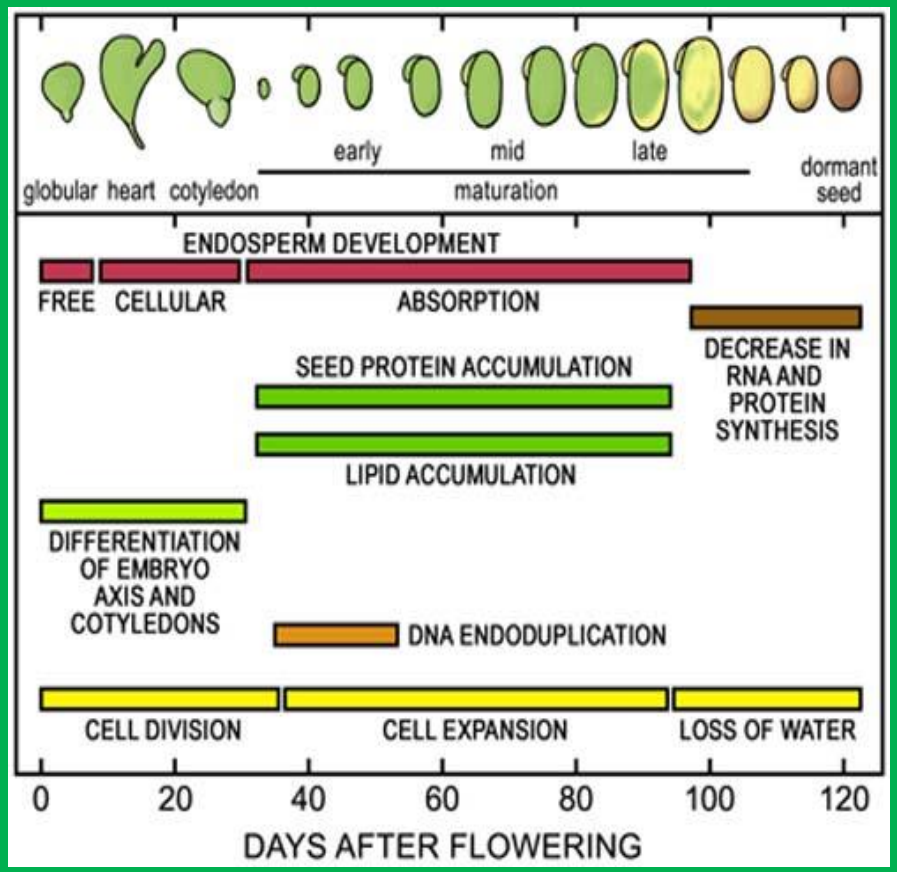
Exploring Soybean Transcript Polymorphisms to Identify Gene Variants and Functional Markers for Seed Quality Improvement

Yong-Qiang (Charles) An
USDA-ARS at Danforth Plant Science Center, St. Louis, MO

Soybean Breeder Workshop 2014

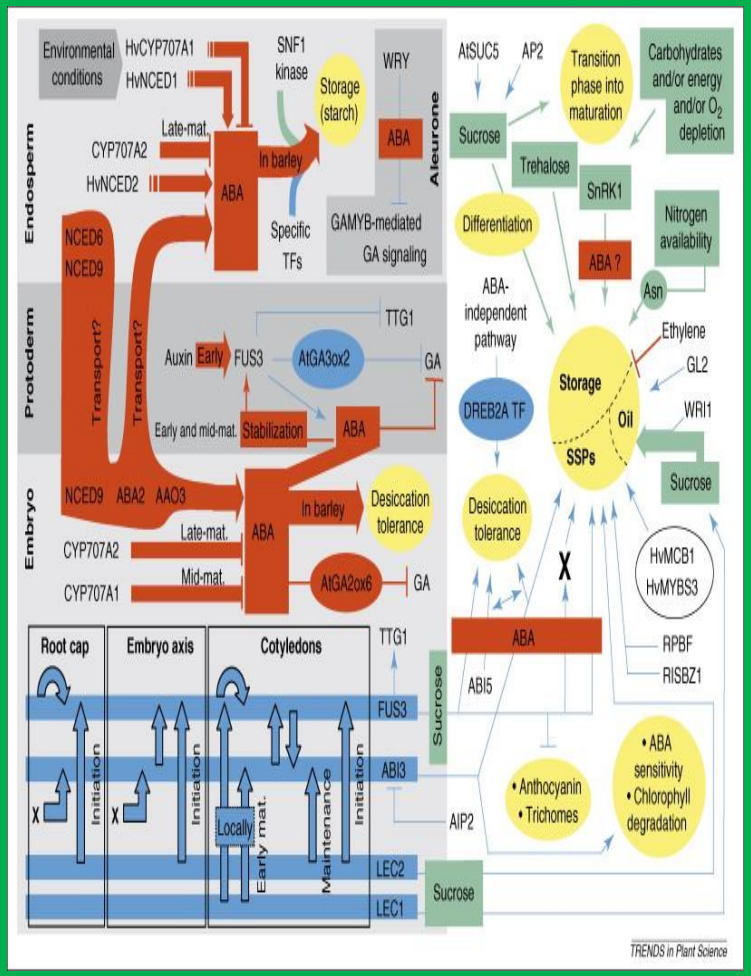
Why to Study Seeds



(Le et.al 2007)

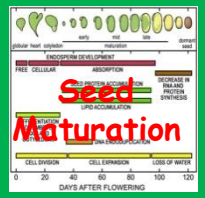
- Important to seed quality and yield
- Seed filling and reserve production are adaptive traits
- Ideal for genetic engineering to improve seed quality without affecting plant normal growth.

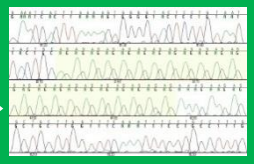
Soybean Seed Networks



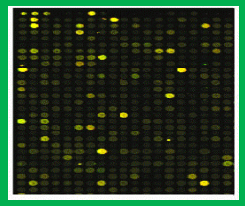
- Accomplished through concerted and interactive activities of many gene products and pathways.
- Breeders are mainly optimizing biological networks through mixing the gene variants and selecting the best of combinations.
- Understanding of gene networks and collection of gene variants and functional markers

Overall Strategies

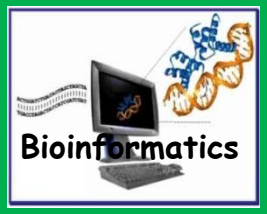




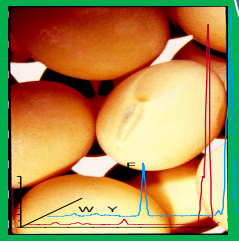
Sequences



Expression



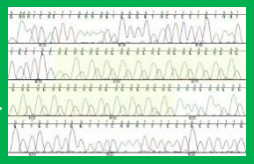
Gene Network
Gene variants
Functional Markers



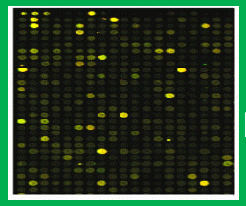
Phenotypic profile
(oil and proteins)

Biotech and Breeding

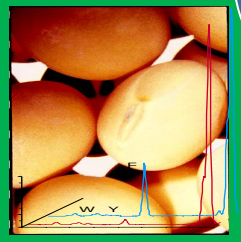
Explore Soybean Seed Diversity



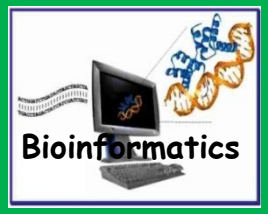
Sequences



Expression



Phenotypic profile
(oil and proteins)



Gene variants
Functional Markers

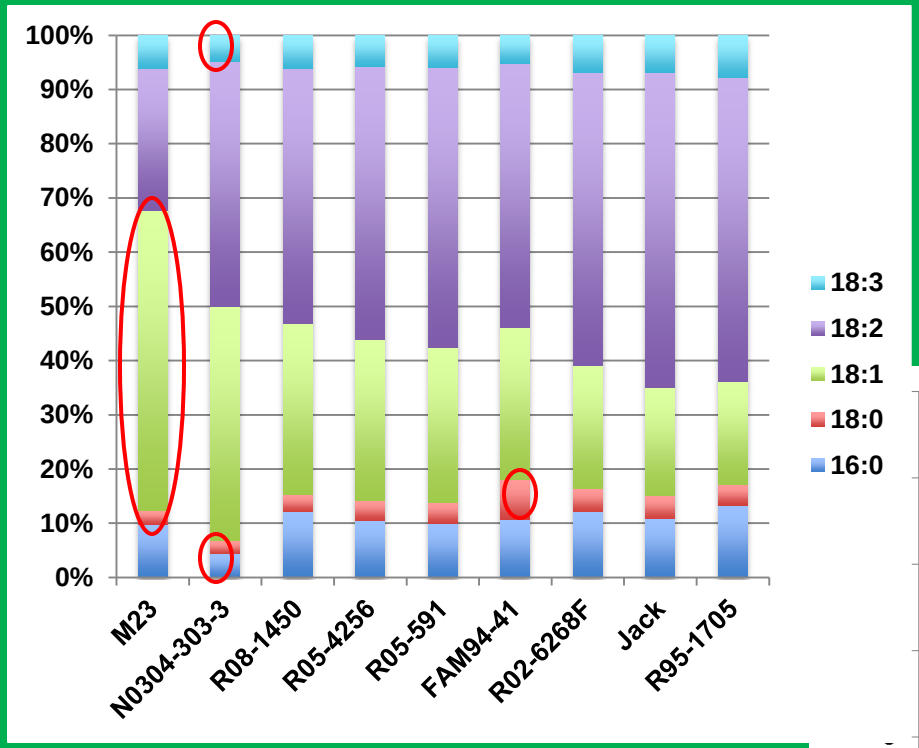
Explore Seed Diversity By Sequencing Transcriptomes

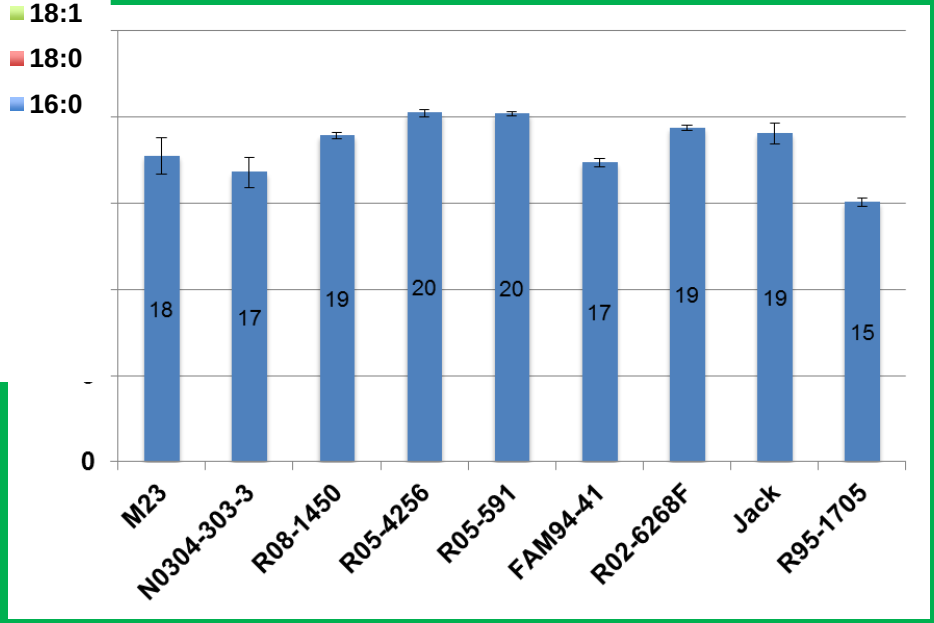
- Wild soybean: Glycine soja
- Ancestral landraces
- Milestone cultivars
- NAM parents
- Various seed mutants

Proof-of-concept study

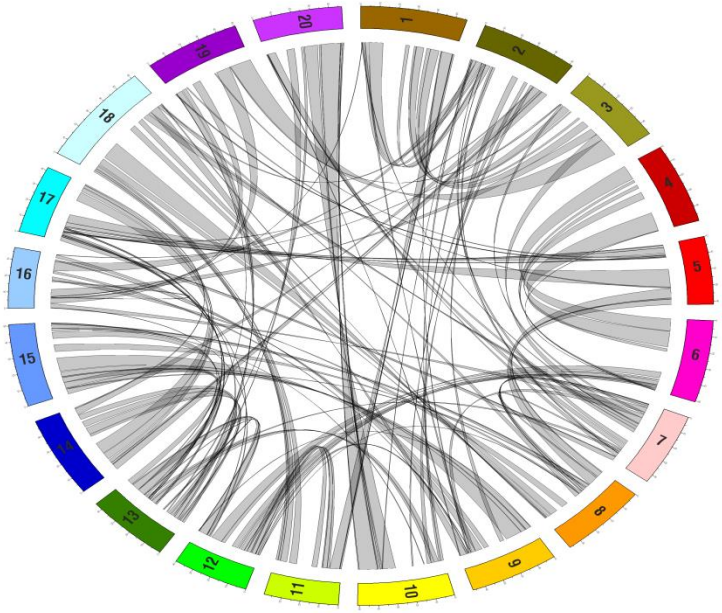
- Sequencing nine lines varying in oil composition and content

Fatty Acid Profiles and Oil Content





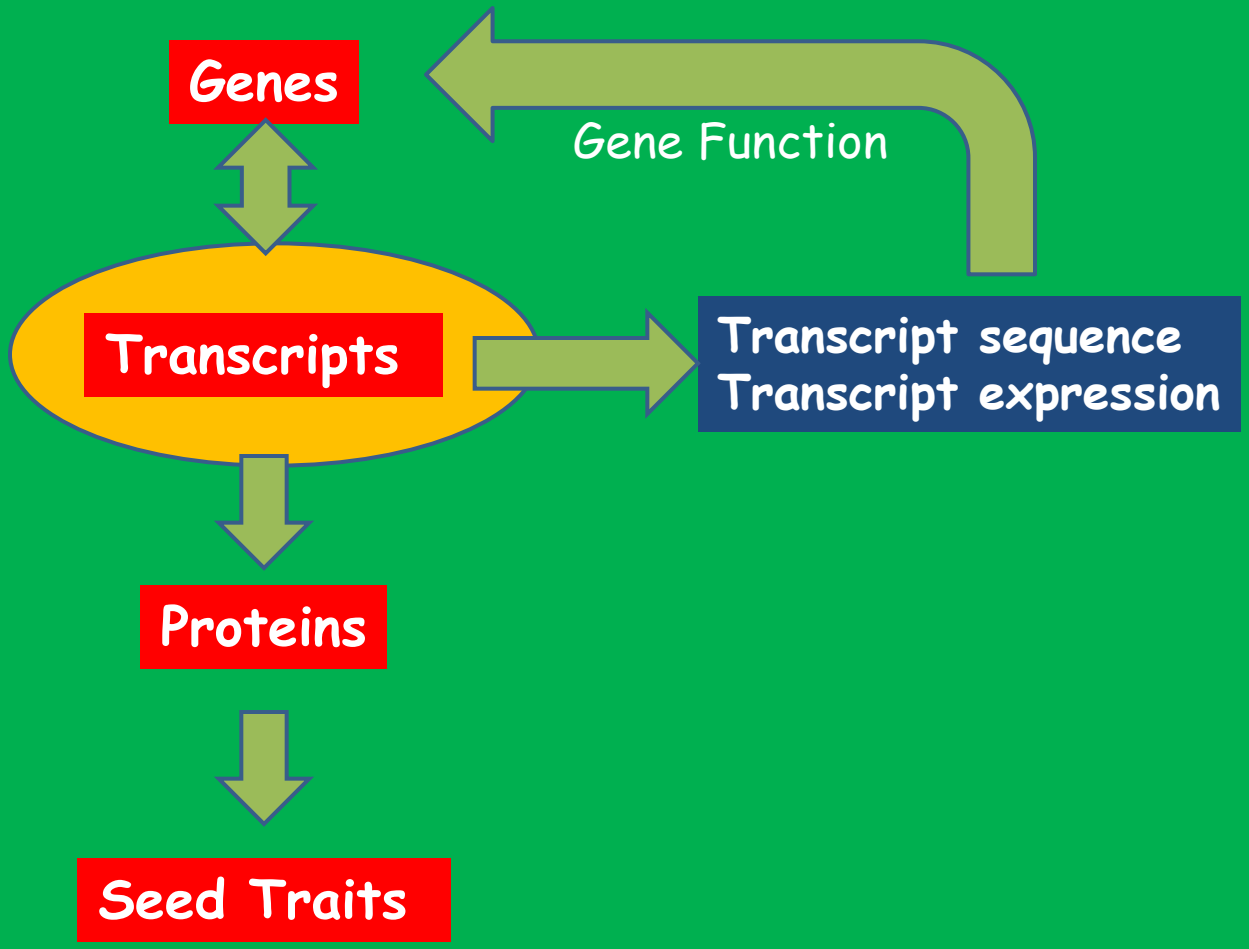
The Complex Soybean Genome



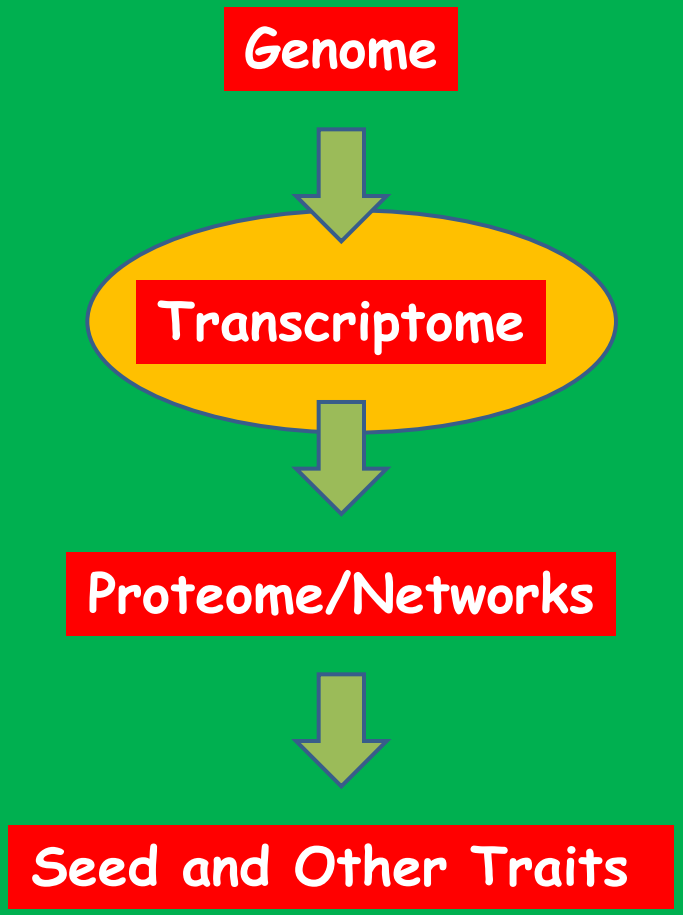
- Whole genome duplications
  - 59 mya early-legume duplication
  - 13 mya Glycine-specific duplication (allotetraploidy event)
- Nearly 75% of the genes present in multiple copies

- Genome size: 1.1 gb
- 59% of genome transposons
- Less than 10% genome transcribed into RNAs (Transcriptomes)

Central Dogma of Molecular Biology



Exploring Diversity by Sequencing Seed Transcriptomes



- Focus on functional gene components (not random)
- genes functionally related to seed traits
- Transcript sequence and expression at same time (two attributes of a gene function)
- Less than 10% of the genome

- Not transcribed functional sequences:
  - Promoter sequences
  - Not expressed gene sequences

- Transcript Polymorphisms:
  - Transcript sequences (SNPs and indels)
  - Transcript expression
  - Transcript splicing

Summary of Transcriptome Sequencing

Total no. of sequenced reads (in million)	34.6
Total length of transcripts covered (in million nt)	72.0
Total length of annotated transcripts covered (in million nt)	52.6
Percentage of genome sequence covered	7%
Percentage of annotated transcript sequence covered	55%
Average depth of coverage per transcript nt	21
Average depth of coverage per annotated transcript nt	27
No. of genes identified (in thousand)	33.3
No of novel genes identified	755
No. of annotated genes identified (in thousand)	32.5
Percentage of annotated genes identified (%)	60%

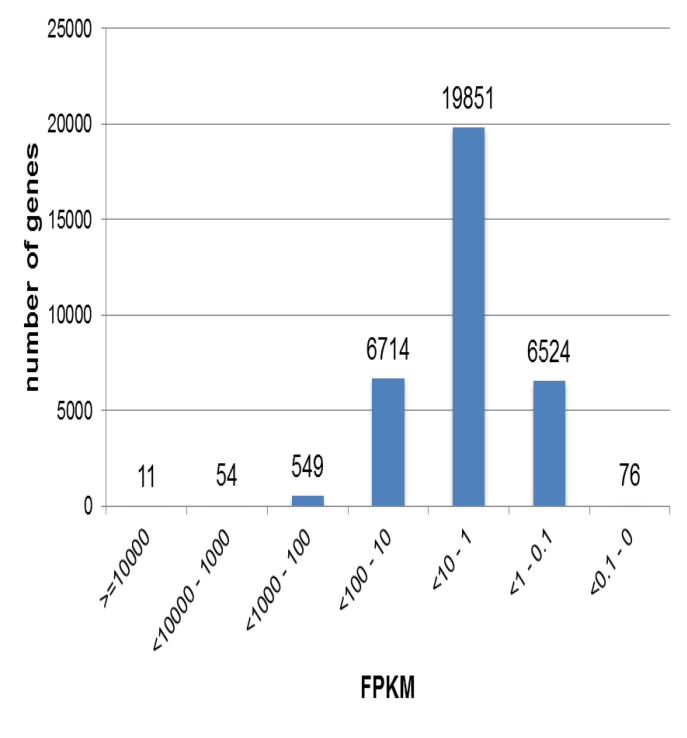
Goettel et. al BMC Genomics (in press)

Summary of Transcriptome Sequencing

	Average
Total no. of sequenced reads (in million)	34.6
Total length of transcripts covered (in million nt)	72.0
Total length of annotated transcripts covered (in million nt)	52.6
Percentage of genome sequence covered	7%
Percentage of annotated transcript sequence covered	55%
Average depth of coverage per transcript nt	21
Average depth of coverage per annotated transcript nt	27
No. of genes identified (in thousand)	33.3
No of novel genes identified	755
No. of annotated genes identified (in thousand)	32.5
Percentage of annotated genes identified (%)	60%

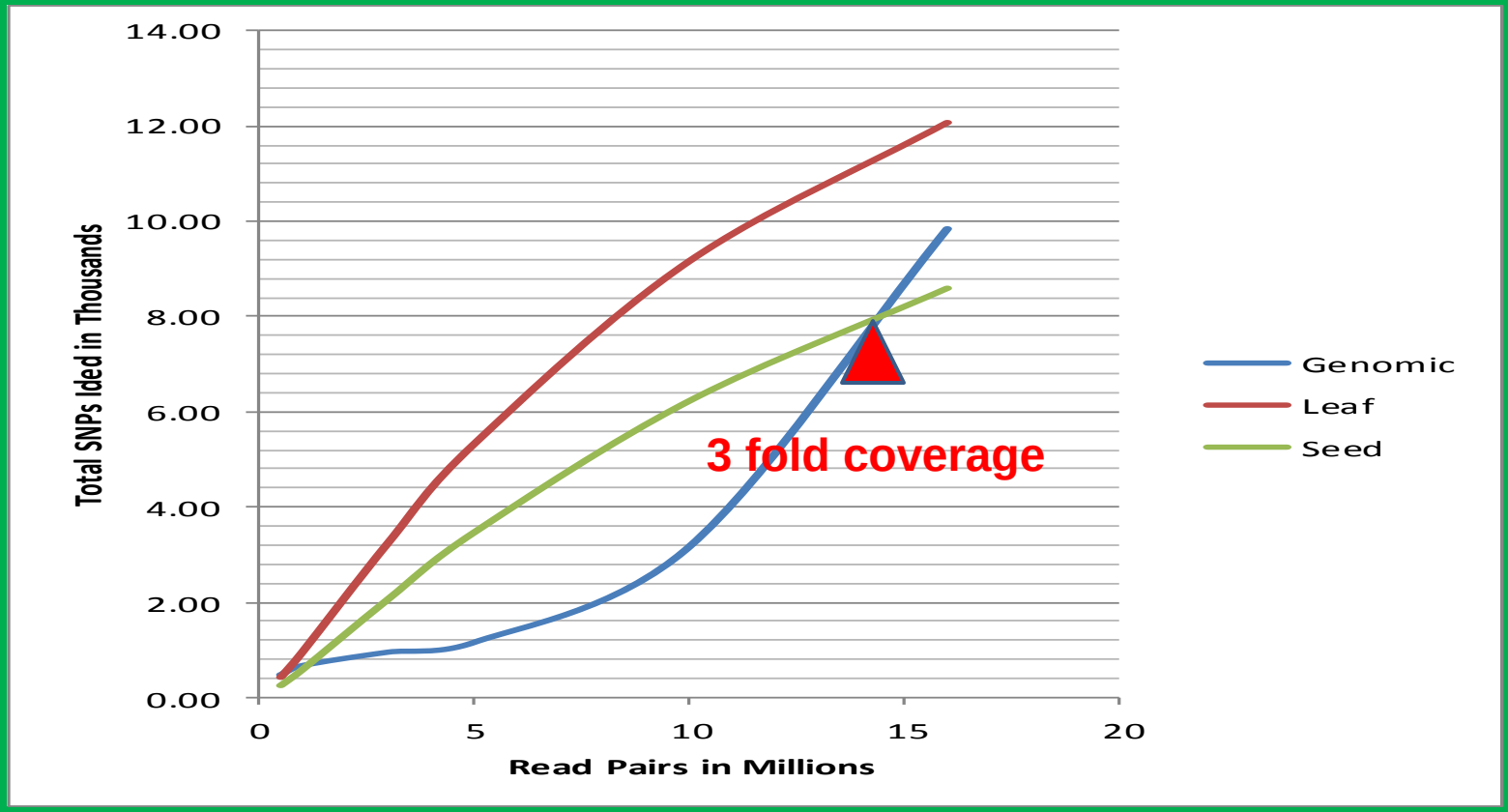
Goettel et. al BMC Genomics (in press)

Seed Transcriptomes



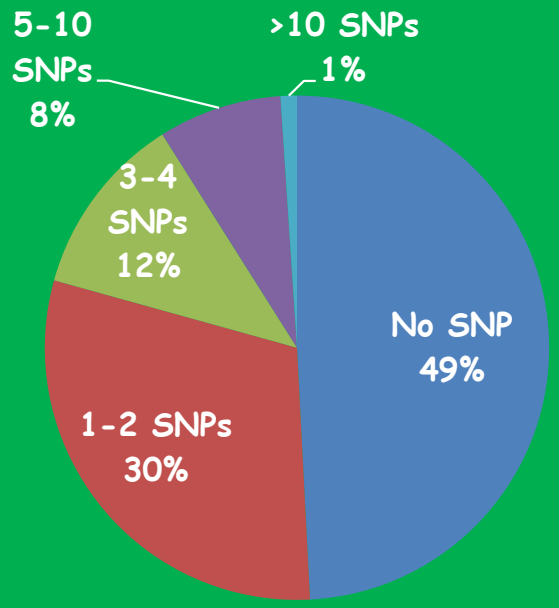
Gene Functions	no. of genes	% of total mRNA
seed storage proteins	10	19.2
protease inhibitors	3	4.1
proteases	4	3.9
acyl lipid enzymes	4	3
oil body proteins	7	2.8
LEA /dehydrins	17	4.6
metallothioneins	3	1.4
ribosomal protein	2	0.3
lectins	2	0.8
miscellaneous	16	5.8
unknown protein	11	4.3
Total	79	50.1

SNP Discovery by Genome and Transcriptome Sequencing



Transcript Sequence Polymorphisms

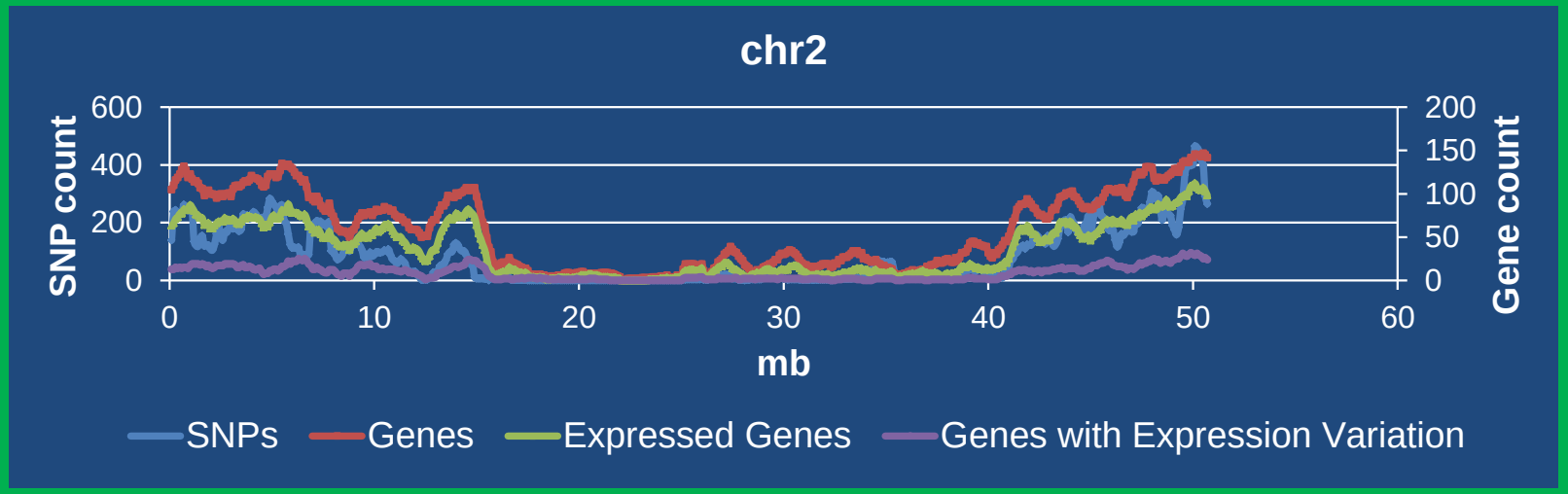
Soybean Line	Average	Total
Total Transcript SNPs	23,669	48,792
Overlapped with dbSNP	81%	82%
Validated by genome sequencing	90%	
Overlapped with SoySNP50K	5%	5%
Total No. of Indels	952	1,693
Indels overlapping with dbSNP	338	578
overlapped with dbSNP	36%	34%

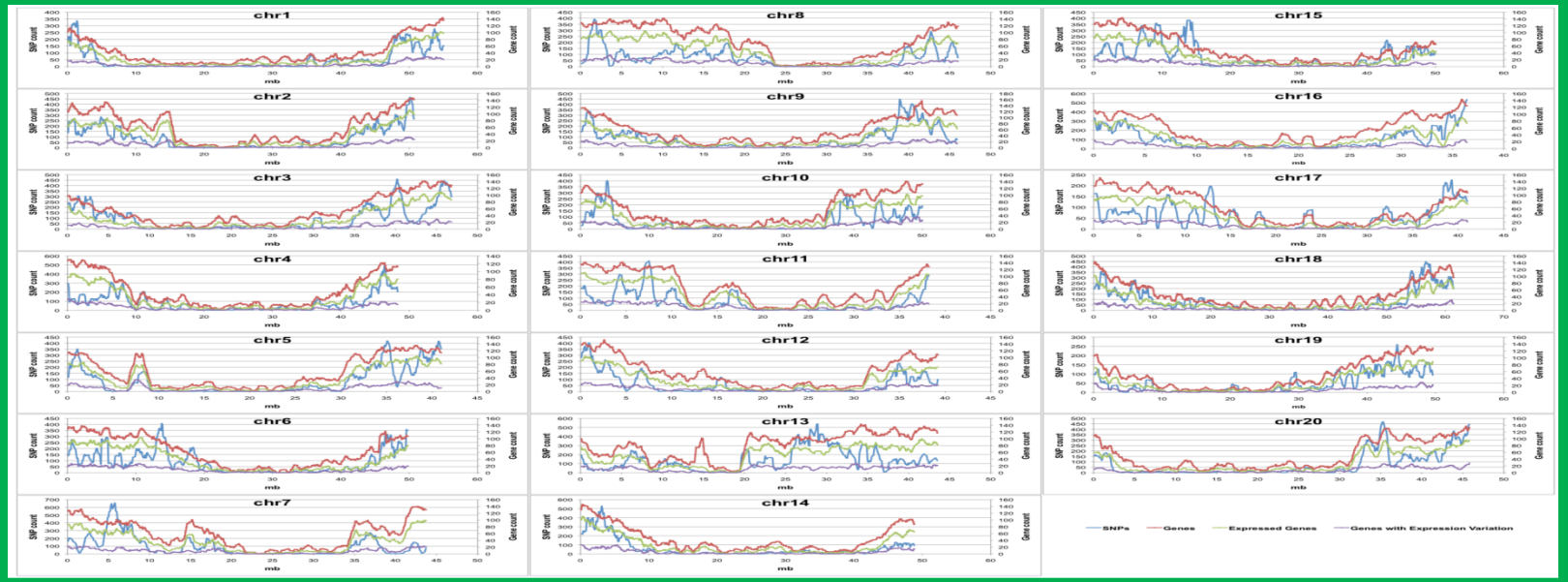


Genes Containing SNPs

Average: 1 SNP/1.5kb Transcript Sequence

Distribution of Transcript Polymorphisms





Functional Mapping of Transcript SNPs and Indels

Soybean Line	Jack	FAM94-41	M23	N0304-303	Average	Total
SNPs in annotated Transcript	14271	23757	22450	22311	20919	43283
SNPs in UTRs	4156	7226	6659	6657	6264	12748
SNPs in CDS	10115	16531	15791	15654	14655	30535
Synonymous SNPs	5365	8731	8366	8329	7715	16196
Non-synonymous SNPs	4759	7814	7441	7338	6955	14372
SNPs eliminating start codons	6	11	9	8	9	17
SNPs causing premature termination codons	29	70	68	67	59	125
SNPs eliminating termination codons	24	28	29	31	27	49
SNPs in splice sites	220	369	342	323	309	606
INDELs in CDS	134	198	166	164	170	241
INDELs in termination codons	2	3	3	3	3	4
INDELs in splice sites	16	33	26	30	26	35

Functional Mapping of Transcript SNPs and Indels

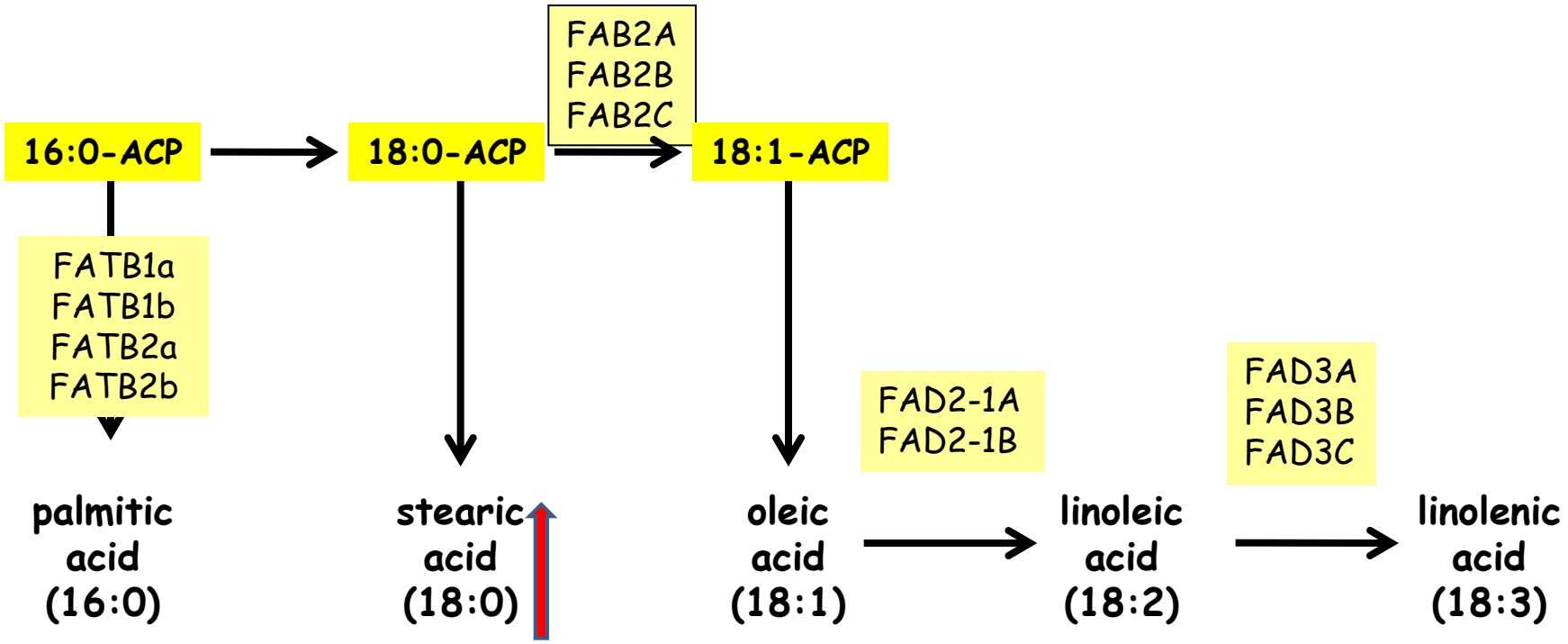
Soybean Line	Jack	FAM94-41	M23	N0304-303	Average	Total
SNPs in annotated Transcript	14271	23757	22450	22311	20919	43283
SNPs in UTRs	4156	7226	6659	6657	6264	12748
SNPs in CDS	10115	16531	15791	15654	14655	30535
Synonymous SNPs	5365	8731	8366	8329	7715	16196
Non-synonymous SNPs	4759	7814	7441	7338	6955	14372
SNPs eliminating start codons	6	11	9	8	9	17
SNPs causing premature termination codons	29	70	68	67	59	125
SNPs eliminating termination codons	24	28	29	31	27	49
SNPs in splice sites	220	369	342	323	309	606
INDELs in CDS	134	198	166	164	170	241
INDELs in termination codons	2	3	3	3	3	4
INDELs in splice sites	16	33	26	30	26	35

Functional Mapping of Transcript SNPs in Oil-Related Genes and Pathways

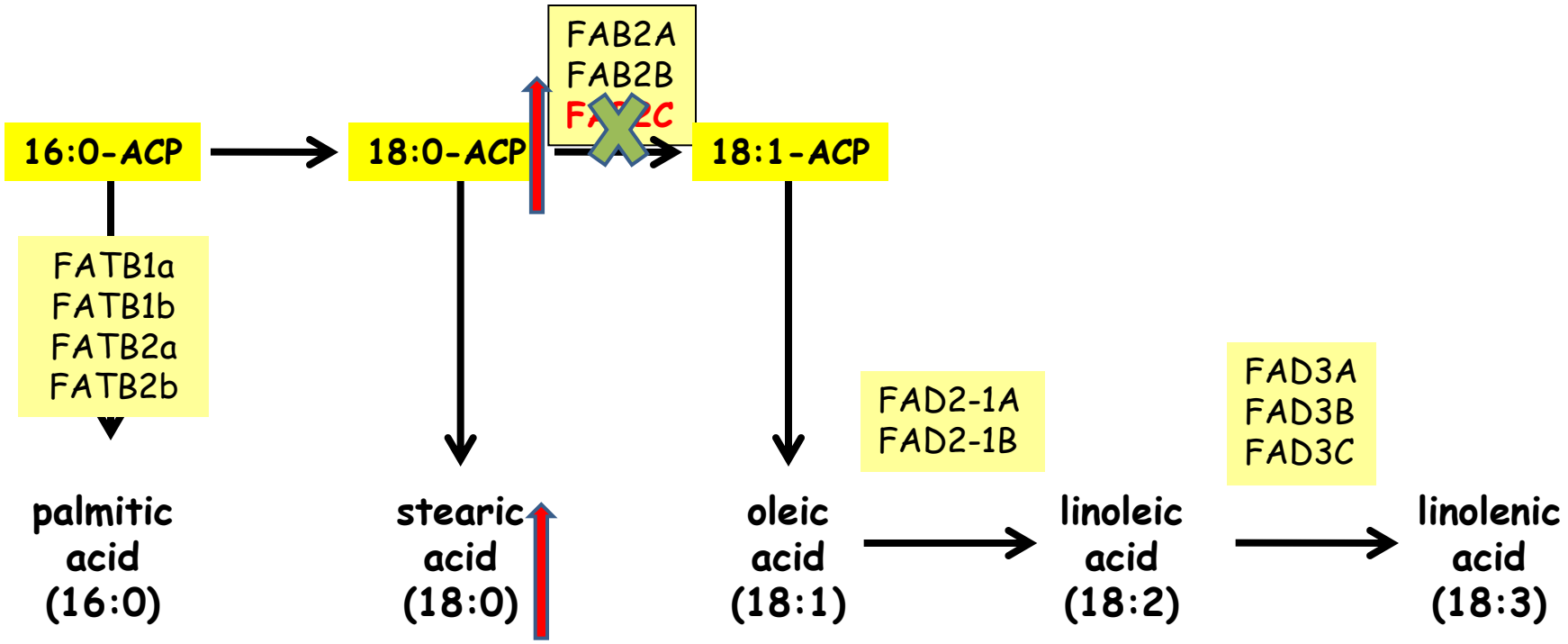
Functional Categories	Genes	Total SNPs	nsSNPs
1. Synthesis of fatty acids in plastids	73	107	24
2. Synthesis of membrane lipids in plastids	51	70	14
3. Synthesis of membrane lipids in endomembrane system	103	133	33
4. Metabolism of acyl lipids in mitochondria	72	49	12
5. Synthesis and storage of oil	33	38	4
6. Degradation of storage lipids and straight fatty acids	143	142	33
7. Lipid signalling	297	305	68
8. Fatty acid elongation and wax and cutin metabolism	84	30	12
9. Miscellaneous	302	327	87
Transcript Factor Genes	1287	1525	439
Transcript Factor Hub Genes	279	414	118

Transcript	reference	Code position	variant	Functional prediction	Gene Functions
Glyma07g09370.1	P	24	T	deleterious	Enoyl-CoA hydratase/isomerase
Glyma10g39540.1	R	102	C	deleterious	long-chain acyl-CoA synthetase 6
Glyma17g03130.1	F	257	L	deleterious	alpha/beta-Hydrolases superfamily protein
Glyma05g33310.1	D	256	G	deleterious	Amidase family protein
Glyma16g08960.3	T	137	M	deleterious	beta-hydroxyisobutyryl-CoA hydrolase 1
Glyma10g32660.2	R	267	C	deleterious	cytidinediphosphate diacylglycerol synthase 2
Glyma11g13070.1	F	381	L	deleterious	allene oxide synthase
Glyma18g42540.1	G	434	R	deleterious	sulfoquinovosyldiacylglycerol 2
Glyma14g04590.1	P	351	L	deleterious	Protein of unknown function
Glyma14g04590.1	H	360	L	deleterious	Protein of unknown function
Glyma04g11550.1	S	###	L	deleterious	acetyl-CoA carboxylase 1
Glyma18g02110.1	I	###	N	deleterious	peroxisomal ABC transporter 1
Glyma12g04990.1	Q	309	L	deleterious	transducin family protein
Glyma01g39920.1	N	210	Y	deleterious	acyl-CoA-binding domain 3
Glyma13g42310.1	E	393	D	deleterious	lipoxygenase 1
Glyma03g07540.3	T	155	A	deleterious	acyl-CoA oxidase 4
Glyma04g14650.1	G	87	S	deleterious	acyl-CoA-binding protein 6
Glyma20g37940.1	Y	451	C	neutral	acyl-CoA binding protein 4
Glyma18g03090.1	A	145	S	neutral	phospholipase C 2
Glyma01g39350.1	S	20	C	neutral	alpha/beta-Hydrolases
Glyma14g27990.1	D	77	N	neutral	stearoyl-acyl-carrier-protein desaturase protein
Glyma07g07750.2	D	89	N	neutral	alpha/beta-Hydrolases protein
Glyma08g16620.1	D	203	G	neutral	phosphatidylinositol 4-OH kinase beta1
Glyma01g01910.1	P	37	T	neutral	Plastid-lipid associated protein PAP

Elevated Stearic Acid Content *in FAM94*



Deleterious Mutation of *FAB2C* in FAM94



SNPs in *FAB2C* (Glyma14g27990.1)

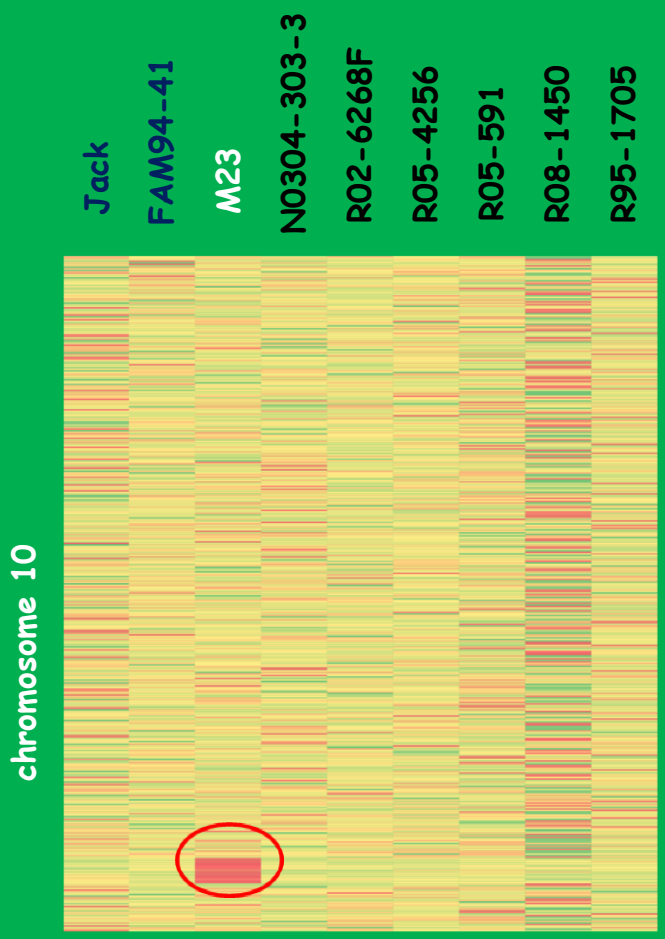
	Variant Type	Jack	FAM94	M23	N0304-303-3	R02-6268F	R05-4256	R05-591	R08-1450	R95-1705
codon 77	Non-Synonymous	GAC	AAC	GAC	GAC	GAC	GAC	GAC	GAC	GAC
amino acid		Asp	Asn	Asp	Asp	Asp	Asp	Asp	Asp	Asp
codon 310	Synonymous	GAA	GAA	GAG	GAA	GAG	GAA	GAA	GAG	GAA
amino acid		Glu	Glu	Glu	Glu	Glu	Glu	Glu	Glu	Glu

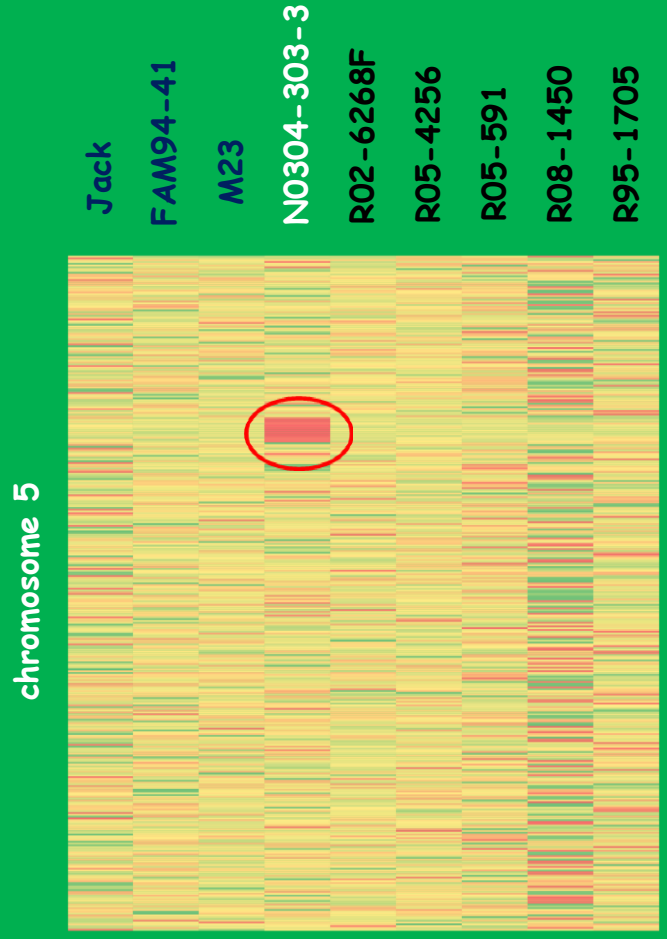
Transcript Expression Polymorphisms
(8037 genes)

Gene ID	Genes	Mean FPKM	Max Fold Change	Soybean Line
Glyma07g06960	LEA	114.02	-850.71	FAM94-41
Glyma07g28940	BURP domain protein	20.77	-420.71	R08-1450
Glyma06g04740	Gibberellin-regulated protein	56.42	-317.91	R08-1450
Glyma01g39590	unknown	13.64	-300.07	R08-1450
Glyma14g03685	unknown	23.03	-279.51	R08-1450
Glyma02g06882	unknown	225.76	-261.86	R08-1450
Glyma13g02960	EamA-like transporter	33.83	-242.83	R08-1450
Glyma20g01430	unknown	38.56	-239.35	R08-1450
Glyma05g22770	ACT domain repeat 1	11.54	-231.64	M23
Glyma04g37520	Nodulin MtN3 family protein	11.33	-224.46	R08-1450

Function Category of Acyl Lipid Genes	Acyl Lipid Genes	Expressed	Varied	exp/all	Varied /Exp
Synthesis of fatty acids in plastids	78	70	11	90%	16%
Synthesis of membrane lipids in plastids	54	31	13	57%	42%
Synthesis of membrane lipids in endomembrane system	101	89	19	88%	21%
Metabolism of acyl lipids in mitochondria	33	30	5	91%	17%
Synthesis and storage of oil	36	31	4	86%	13%
Degradation of storage lipids and straight fatty acids	142	104	29	73%	28%
Lipid signalling	295	191	52	65%	27%
Fatty acid elongation and wax and cutin metabolism	84	38	11	45%	29%
Miscellaneous	267	185	44	69%	24%
Total	**1090**	**769**	**188**	**74%**	**24%**

Identification of Large DNA Deletions

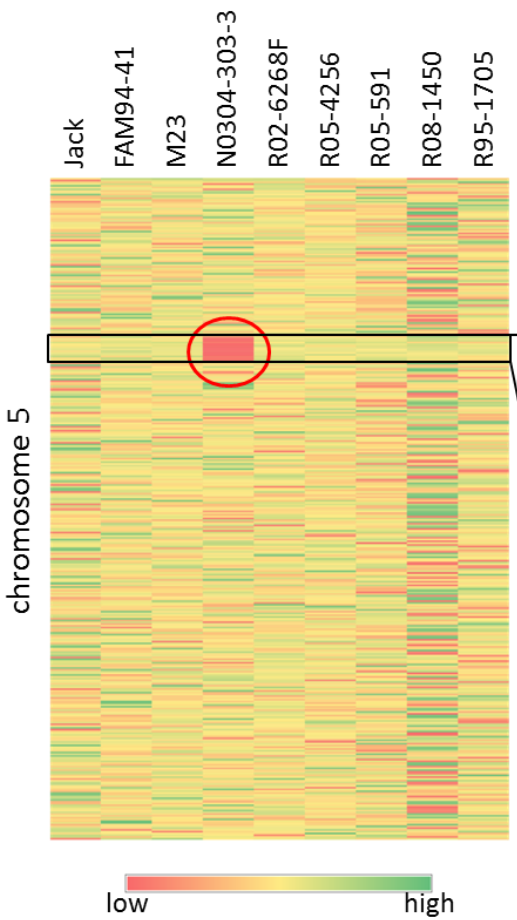




164 kb Deletion in M23

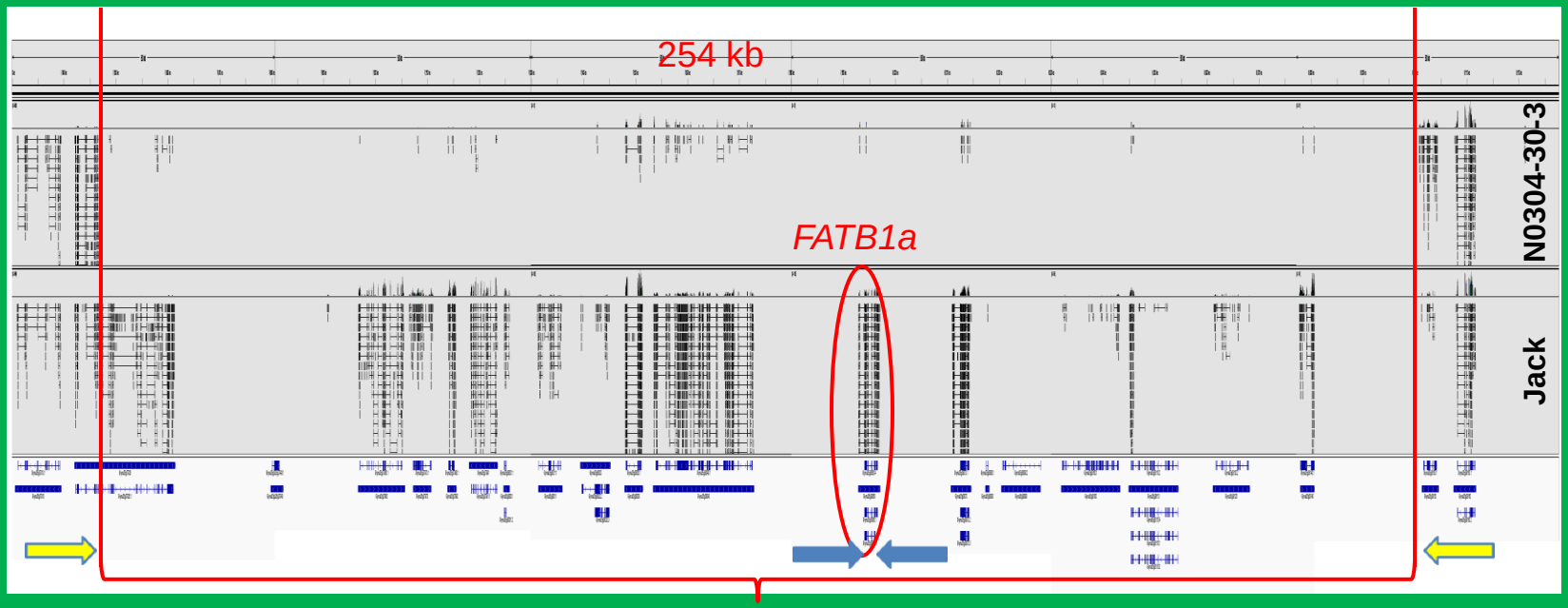
gene models	annotation	mean FPKM	mean abs Z-score	M23 Z-score
Glyma10g42360	Kow domain-containing transcription factor 1	5.72	0.84	0.34
Glyma10g42360	Kow domain-containing transcription factor 1	4.76	0.85	-0.08
Glyma10g42380	Protein phosphatase 2A	0.43	0.87	0.44
Glyma10g42390	Pectin lyase-like superfamily protein	0.96	0.78	0.29
Glyma10g42420	Inner centromere protein, ARK binding region	0.29	0.73	-0.60
Glyma10g42440	Cationic amino acid transporter 7	5.07	0.56	-1.92
Glyma10g42450	Duplicated homeodomain-like	4.23	0.48	-2.11
Glyma10g42461	F1F0-ATPase inhibitor protein	145.50	0.35	-2.57
Glyma10g42470	**FAD2-1A, fatty acid desaturase 2**	**430.73**	**0.44**	**-2.40**
Glyma10g42480	Vesicle-associated membrane protein 727	40.82	0.34	-2.60
Glyma10g42490	Clathrin adaptor complexes medium subunit	11.96	0.33	-2.65
Glyma10g42500	Hemerythrin HHE cation binding domain	21.92	0.32	-2.48
Glyma10g42510	Polynucleotidyl transferase	5.52	0.37	-2.54
Glyma10g42520	Unknown	5.10	0.39	-2.34
Glyma10g42530	Nucleic acid-binding, OB-fold-like protein	2.71	0.55	-2.03
Glyma10g42540	Ferredoxin 3	8.14	0.48	-2.42
Glyma10g42551	Phosphotyrosine protein phosphatase	2.09	0.43	-2.43
Glyma10g42560	ARM repeat superfamily protein	3.02	0.40	-2.45
Glyma10g42580	ACT domain-containing protein	19.33	0.33	-2.62
Glyma10g42590	Unknown	3.18	0.43	-2.50
Glyma10g42600	Radical SAM superfamily protein	1.31	0.44	-2.18
Glyma10g42630	GHMP kinase family protein	4.93	0.33	-2.63
Glyma10g42650	20S proteasome alpha subunit E2	20.11	0.32	-2.49
Glyma10g42660	Indeterminate(ID)-domain 5	11.64	0.88	-0.29
Glyma10g42680	UDP-glucosyl transferase 73B5	9.11	0.58	1.01
Glyma10g42691	GTP binding	4.35	0.81	-0.54
Glyma10g42700	Nucleotide-diphospho-sugar transferase	4.55	0.84	-0.10

Identification of 254 kb Deletion in N0304



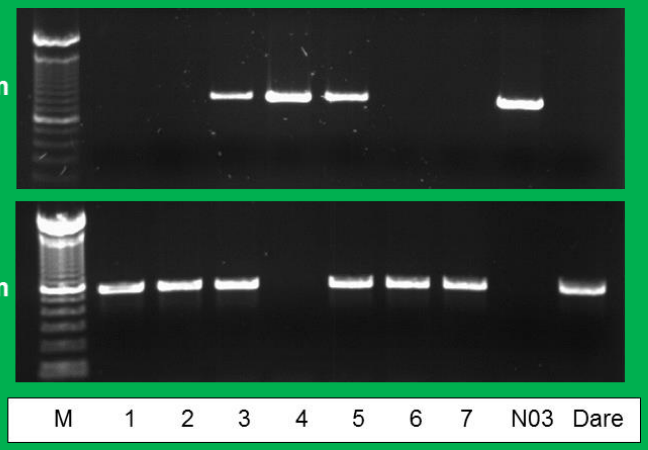
gene name	annotation	mean FPKM	N0304- 303-3 FPKM	mean abs Z-score	N0304-303-3 Z-score
Glyma05g07840	Nucleic acid-binding, OB-fold-like protein	9.38	13.14	0.69	1.62
Glyma05g07870	Arogenate dehydrogenase	2.93	3.21	0.80	0.56
Glyma05g07875	hAT family dimerisation domain	0.48	0.77	0.78	1.15
Glyma05g07880	Oleosin family protein	96.69	152.59	0.86	1.23
Glyma05g07910	Transcription factor TFIIE, alpha subunit	3.07	2.61	0.85	-0.13
Glyma05g07926	TBP-associated factor 2	6.56	3.66	0.59	-2.02
Glyma05g07960	Transmembrane Fragile-X-F-associated protein	6.08	0.05	0.32	-2.58
Glyma05g07970	Unknown	2.98	0.03	0.42	-2.47
Glyma05g07980	homeodomain-like protein	4.67	0.21	0.55	-2.28
Glyma05g07991	Histone deacetylase	7.38	0.25	0.34	-2.60
Glyma05g08001	Histone deacetylase	5.81	0.00	0.40	-2.50
Glyma05g08011	methyltransferases	2.96	0.00	0.33	-2.62
Glyma05g08020	Plant protein of unknown function (DUF936)	2.18	0.04	0.72	-1.72
Glyma05g08030	Unknown	38.22	0.92	0.33	-2.62
Glyma05g08040	DNAJ heat shock protein	12.74	0.28	0.38	-2.54
Glyma05g08060	**Fatty acyl-ACP thioesterases B FATB**	**18.30**	**0.11**	**0.34**	**-2.61**
Glyma05g08070	phosphatase 2A regulatory B subunit	13.48	0.35	0.32	-2.60
Glyma05g08110	WD40/YVTN repeat-like-containing domain	1.00	0.03	0.49	-2.30
Glyma05g08120	Sodium Bile acid symporter family	1.39	0.00	0.44	-2.37
Glyma05g08140	Leucine-rich repeat protein kinase	3.68	0.05	0.46	-2.36
Glyma05g08150	Homeodomain-like superfamily protein	0.43	2.53	0.33	2.56
Glyma05g08160	RNA-bindingprotein	5.99	5.55	0.79	-0.13
Glyma05g08190	3-ketoacyl-CoA synthase 11	1.81	1.97	0.79	0.68
Glyma05g08200	Transducin/WD40 repeat-like protein	14.60	18.07	0.72	1.24
Glyma05g08220	RING/FYVE/PHD zinc finger protein	4.32	3.63	0.87	-0.76

Correlation of FATB1a and Deletion with Low Palmitic Acid



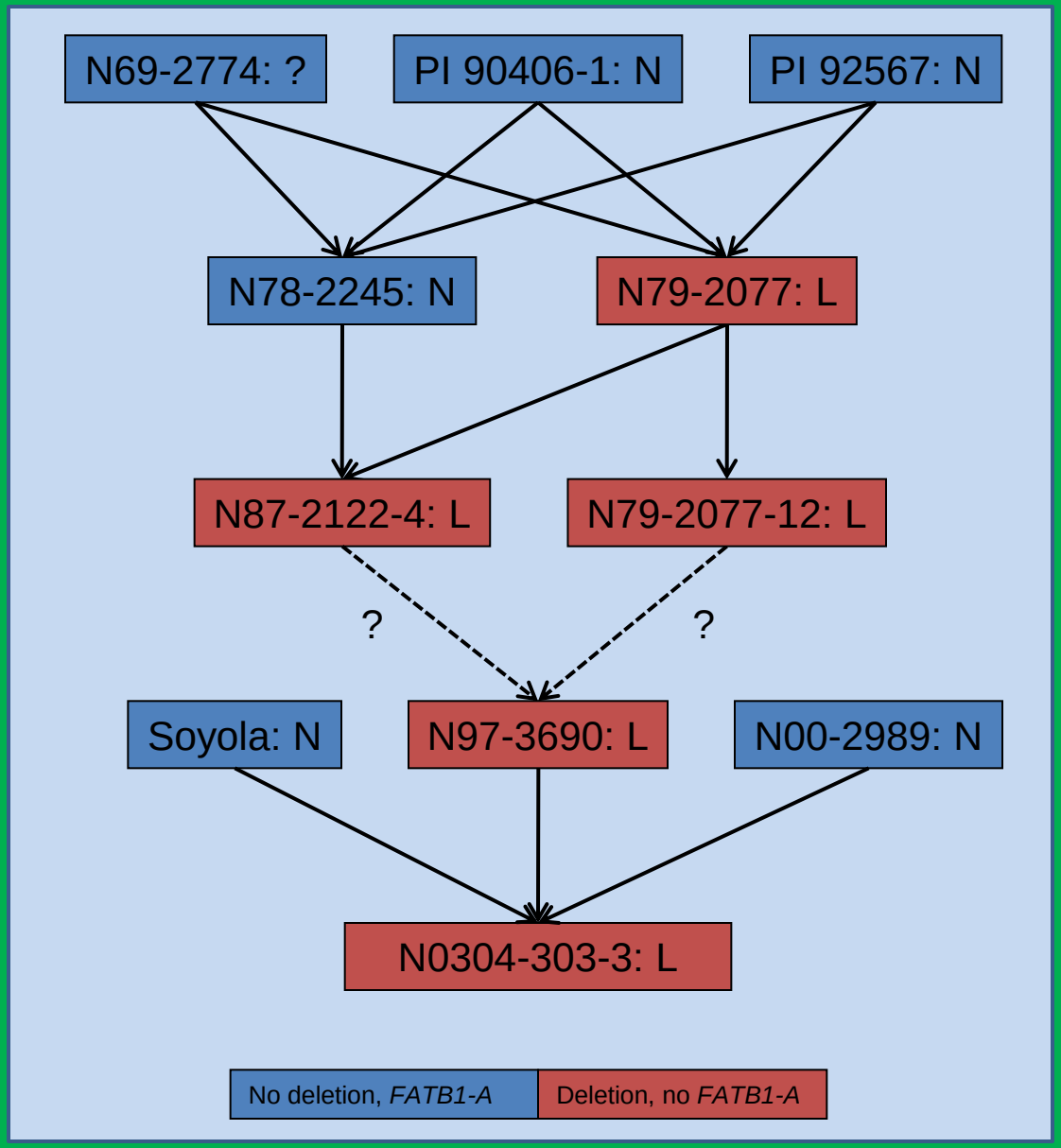
PCR amplification across deletion

PCR amplification of *FATB1a*



lane no.	lines	16:0 (%)
1	Soyola	10.7
2	N97-3363-4	7.7
3	N79-2077-12	5.4
4	N97-3690	4.4
5	N87-2122-4	5.3
6	PI 123440	9.6
7	N00-2989	11.3
indicated	N0304-303-3	4
indicated	Dare	10.1

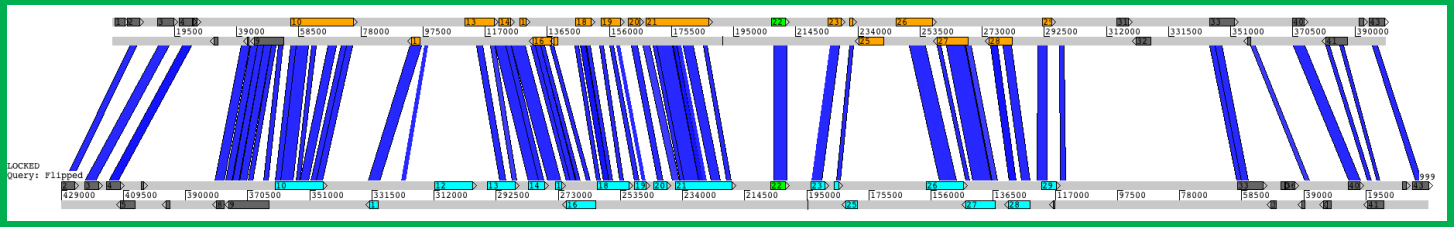
Pedigree and Origin of the Deletion



Duplicated Regions of the N0304 Deletion

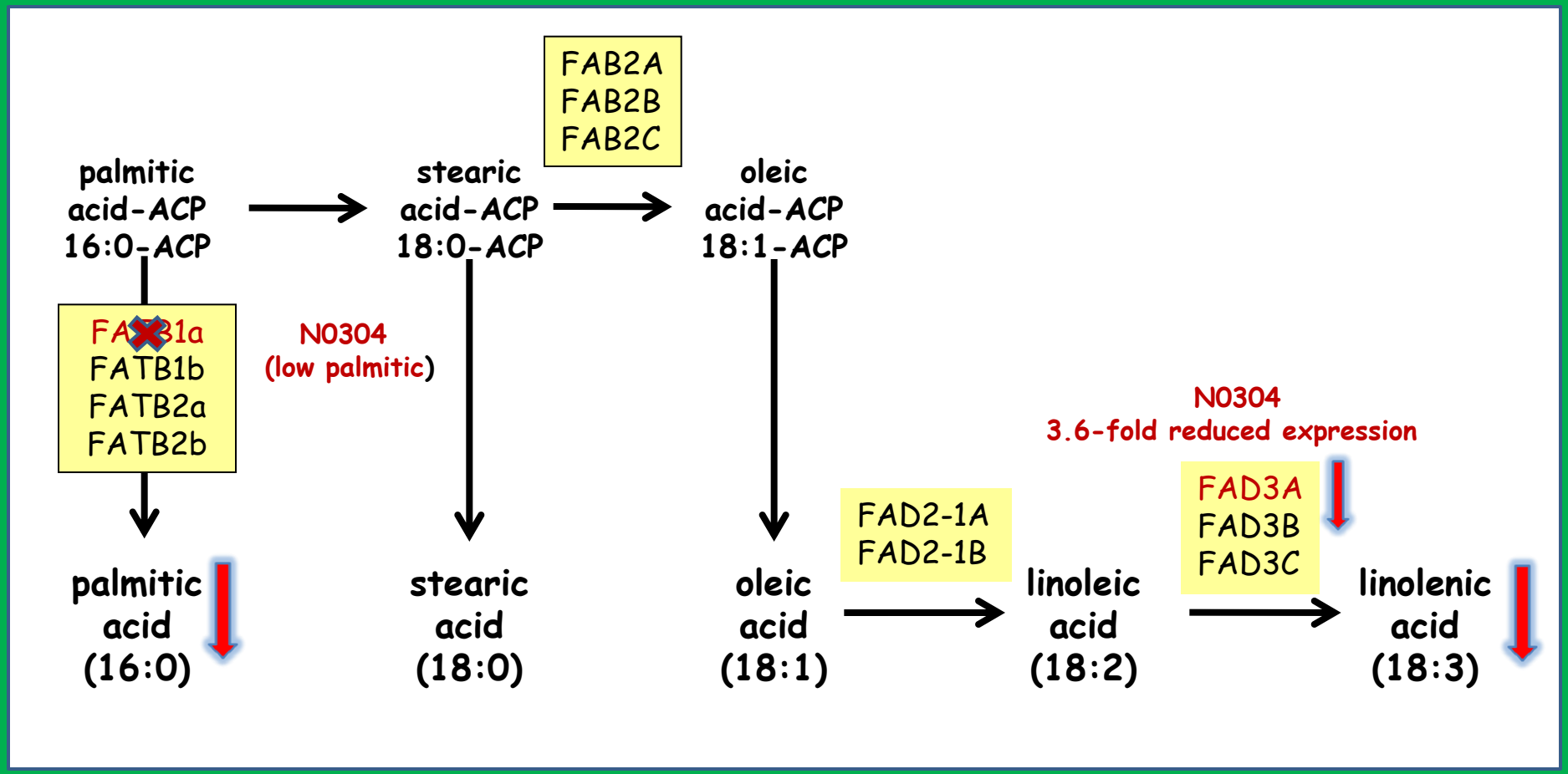
Chr5

chr17

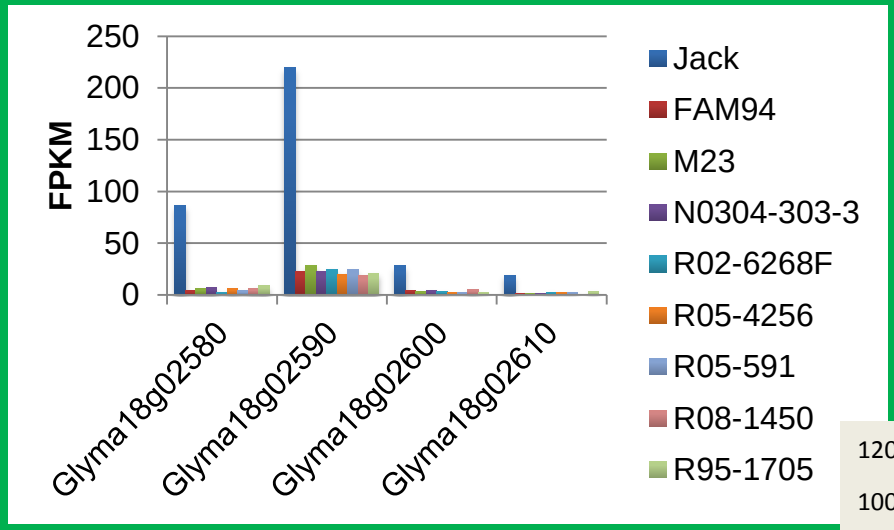


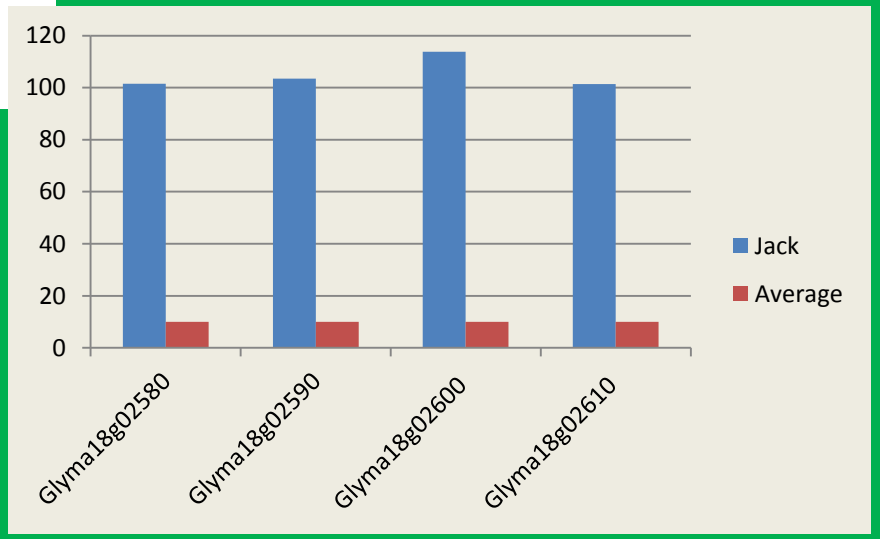
Encoded Proteins	Deleted Genes (chr5)	genes chr17	protein similarity	N0304 FPKM chr17	mean FPKM chr5	mean FPKM chr17	fold diff. chr5-chr17	Ka/Ks
TCP family transcription factor	Glyma05g07943	Glyma17g13065	88%	0.0	0.0	0.0		0.4
MUTS homolog 2		Glyma17g13050		7.0		4.8		
Fragile-X-F-associated protein	Glyma05g07960	Glyma17g13040	98%	8.9	6.1	9.7	1.6	0.3
Unknown	Glyma05g07970	Glyma17g13030	90%	2.3	3.0	2.2	1.4	0.5
Duplicated homeodomain-like protein	Glyma05g07980	Glyma17g13010	94%	2.3	4.7	3.8	1.2	0.3
Histone deacetylase 15 (C-terminus)	Glyma05g07991	Glyma17g13000	95%	2.5	7.4	2.7	2.7	0.3
Histone deacetylase 15 (N-terminus)	Glyma05g08001	Glyma17g13000			5.8			0.2
methyltransferases protein	Glyma05g08011	Glyma17g12985	92%	4.2	3.0	4.2	1.4	0.7
Plant protein of unknown function (DUF936)	Glyma05g08020	Glyma17g12970	81%	1.2	2.2	1.0	2.3	0.2
Unknown	Glyma05g08030	Glyma17g12960	90%	15.4	38.2	14.9	2.6	0.2
DNAJ heat shock N-terminal domain-containing	Glyma05g08040	Glyma17g12950	95%	20.1	12.7	11.9	1.1	0.1
FATB1a (chr5), FATB1b (chr17)	Glyma05g08060	Glyma17g12940	98%	5.5	18.3	5.2	3.5	0.2
Protein phosphatase 2A regulatory B subunit	Glyma05g08070	Glyma17g12930	93%	11.7	13.5	12.2	1.1	0.3
Unknown	Glyma05g08080	Glyma17g12925		0.0	0.0	0.0		
Protein of unknown function DUF92	Glyma05g08090	Glyma17g12920	81%	2.3	0.0	3.0		0.5
Pleiotropic drug resistance 4	Glyma05g08100	Glyma17g12910	98%	0.6	0.0	0.7		0.1
WD40/YVTN repeat-like-containing domain	Glyma05g08110	Glyma17g12900	85%	0.0	1.0	0.2	4.8	0.6
Sodium Bile acid symporter family	Glyma05g08120	Glyma17g12890	96%	0.1	1.4	0.5	2.6	0.3
Leucine-rich repeat protein kinase	Glyma05g08140	Glyma17g12880	87%	4.0	3.7	3.5	1.0	0.2
Average			**91%**	**4.9**	**6.7**	**4.5**	**2.1**	**0.3**

Deletion of FATB1a and Reduced Expression of FAD3A in N0304

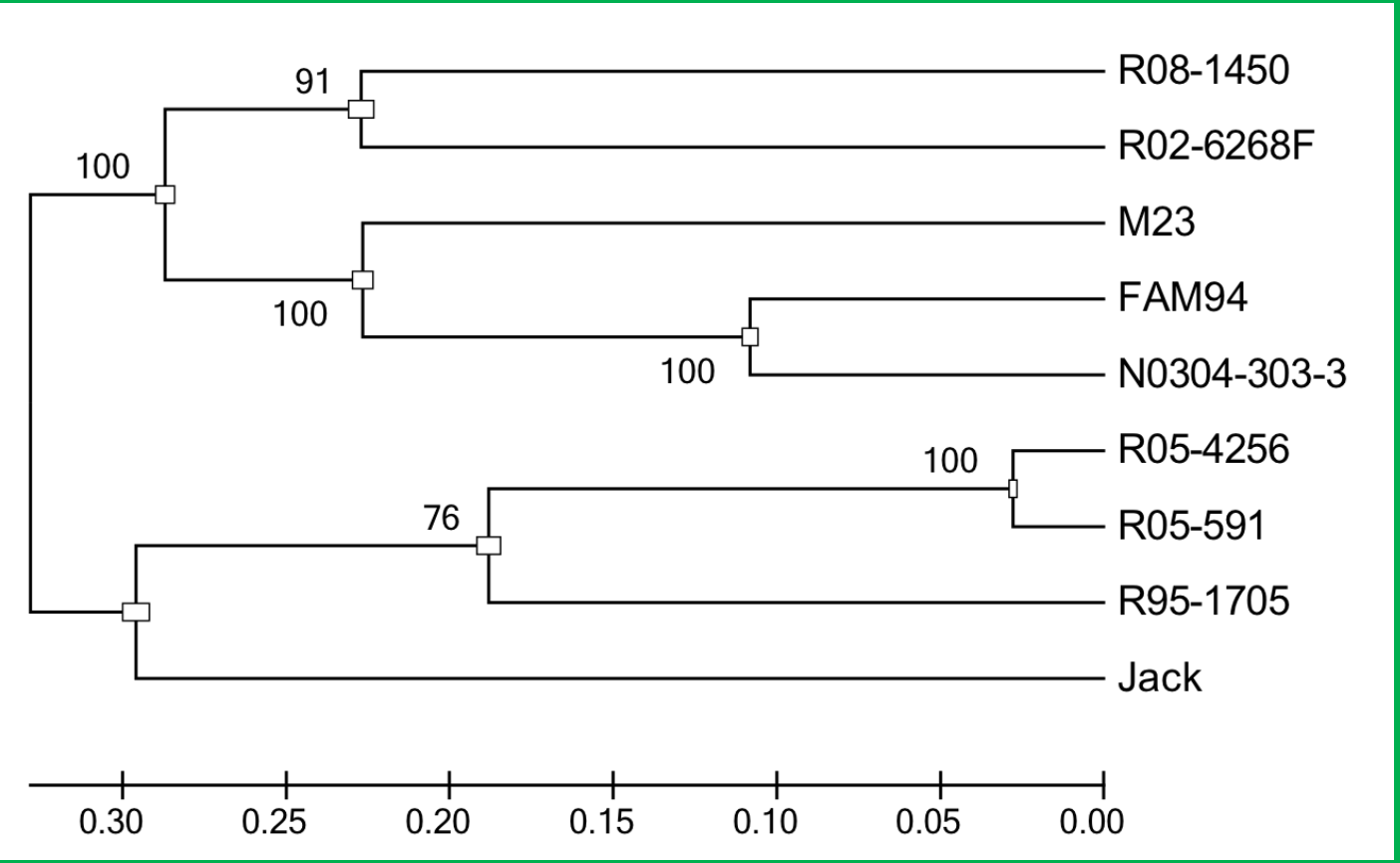


Increased Expression of Clustered Genes at *Rhg1* Locus in Jack

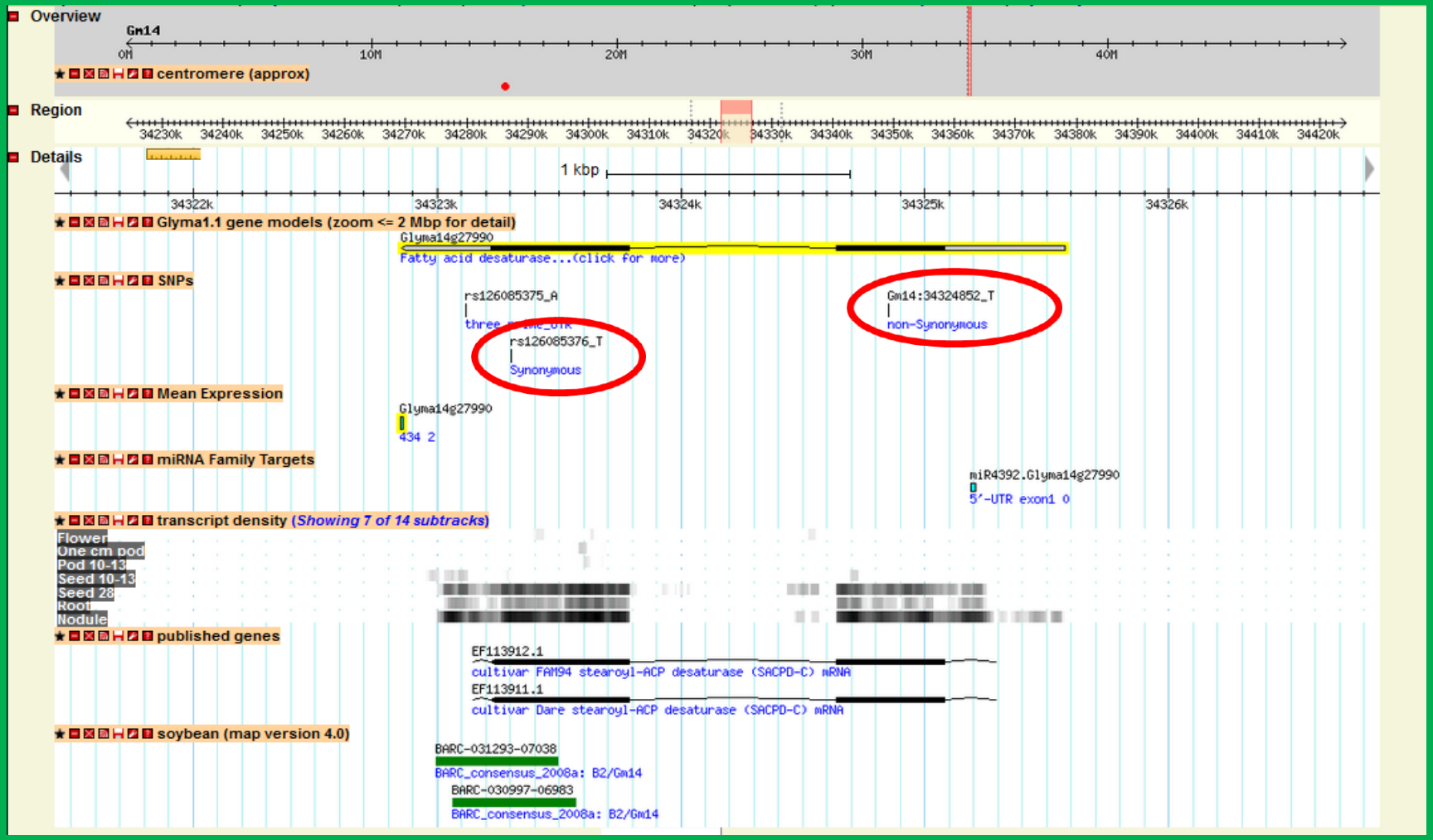




Phylogenetic Relationship

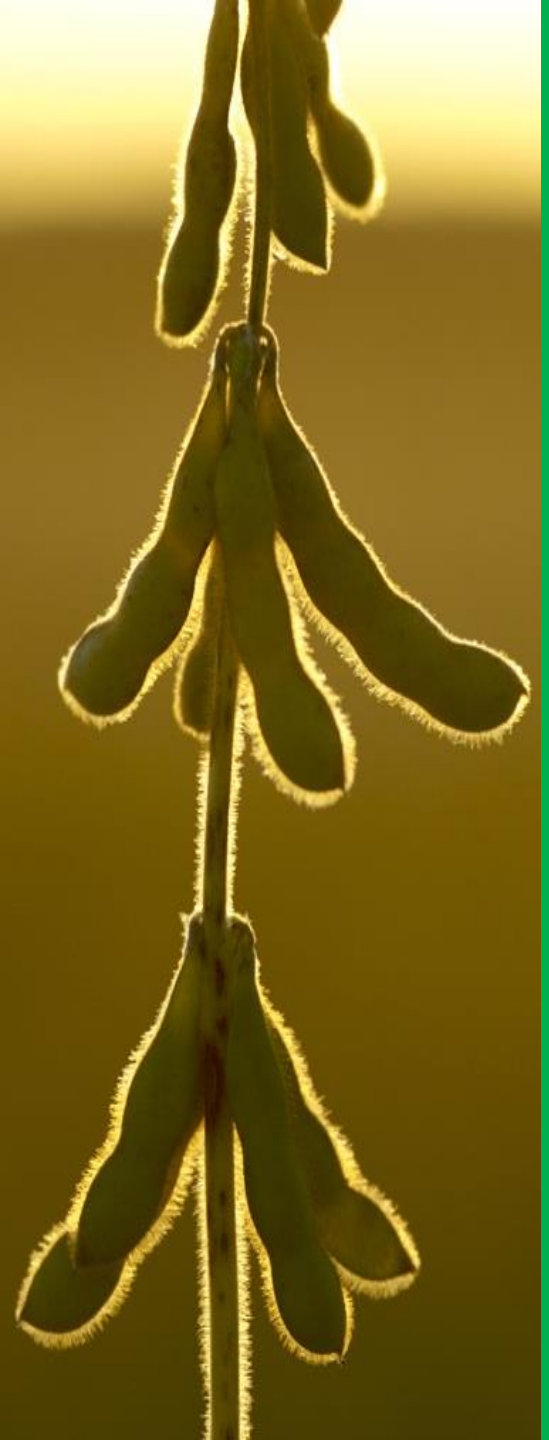


Relational Database and G-Browser to Integrate with Other Data



Take-home Message

- Transcriptome sequencing is an effective approach to identify both transcript sequence (SNPs and Indels) and expression polymorphisms/variations.
- Developed a bioinformatics platform to discover transcription polymorphisms and predict their biological effects.
- Generated a collection of seed transcript polymorphisms as "functional" markers for genetic mapping, or to develop into functional makers for marker-assisted breeding.
- The approach and the collection of seed transcript polymorphisms can be used for other seed traits and non-seed traits.



Acknowledgements:

Wolfgang Goettel
Rick Meyer
Eric Xia
Zhenghai Zhang

Collaboration:

Greg UpChurch/Joe Burton and Earl Taliercio(NC)
Pengying Chen (AK),
Mingli Wang (GA)
Randy Shoemaker (Iowa) and Kristin Bilyeu (MO)

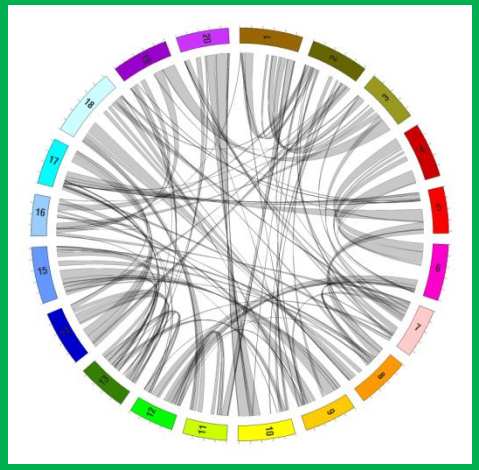
Henry Nguyen (MO), Zongrang Liu (WV) and Sam Wang (MO),

Funding:

USB UNITED SOYBEAN BOARD
Progress Powered by U.S. Farmers

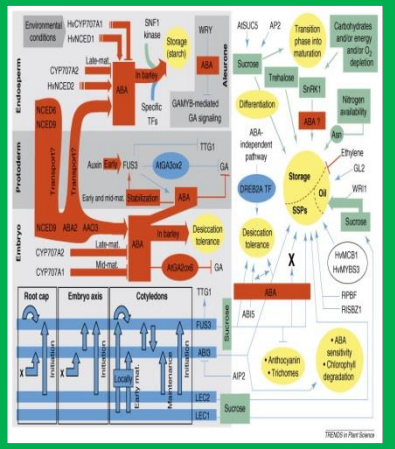
USDA

Genome Engineering to Biological Network Optimization For Seed Quality Improvement



Transcriptome Analysis
of Soybean Genetic Diversity

Transgenic and breeding



Gene Variants and Functional Markers

Breeding: Mix two set of gene variants (gene function variations) and select the best combination of the gene variants that produces the optimized biological networks

Opportunities and Challenges to Biotech and Breeding