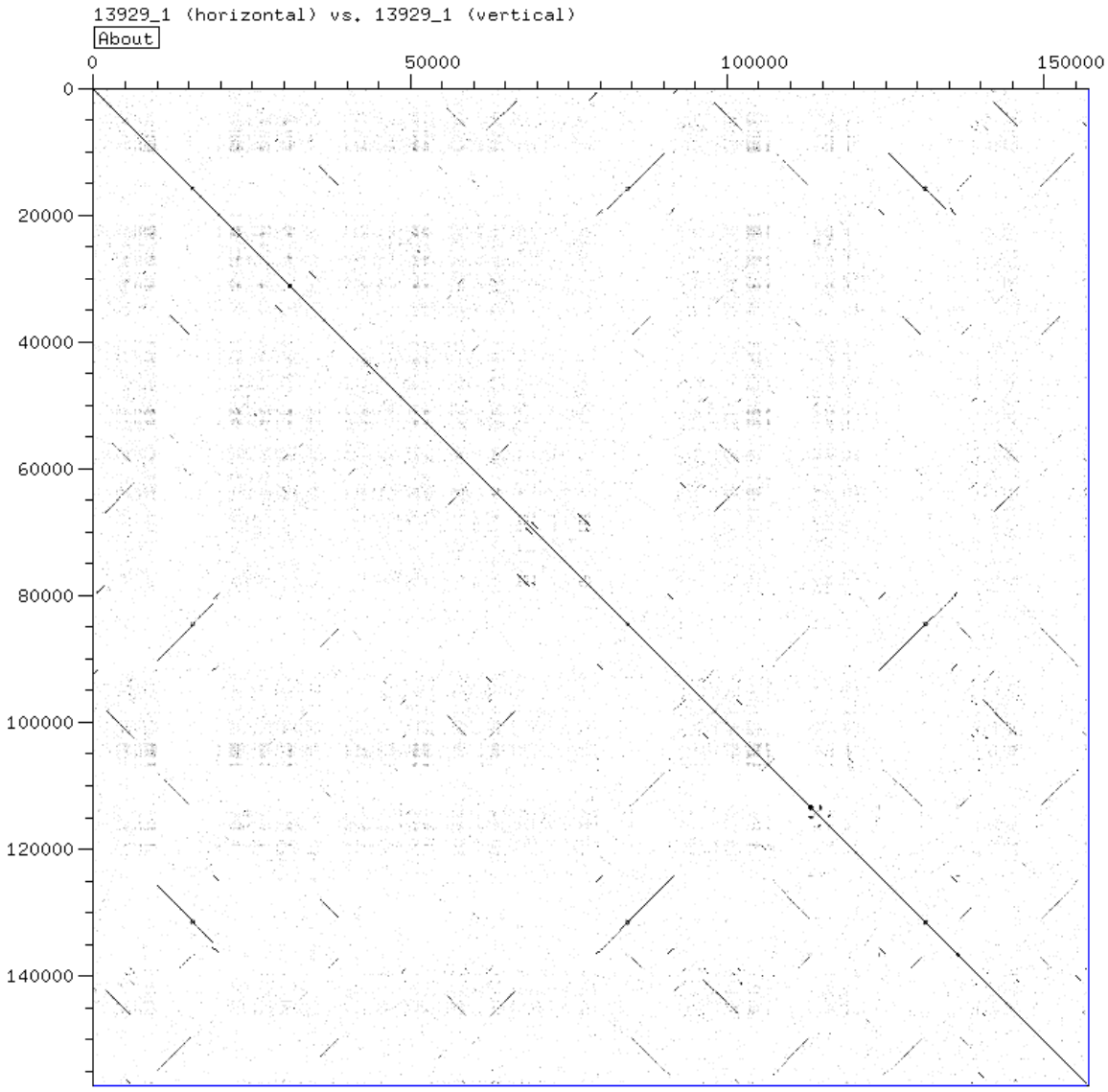
- Very short contigs N50=15.8 kb
- Short scaffolds N50=391.3 kb
- Lots of contigs and scaffolds

```
Main genome scaffold total: 10,132
Main genome contig total:   46,828
Main genome scaffold sequence total: 486.9 MB
Main genome contig sequence total:   430.4 MB (-> 11.6% gap)
Main genome scaffold N/L50: 279/391.3 KB
Main genome contig N/L50:   7,457/15.8 KB
Number of scaffolds > 50 KB: 1,601
% main genome in scaffolds > 50 KB: 87.4%
```

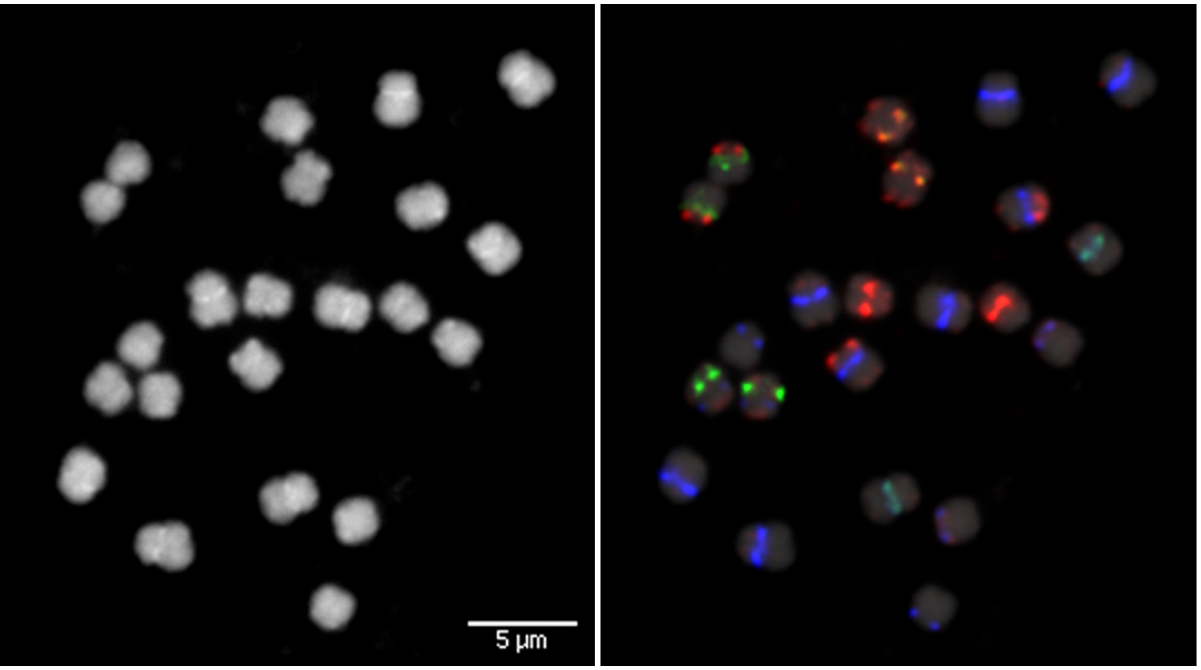
Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Total Scaffold Length	Total Contig Length	Scaffold Contig Coverage
All	10,132	46,828	486,869,582	430,369,104	88.40%
1 kb	10,132	46,828	486,869,582	430,369,104	88.40%
2.5 kb	6,680	43,376	479,209,664	422,709,247	88.21%
5 kb	4,016	40,684	470,999,681	414,550,589	88.02%
10 kb	3,138	39,188	465,025,927	409,783,158	88.12%
25 kb	2,305	36,809	451,343,582	398,976,275	88.40%
50 kb	1,601	33,182	425,717,052	379,259,827	89.09%
100 kb	986	28,047	382,245,716	344,554,929	90.14%
250 kb	454	20,301	299,185,806	273,635,738	91.46%
500 kb	198	13,021	207,947,599	193,037,241	92.83%
1 mb	66	6,757	116,138,930	108,820,797	93.70%
2.5 mb	10	1,730	31,586,117	29,910,282	94.69%
5 mb	0	0	0	0	0.00%

Why are there so many contigs and scaffolds?

- Shorter 454 reads give us less assembled repeat content
- *Phaseolus* has a lot of repeat content to assemble!
- Finished clones show significant ancestral and recent transposon activity

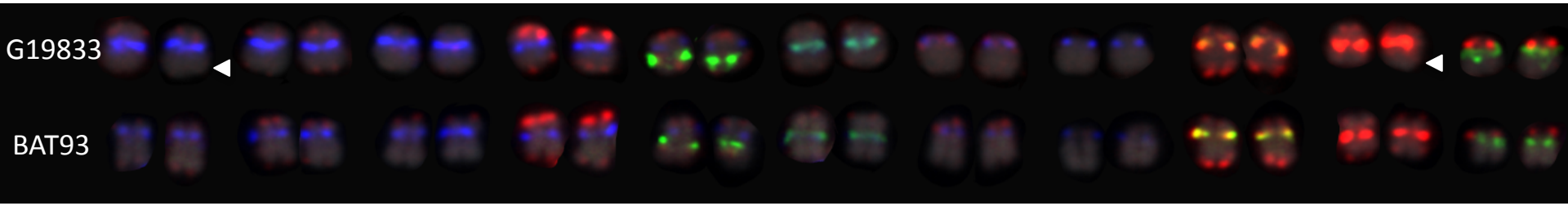


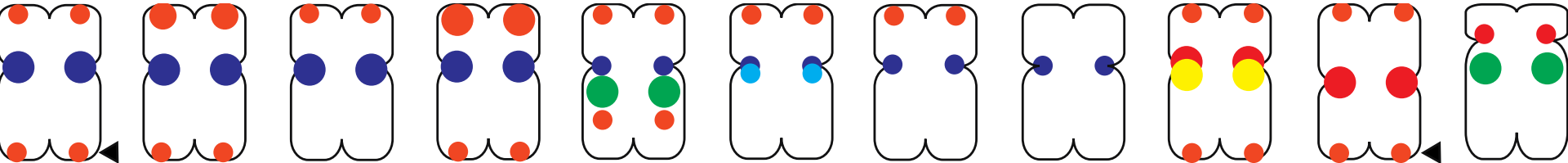
Structural differences due to repeats



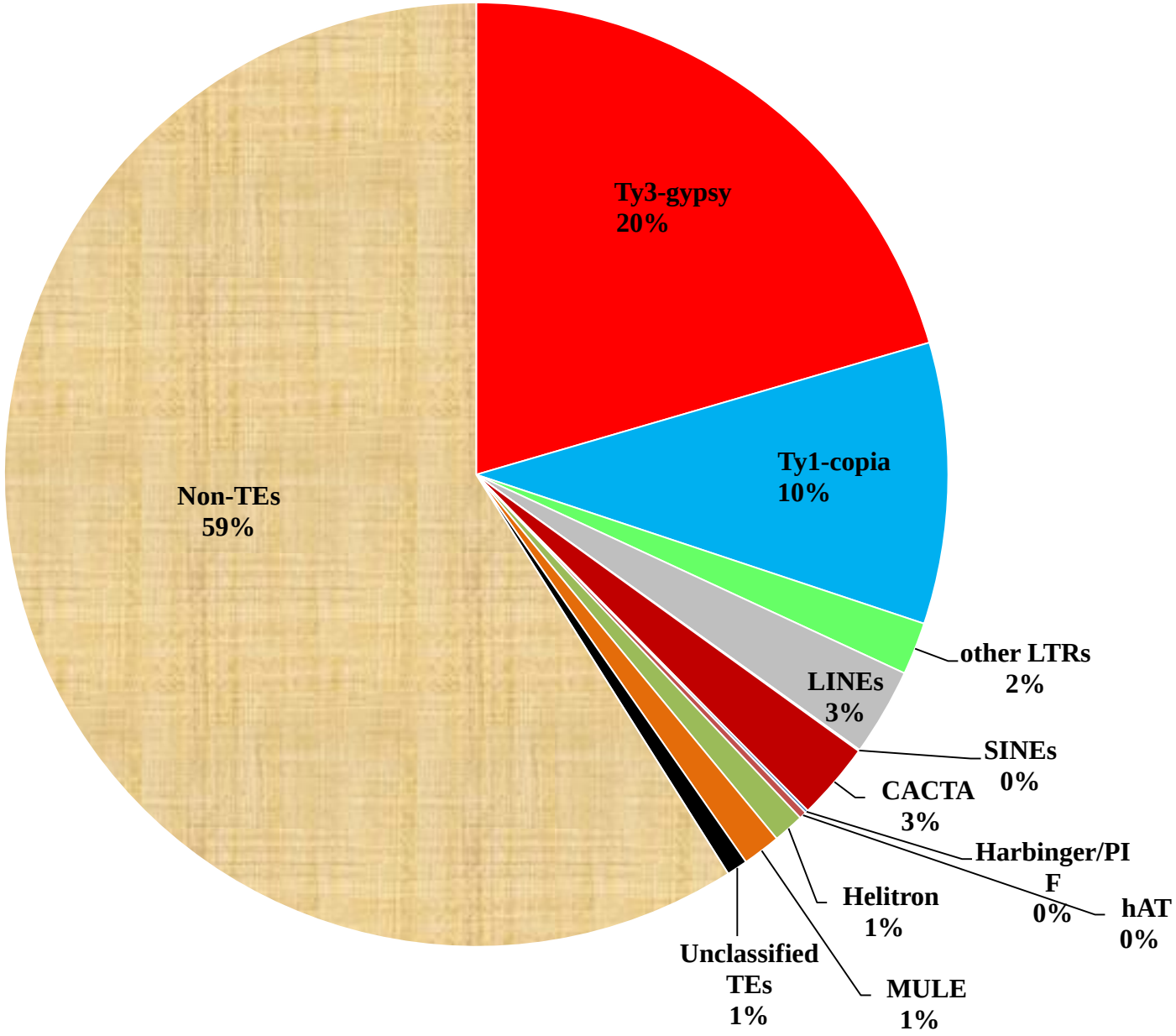
Probe cocktail

- CB100 CB100A
- CB110 CB111
- Subtelomeric repeat, *khipu*
- 5S rDNA





Transposon Content of V0.9



Annotation results for V0.9

- **Loci**
  - 26,374 total loci containing protein-coding transcripts
- **Alternative Transcripts**
  - 4,347 total alternatively spliced transcripts
- **For primary transcripts**:
  - Average number of exons 5.6
  - Median exon length 160
  - Median intron length 202
  - Number of complete genes 25,457
  - Number of incomplete gene with start codon 279
  - Number of incomplete gene with stop codon 597

Additional Data for V1.0

- 3 BES and 2 fosmid end sequence libraries
- 6 linear long read 454 runs from two libraries (4.5x coverage, 417.9 HQ bps average)
- Paired V3 Illumina from 2 libraries (2x100, 136 GBs)

Library	Reads	Total Bases (MB)	Mean Length (bp)	Repeat Content	Mean V1.0 Insert Size
PVA	89,017	62.5	781	12%	126,959
PVB	92,160	94.1	1,006	19%	135,292
PVC	81,408	83.3	1,017	22%	121,960
VUL	88,320	83.7	936	28%	34,956
VUK	240,384	242.7	995	28%	36,001

Assembly of V1.0

- De-replication of 454 pairs and organelle identification
- Pre-correction of 454 data for insertion/deletion errors using V3 Illumina reads
- Assembly with modified Arachne2
- Removal and cleaning of small contigs
- Chromosome construction, genetic map integration
- Post correction of remaining consensus errors

Anticipated V1.0 Result

465.7 MB incorporated in 11 chromosomes, 98.7% of total bases.

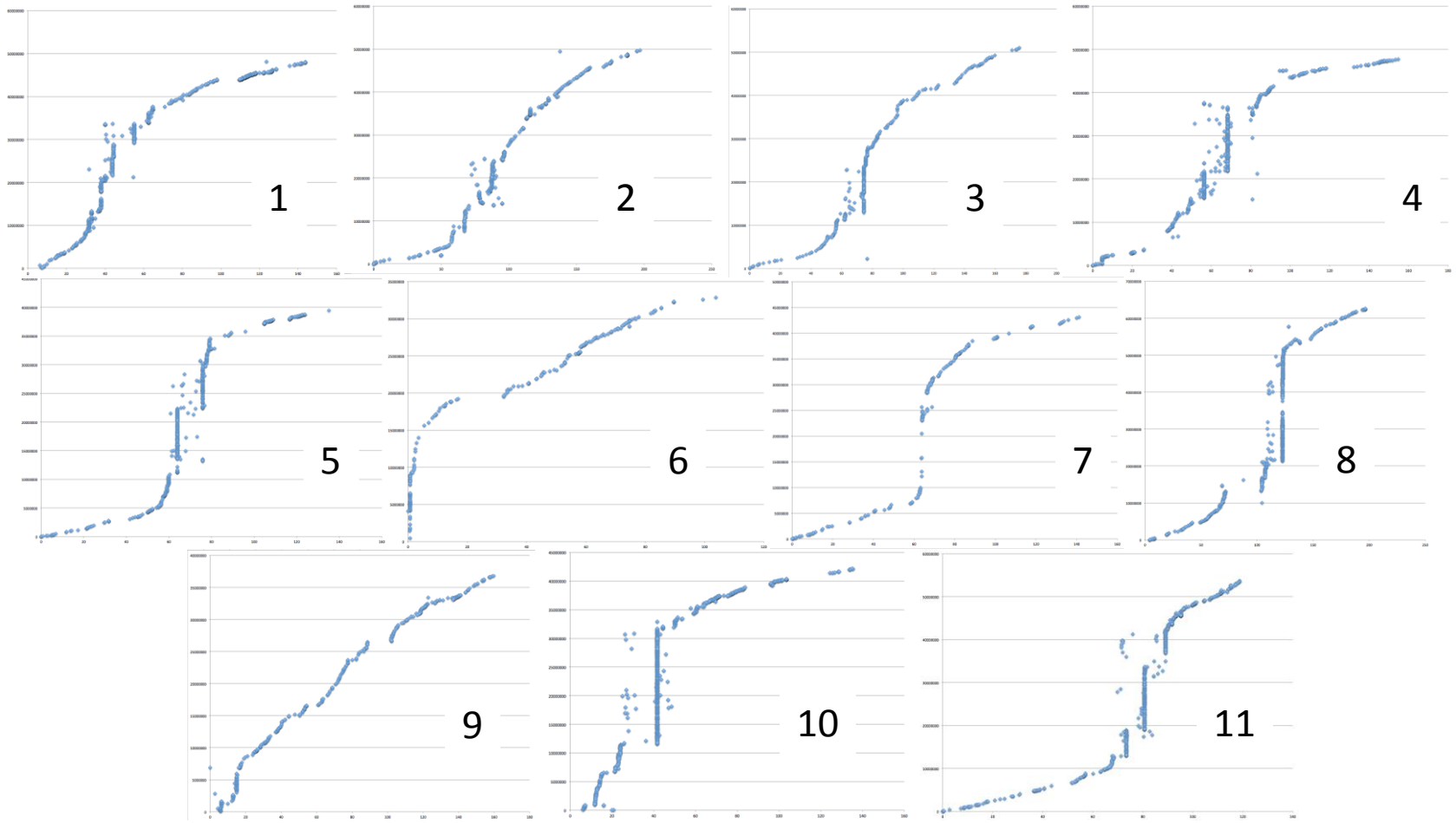
Contig N50 = 38.9 Kb

12 Mb in scaffolds that couldn't be orientated

```
Main genome scaffold total: 1422
Main genome contig total:   42364
Main genome scaffold sequence total: 520.3 MB
Main genome contig sequence total:   471.8 MB (->  9.3% gap)
Main genome scaffold N/L50: 5/48.2 MB
Main genome contig N/L50:   3316/38.9 KB
Number of scaffolds > 50 KB: 39
% main genome in scaffolds > 50 KB: 99.1%
```

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Total Scaffold Length	Total Contig Length	Scaffold Contig Coverage
All	1,422	42,364	520,276,899	471,833,293	90.69%
1 kb	895	41,837	519,884,055	471,440,449	90.68%
2.5 kb	457	41,390	519,244,367	470,803,692	90.67%
5 kb	313	41,164	518,703,264	470,355,875	90.68%
10 kb	141	40,798	517,565,106	469,448,059	90.70%
25 kb	70	40,500	516,574,630	468,661,103	90.72%
50 kb	39	40,267	515,477,007	468,097,278	90.81%
100 kb	25	40,137	514,537,663	467,480,431	90.85%
250 kb	17	39,931	512,970,929	466,823,772	91.00%
500 kb	15	39,812	512,279,041	466,638,115	91.09%
1 mb	12	39,696	510,348,853	464,817,263	91.08%
2.5 mb	12	39,696	510,348,853	464,817,263	91.08%
5 mb	11	39,445	505,901,054	460,534,793	91.03%

Preliminary map integration for V1.0

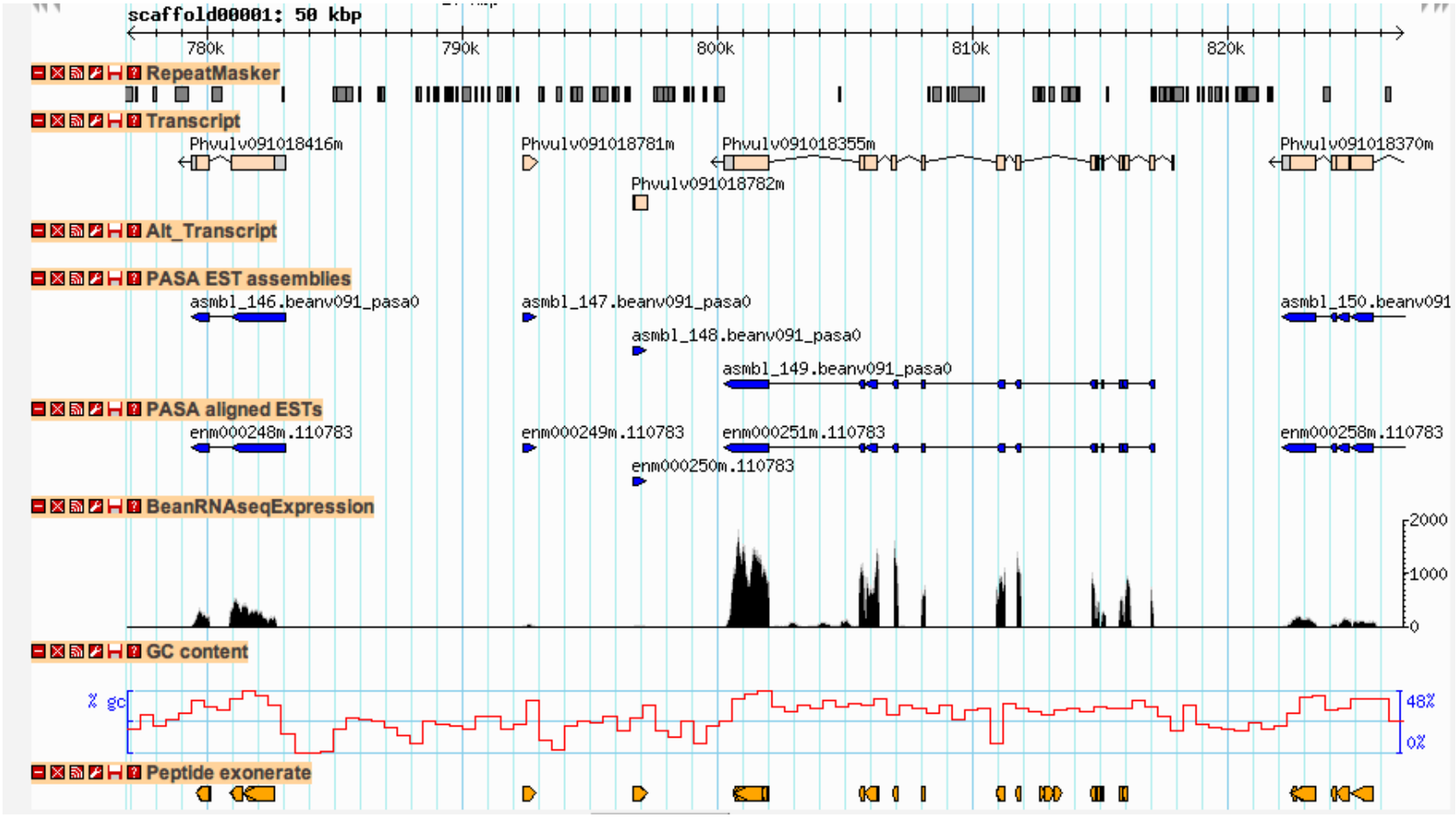


- 6,926/7,018 markers place in the V1.0 genome from the Stampede x Redhawk SNP map
- 81 breaks and 261 joins to make 11 chromosomes

RNA-seq Data Collection for Annotation

Tissue	Reads that can be mapped
Flower Buds	29M
Flowers	52M
Primary Leaves	44M
Young Trifoliates	40M
Roots	76M
Nodules	50M
Young Pods	49M
Stem	77M
Green Mature Pods	100M

Annotation Example V0.9



What is left to do for V1.0?

- Update map based on V1.0 scaffolds
- Build final pseudomolecule set and correct 454 errors
- Final annotation for V1.0
- Analysis and publication

- We will make 1.0 publicly available as soon as the annotation is validated.

Acknowledgements

Sequencing

JGI Production Sequencing Team
HudsonAlpha Production Sequencing Team
Kerrie Barry, JGI – Project Management
Jane Grimwood – HA Laboratory Lead
Erika Lindquist, JGI – RNA
Hope Tice, JGI – Fosmid Libraries
Dave Kudrna, AGI – BAC Libraries

Genetic Mapping

Qijian Song, USDA-Beltsville

Repeat Analysis

Dongying Gao- UGA
Aiko Iwata – UGA

Assembly

Jerry Jenkins, HA
Dave Flowers, HA

Annotation and Display

Shengqiang Shu, JGI
David Goodstein, JGI

Project Leads

Scott Jackson, UGA
Jeremy Schmutz, JGI & HA
Perry Cregan, USDA-Beltsville
Phil McClean, NDSU
Dan Rokhsar, JGI

Funding Sources

USDA-NIFA 2009-01860
DOE DE-AC02-05CH11231
ARRA UC Berkeley

Assembly of the V0.9 Genome

- De-replication of 454 paired libraries
- Organelle and simple sequence screening
- Assembly with Newbler 2.5.3
- Post contamination classification of scaffolds

Annotation Pipeline for V0.9

- RNAseq transcript assemblies were constructed using PERTRAN
- RNAseq transcript assemblies and published ESTs from NCBI were aligned to genome by PASA
- Peptides from soybean, *Arabidopsis thaliana*, poplar, *Medicago truncatula* and grape were BLASTXed to repeatmasked genome and peptides aligned by BLASTX were further aligned by EXONERATE
- Loci were determined from BLAT alignments of PASA EST assemblies and EXONERATE alignments of homologous peptides described above with 2K wiggle room added. Each locus genomic sequence and homologous peptides and EST ORF in the locus were fed into GenomeScan, FGENESH++ and FGENESH_EST for gene prediction. A best gene prediction per locus was selected based on EST assemblies and homologous peptides alignment support. The selected gene predictions were then fed into PASA pipeline where the EST assemblies were obtained for gene model improvement including adding UTRs. PASA improved gene model transcripts were subjected to filtering based on how good the transcript CDS was supported by ESTs and/or homologous peptide, and not overlapped with repeats for more than 20 percent. The filter gene model peptides were assigned PFAM, PANTHER and gene models were further filtered for those with 30% or more of peptides assigned to transposable element domains.

Identifying and Classifying Transposons

- Based on the assembled portion of V0.9:
  - LTR retrotransposons were annotated by the LTR-Finder program
  - Non-LTR retrotransposons, LINEs and SINEs, were recognized by poly A or Poly T motifs as well as by the retrotransposase (LINEs)
  - DNA transposons were analyzed using the conserved domains of transposases from different superfamilies
  - two reported LTR retrotransposons of common bean, **pva1-118d24-re-5** and **Tpv2-6,** also were included