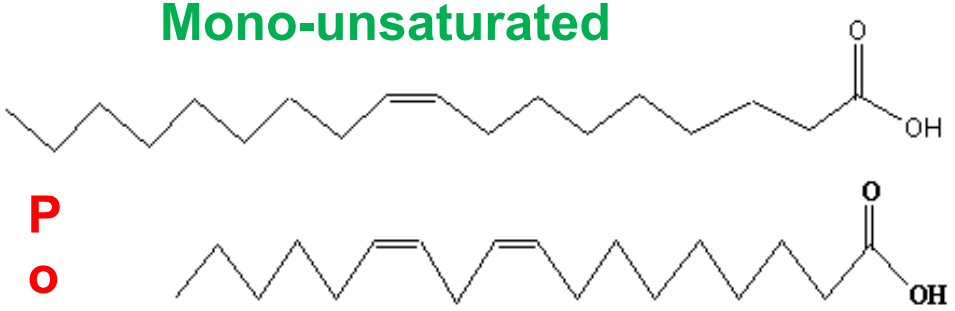
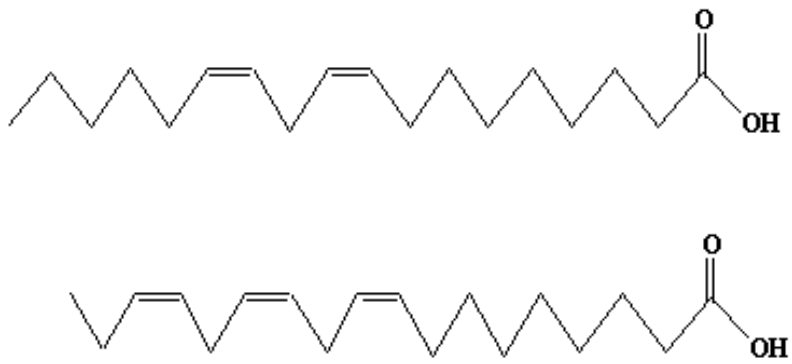
Application of NIR spectroscopy for seed composition improvement in soybean

Jason D. Gillman
USDA-ARS/PGRU
2-14-2017

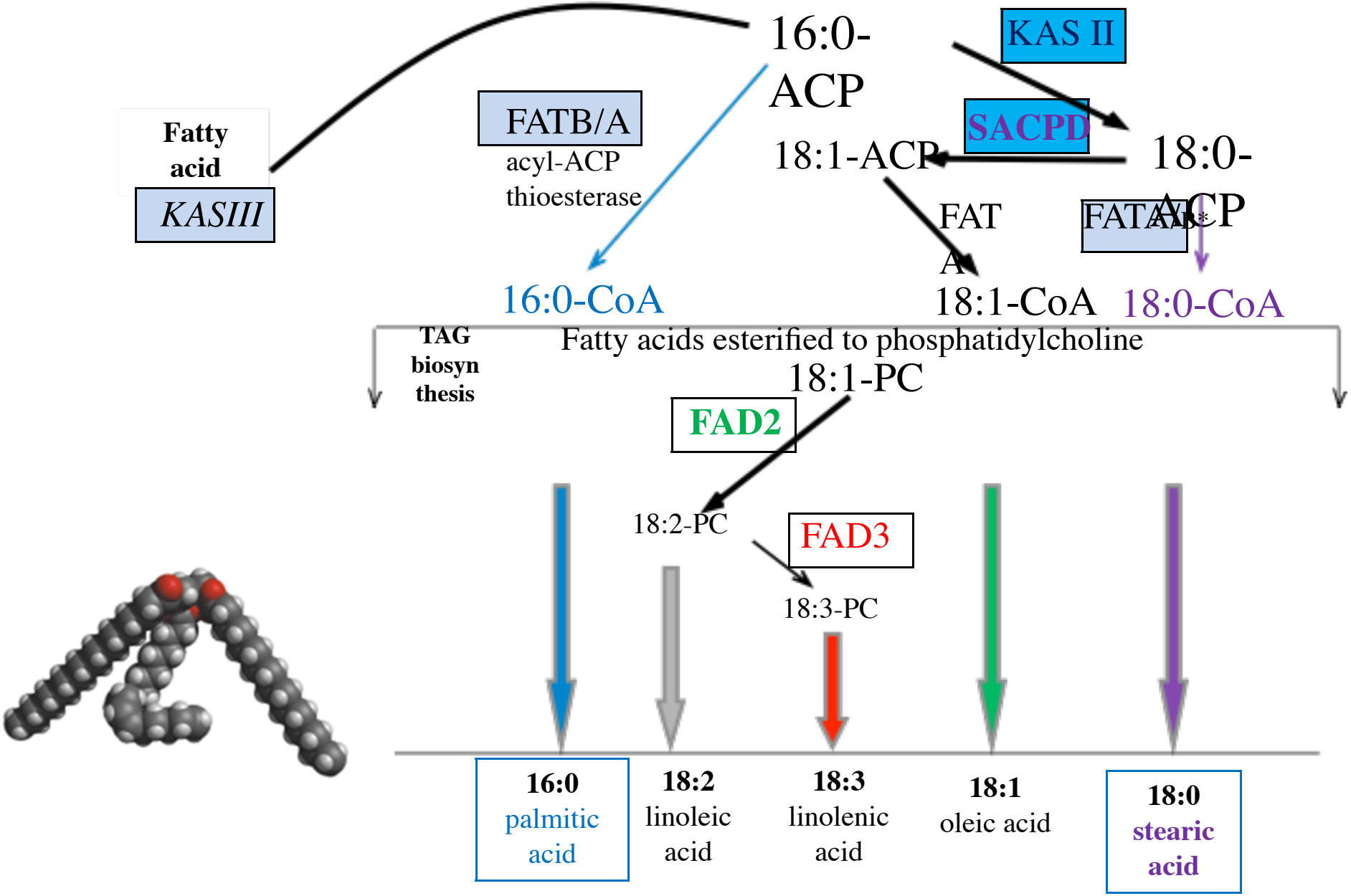
Soybean fatty acid profile





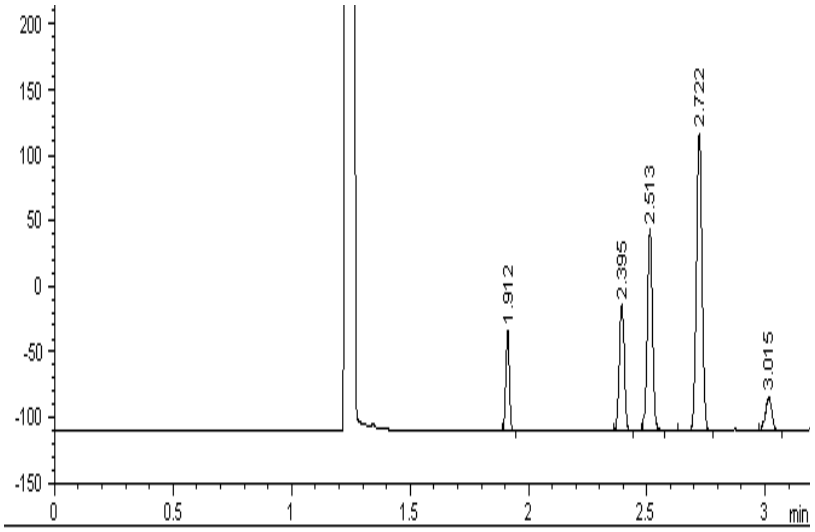
16:0	palmitic	11% ↓
18:0	**stearic**	4% ↑
18:1	oleic	24% ↑
18:2	linoleic	54% ↓
18:3	linoleNic	7% ↓

Very simplified biochemistry of seed oil synthesis

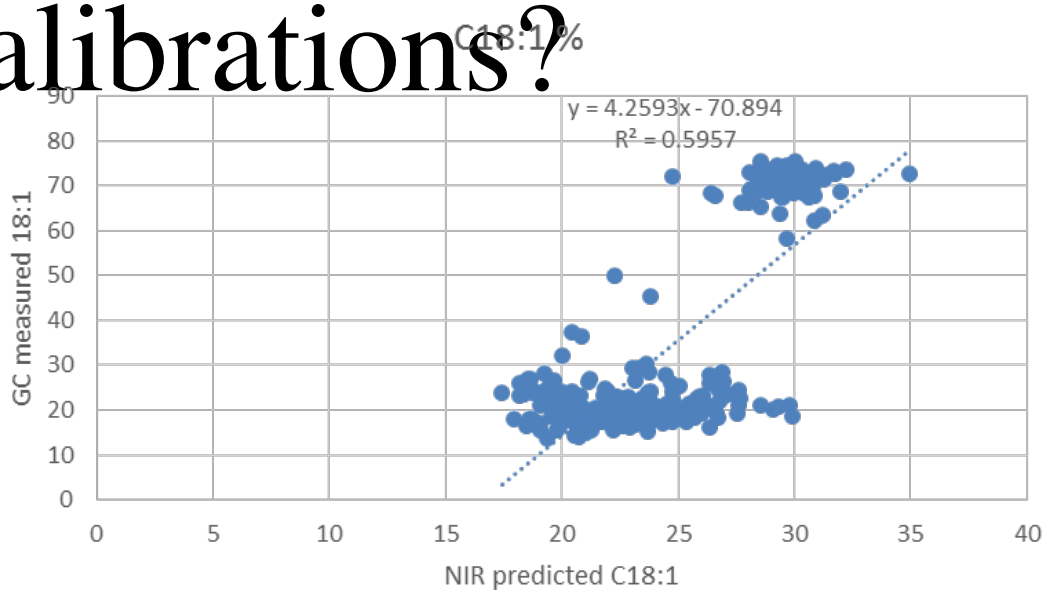


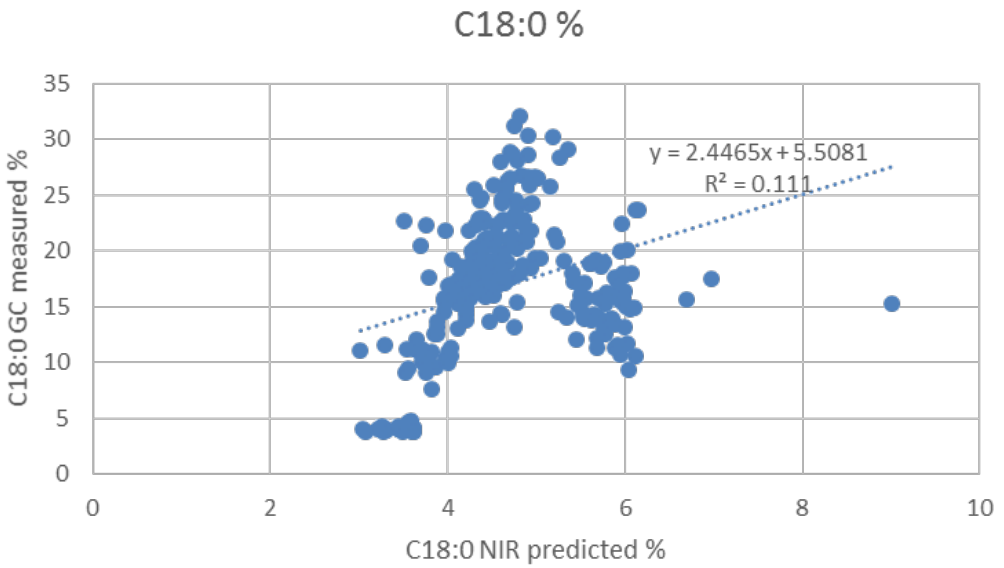
Soybean seed lipid pool is almost completely (~88%) in the form of triacylglycerols

Why go to the bother to create new NIRS calibrations?



Gas chromatography analysis
destructive assay
Relatively slow/non-automatable

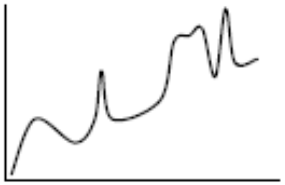




Steps involved during NIR calibration

Identify/produce seed covering a broad phenotypic range, preferably with replication

Spectral Data
X



Regression algorithm

Mathematical relationship (calibration model)

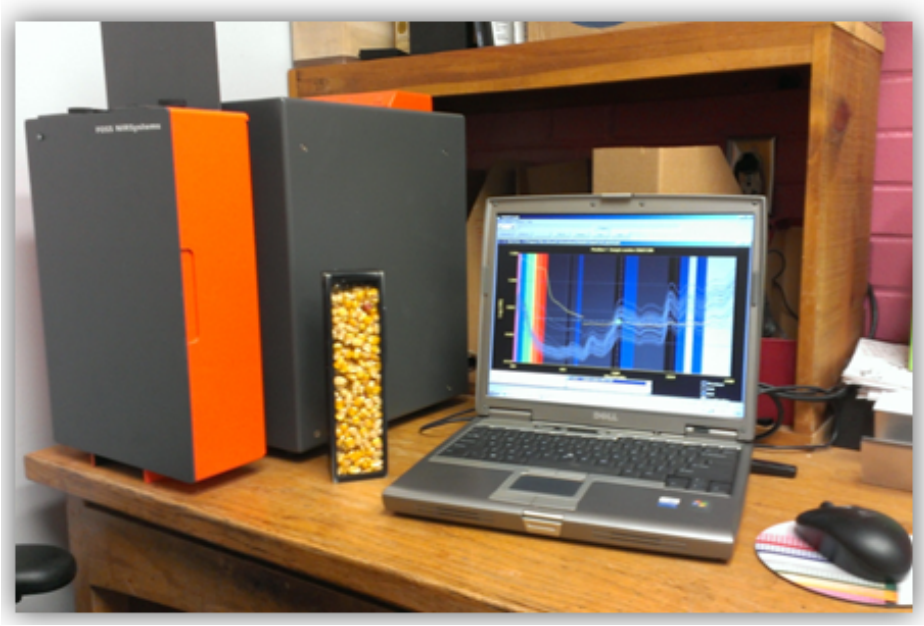
$Y=f(X)$

Constituent concentration
Y

(Obtained by standard wet chemistry methods)

Phenotyping seed/kernel composition traits

- FOSS 6500 Near Infrared Reflectance spectroscopy
- ~50-100 whole seeds per field plot, all plots in RBCD triplicate
- Scan time ~30 seconds



FOSS® 6500 NIR Instrument

Partial Least Squares 1
(UnScrambler® software)

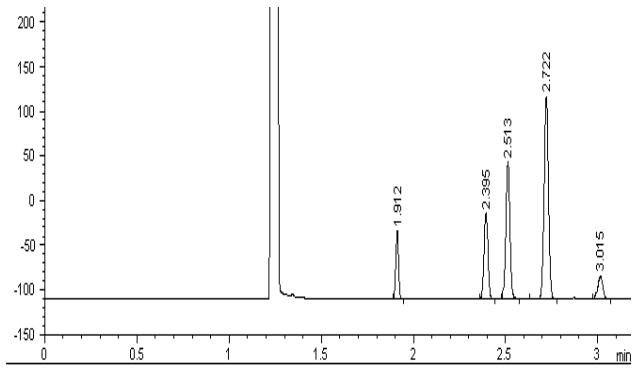
Mention of a trademark, vendor, or proprietary product does not constitute a guarantee or warranty of the product by the USDA and does not imply its approval to the exclusion of other products or vendors that may also be suitable.

Step 1A: Identify and phenotype appropriate samples

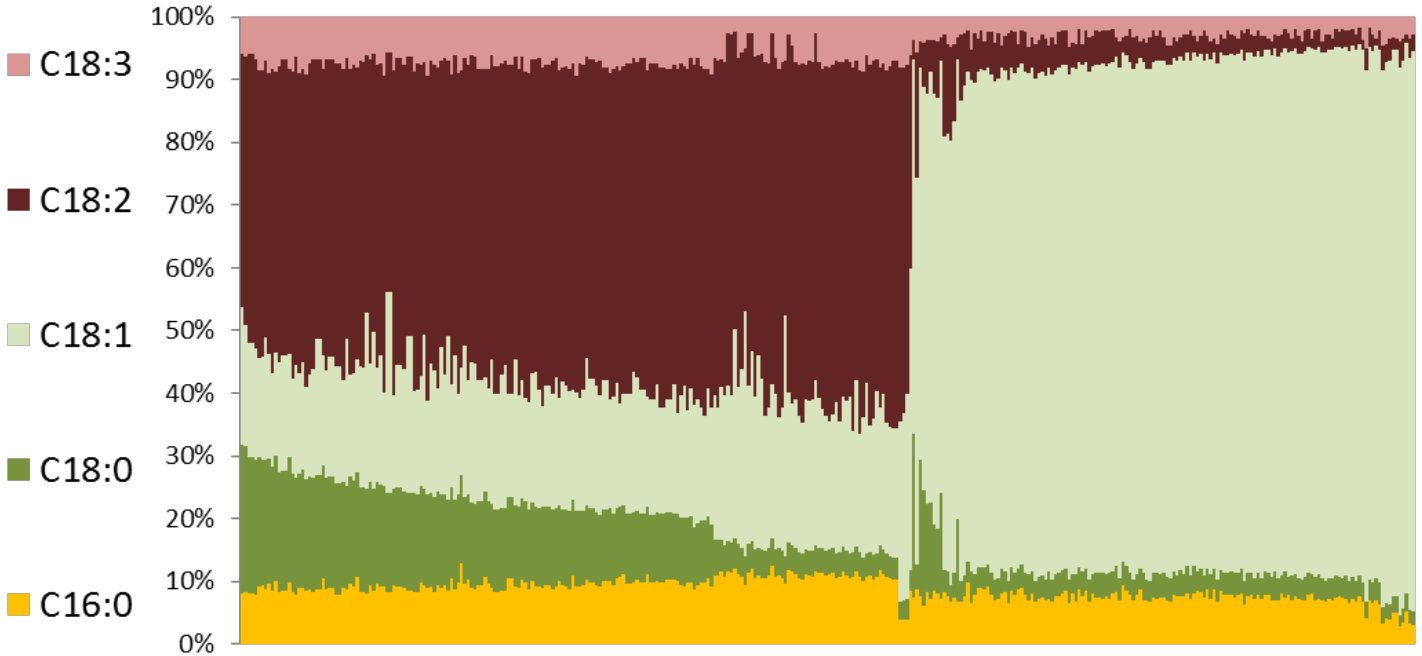
- Elevated Stearic acid 194d x A6 (~11% x 24%)
  - 176 RILs x 3 plot replicates
- Elevated Oleic acid (>60%) and low linolenic acid (<6%)
  - A few had unique combinations (e.g. >10% stearic acid/>70% oleic acid)
  - 23 RILS x 3 replicates in two locations
  - 16 additional RILS x 3 replicates in only one
- Various single mutant lines (↑stearic, ↑↓oleic, ↓palmitic, ↓linolenic)
  - Single replicates across multiple years
- Wild type lines (8) across a range of gene backgrounds and maturities
  - Multiple replicates across multiple years

- Wet chemistry/GC analysis: triplicate per plot

Gas chromatography analysis
destructive assay
Relatively slow/non-automatable



Materials we used



Fatty Acid	n	mean	SD	CV	Range	Difference
C16:0	687	8.91	1.32	0.05	2.78 - 12.62	9.84
C18:0	687	12.30	6.25	0.24	1.85 - 28.04	26.19
C18:1	687	34.65	24.62	0.94	16.05 - 89.44	73.39
C18:2	687	38.43	18.37	0.70	1.24 - 58.66	57.42
C18:3	687	5.71	1.63	0.06	1.75 - 9.45	7.11

Table 1 Fatty Acids measurement of all the Soybean samples in the NIR calibration

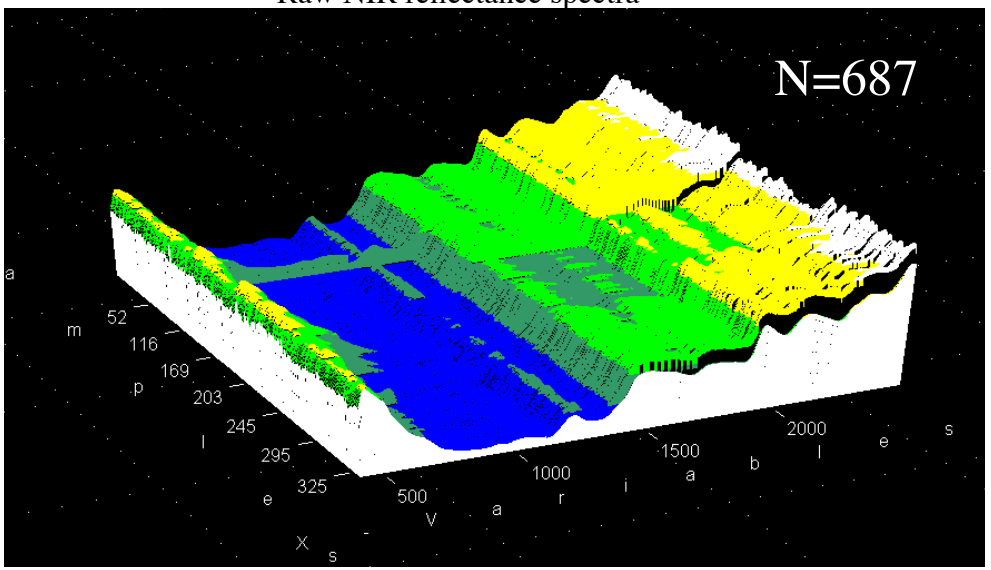
Number of samples (n); Standard deviation (SD); coefficient of variation (CV)

A Karn, C. Heim, S. Flint-Garcia, K. Bilyeu, K.; J. Gillman, J., *JAOCS* ***(2017) 94, 69-76.***

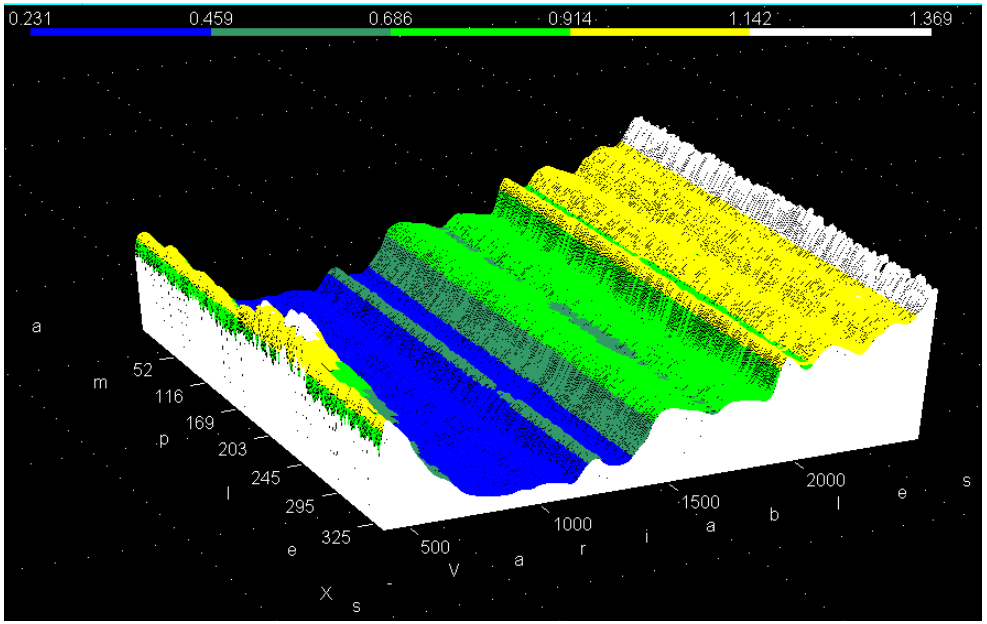
Step 1B. Collecting NIR reflectance data

- Spectra collected from wavelength 400nm – 2490nm with the increment of 10nm
- Removed spectra below 900nm
- Collected spectra were treated with Multiple Scatter Correction (MSC) and 1st derivative (one was also treated with 2nd derivative)
- Why? Reduced the noise caused due to spectral scattering and increase signal intensity

Raw NIR reflectance spectra

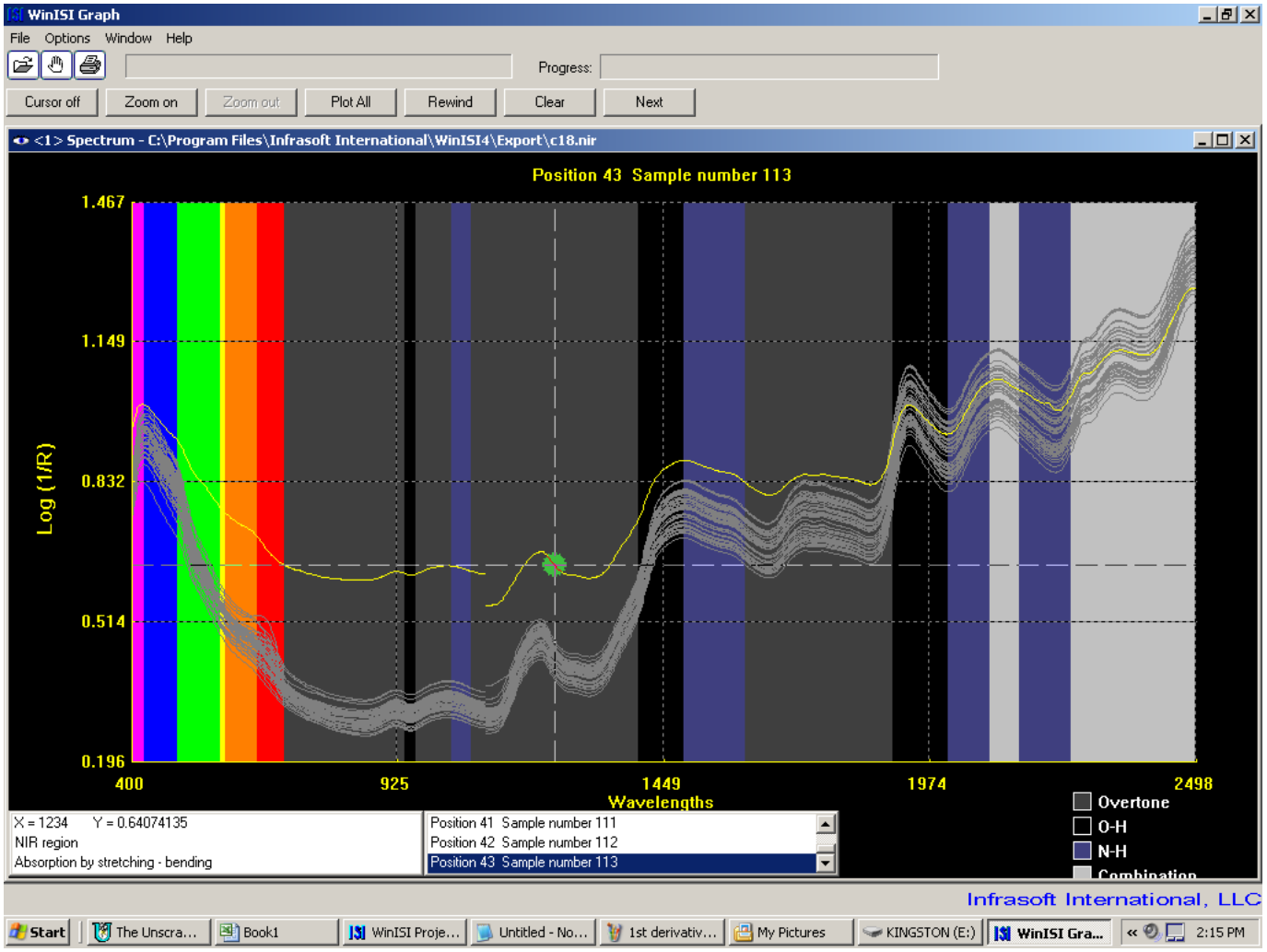


MSC of NIR reflectance spectra

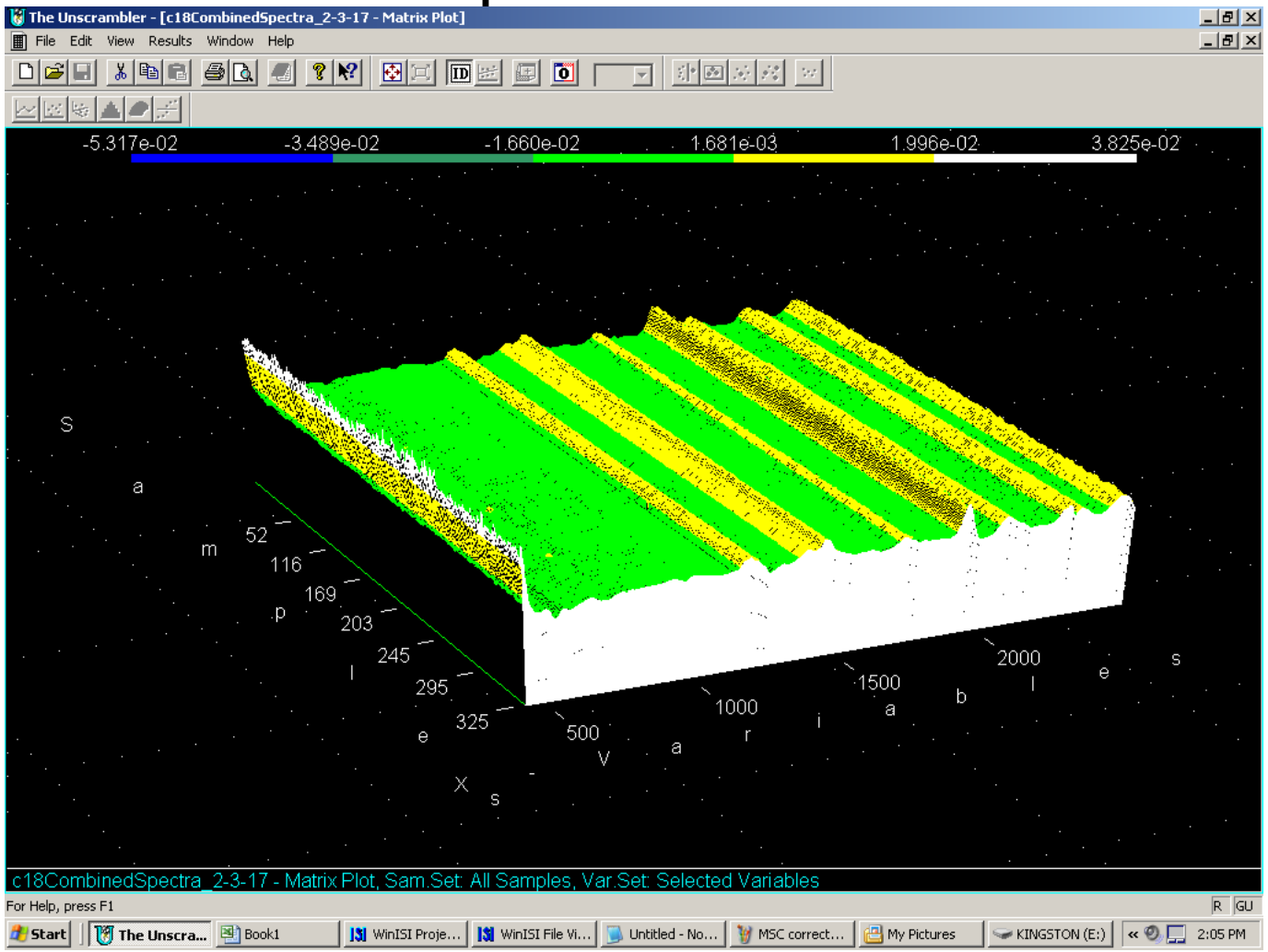


A Karn, C. Heim, S. Flint-Garcia, K. Bilyeu, K.; J. Gillman, J., *JAOCS* ***(2017) 94, 69-76.***

Critical: Inspect your data for outliers

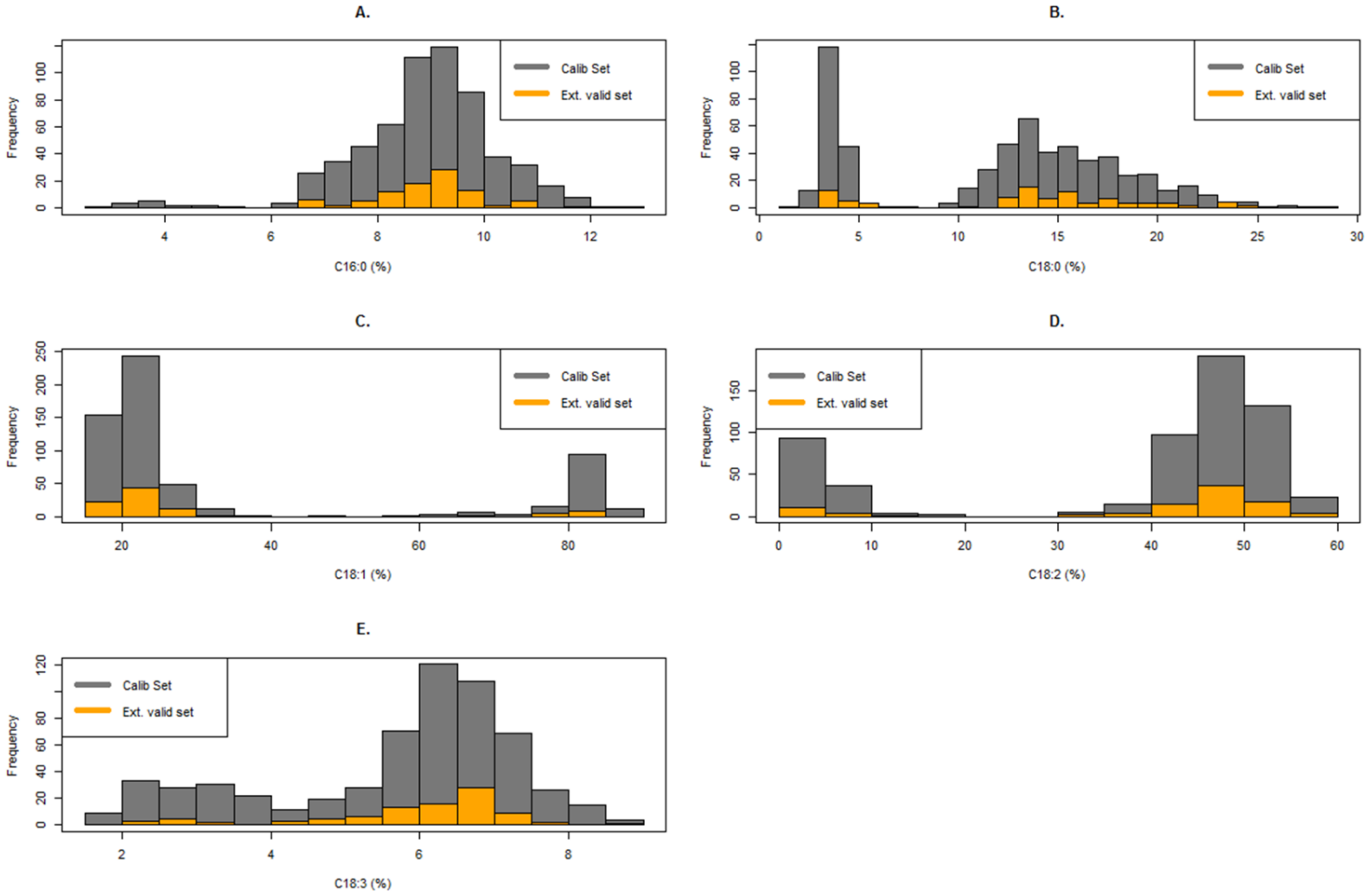


We tested several mathematical processing steps of spectral data



MSC and 1st derivative dramatically improved our predictions
(for one we also did a 2nd derivative)

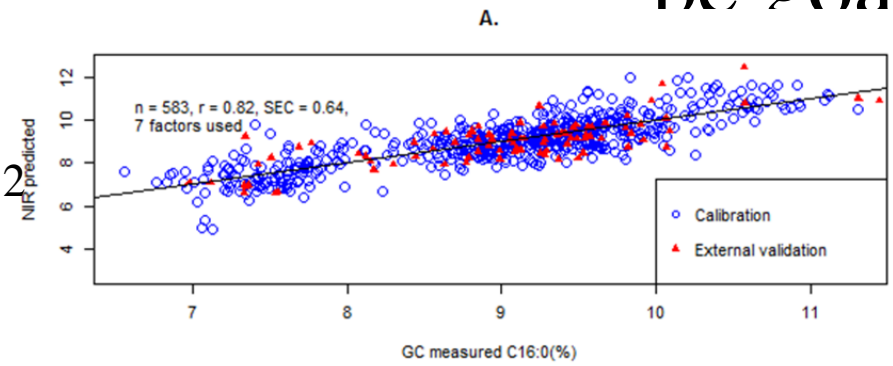
A broad multiply replicated range of phenotypes were incorporated into the NIR calibration

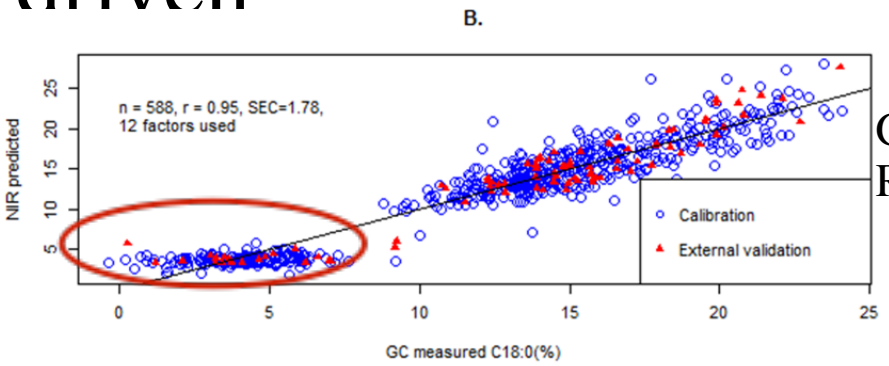


A Karn, C. Heim, S. Flint-Garcia, K. Bilyeu, K.; J. Gillman, J., *JAOCS* ***(2017) 94, 69-76.***

Error and accuracy are relative and models should be goal driven

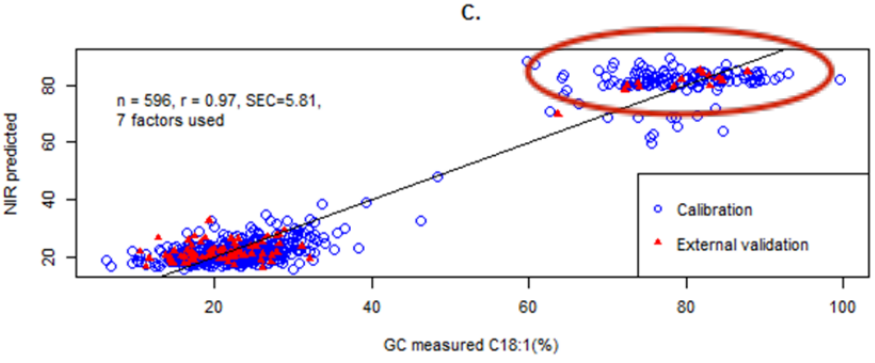
C16:0
R=0.82

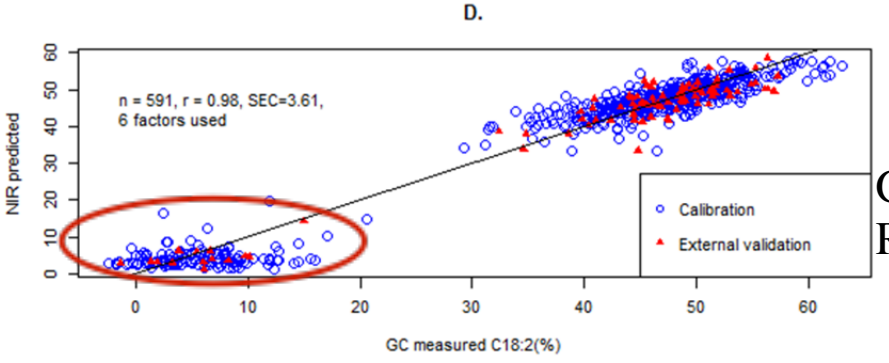




C18:0
R=0.95

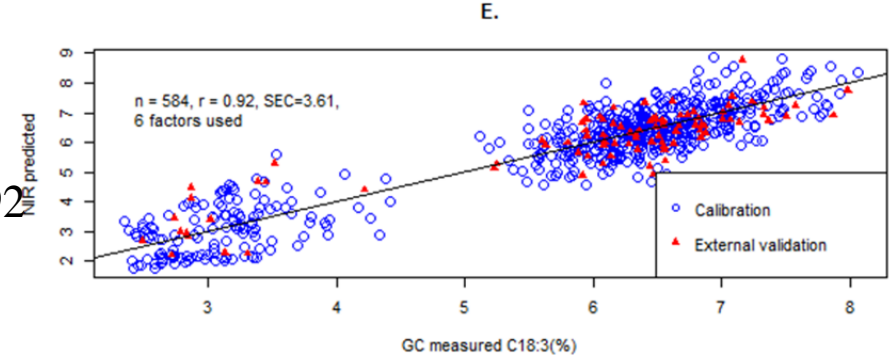
C18:1
R=0.97





C18:2
R=0.98

C18:3
R=0.92



Certain regions the linear relationship falls apart
E.g. low end stearic and high end oleic

A Karn, C. Heim, S. Flint-Garcia, K. Bilyeu, K.; J. Gillman, J., *JAOCS* ***(2017) 94, 69-76.***

We split all samples into two sets – calibration and

Fatty Acids	n	Spectral range	NIR Pretreatment	PLS Factors	SEC	SECV	RMSECV	r
C16:0	583	900 - 2500 nm	MSC; 1 Der	7	0.64	0.67	0.67	0.82
C18:0	588	900 - 2500 nm	MSC; 1 Der	12	1.78	2.17	2.17	0.95
C18:1	596	900 - 2500 nm	MSC; 1 Der	7	5.81	6.14	6.14	0.97
C18:2	591	900 - 2500 nm	MSC; 1 Der	6	3.61	3.73	3.73	0.98
C18:3	584	900 - 2500 nm	MSC; 1 Der	6	0.64	0.66	0.66	0.92

Tab
Nur
vali

Table 3 External validation statistics in NIR models for the estimation of individual fatty acids
Number of samples (*n*); standard error of performance (SEP); Root mean square error for prediction (RMSEP);
Ratio of standard deviation of data to standard error of performance (RPD);
coefficient of correlation (*r*), t-test statistic.

Fatty Acid	n	Mean	Range	SD	SEP	RMSEP	RPD	r	t Sta
C16:0	93	8.97	6.58 – 12.44	1.04	0.66	0.65	1.57	0.77	0.74
C18:0	93	13.45	3.24 – 27.57	6.17	1.85	1.84	3.34	0.95	-0.3
C18:1	93	31.33	16.5 – 84.95	22.00	4.89	4.99	4.50	0.97	-2.2

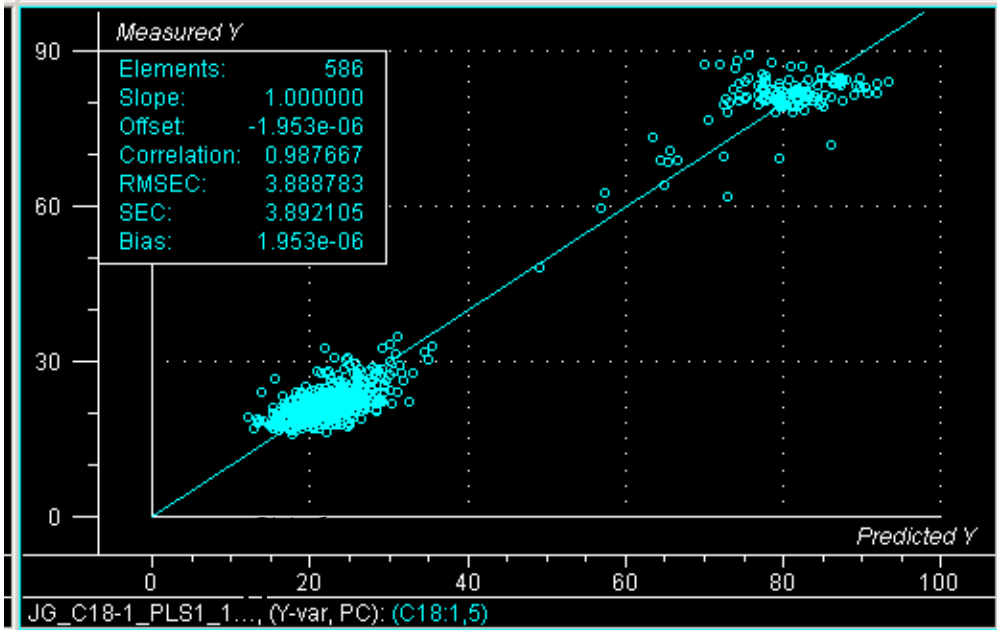
External validation (of at least a subset) is very important when applying calibration on external samples

C18:1 %

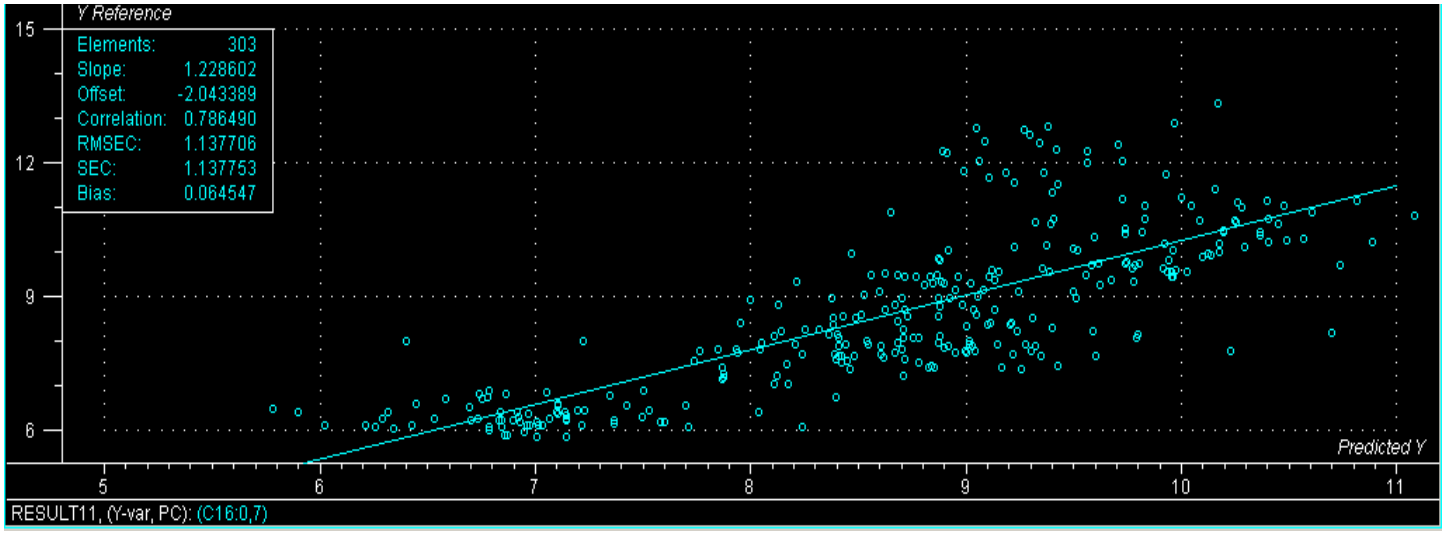
Very predictive due to:

high concentration in seeds

large phenotypic differences

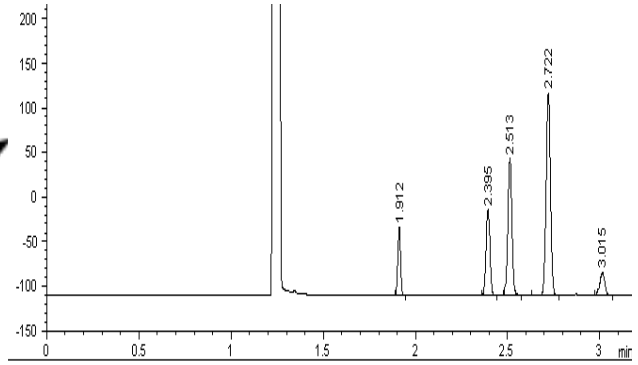


C16:0 %



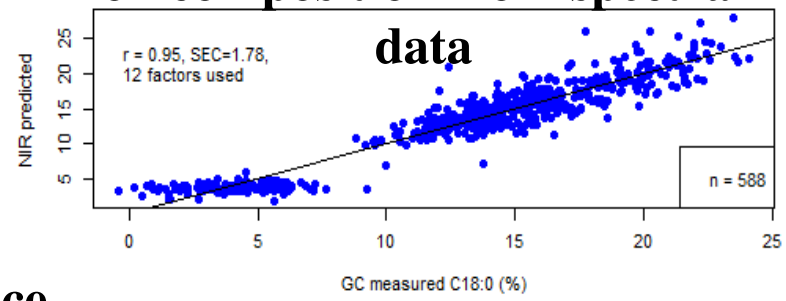
Genes not in calibration set + low end “flatness” = lower correlation coefficient

Gas chromatography analysis
destructive assay
slow/non-automatable

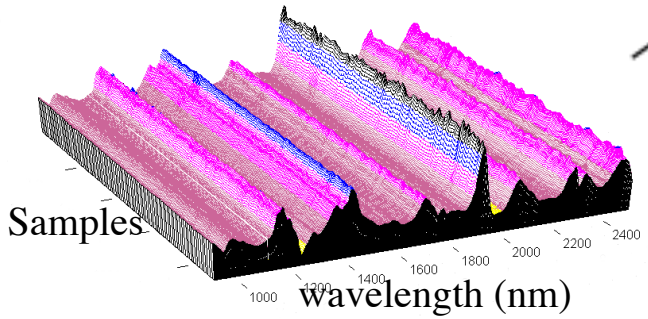


soybean seed

NIRS model accurately predicts oil composition from spectral data



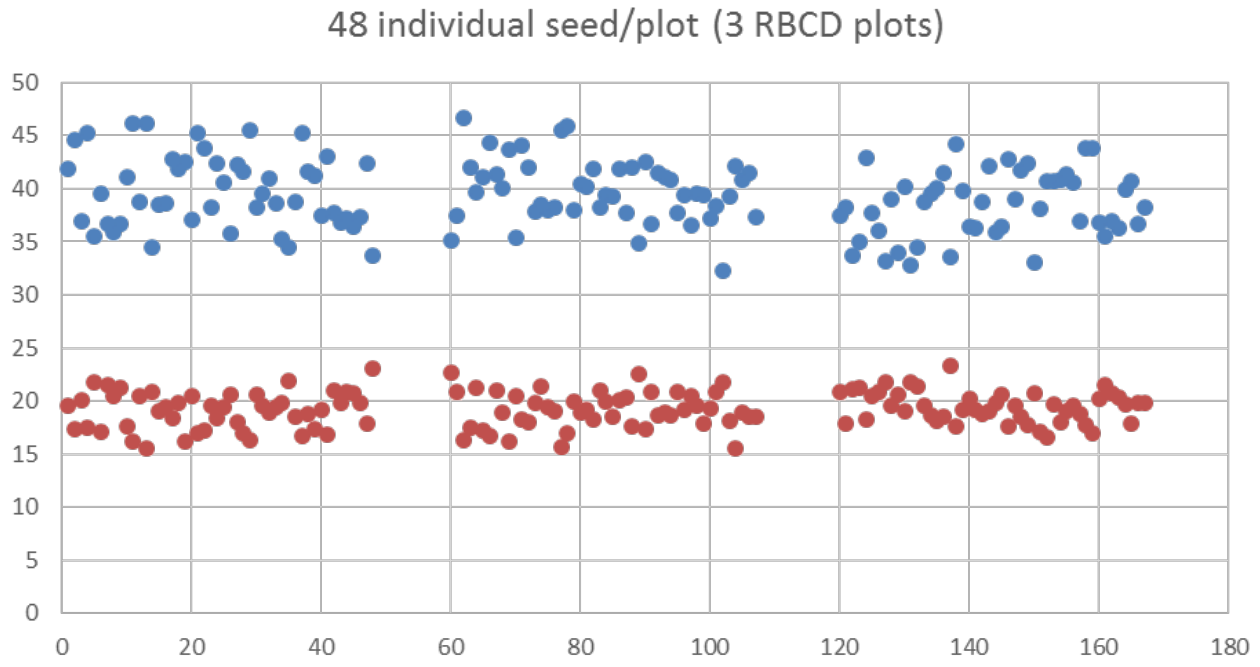
Near Infrared Reflectance
non-destructive
rapid/automatable



A Karn, C. Heim, S. Flint-Garcia, K. Bilyeu, K.; J. Gillman, J., *JAOCS* ***(2017) 94, 69-76.***

Sources of variance in seed quality traits

Genotypic effects
Location effects
Year effects
Replication effects (plot x plot)
Plant x Plant (often ignored)
Seed on a plant (often ignored)



Protein content %

Oil content %

Single seed NIR prediction calibrations have been previously velo

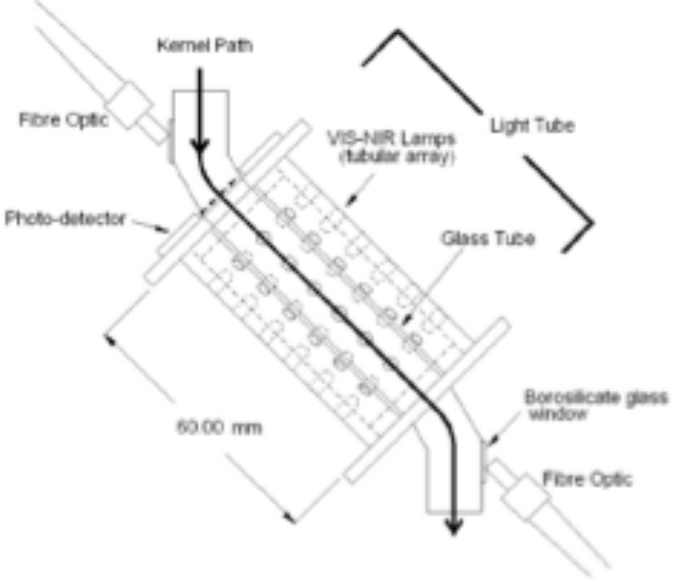


Figure 1. Component assembly used for spectral measurements.

Table 3. PLS Calibration and Validation Statistics for Predicting Soybean Seed Traits with Single Seed NIR

			cross-validation		external validation		
seed trait	spectral pretreatment	factors	R^2	RMSEP	R^2	RMSEP	RPD
% oil	MSC	10	0.98	0.54	0.97	0.47	5.67
% protein	MSC	9	0.84	1.53	0.84	1.48	2.28
density (g/cm^3)	first der.	10	0.72	0.06	0.35	0.07	0.91
weight (mg)	none	10	0.97	9.59	0.94	9.80	5.21
volume (mm^3)	none	9	0.96	8.53	0.94	8.21	4.33
max area (mm^2)	first der.	7	0.84	0.03	0.82	0.03	2.31
length (mm)	first der.	3	0.68	0.49	0.62	0.50	1.68
width (mm)	second der.	7	0.79	0.41	0.65	0.37	1.74
% air space	none	13	0.79	1.62	0.45	1.79	1.25

2 locations
9 genotypes
3 plots/genotype (RBCD)
24 or 48 seed/plot

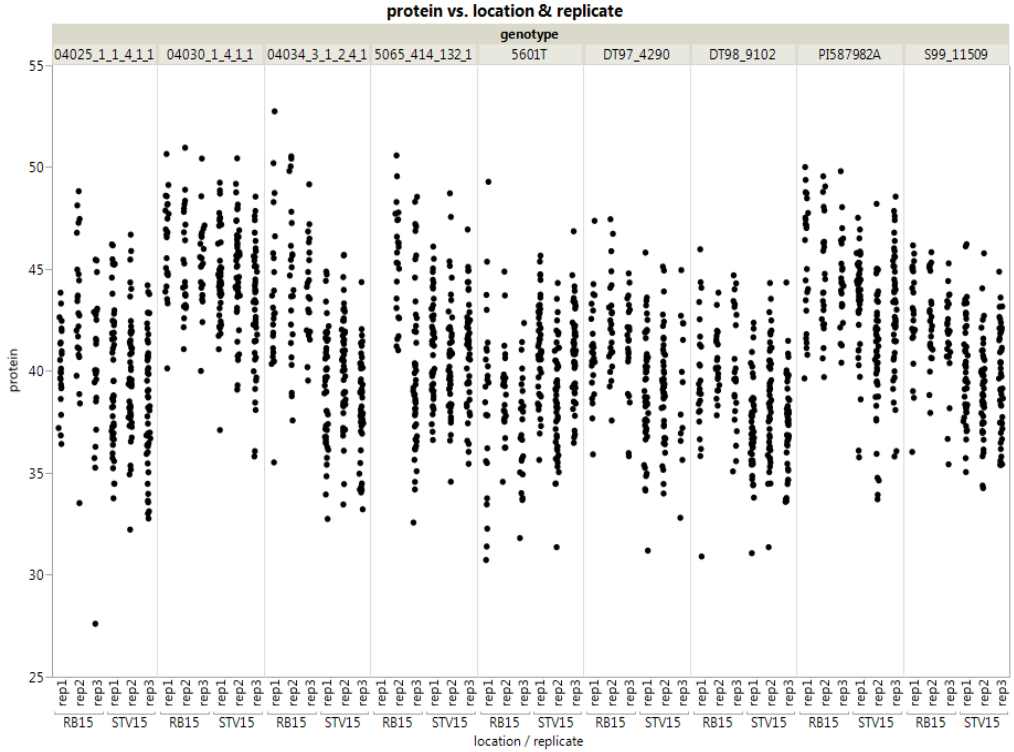
Individual seed were run through the instrument
3 times and spectra were averaged

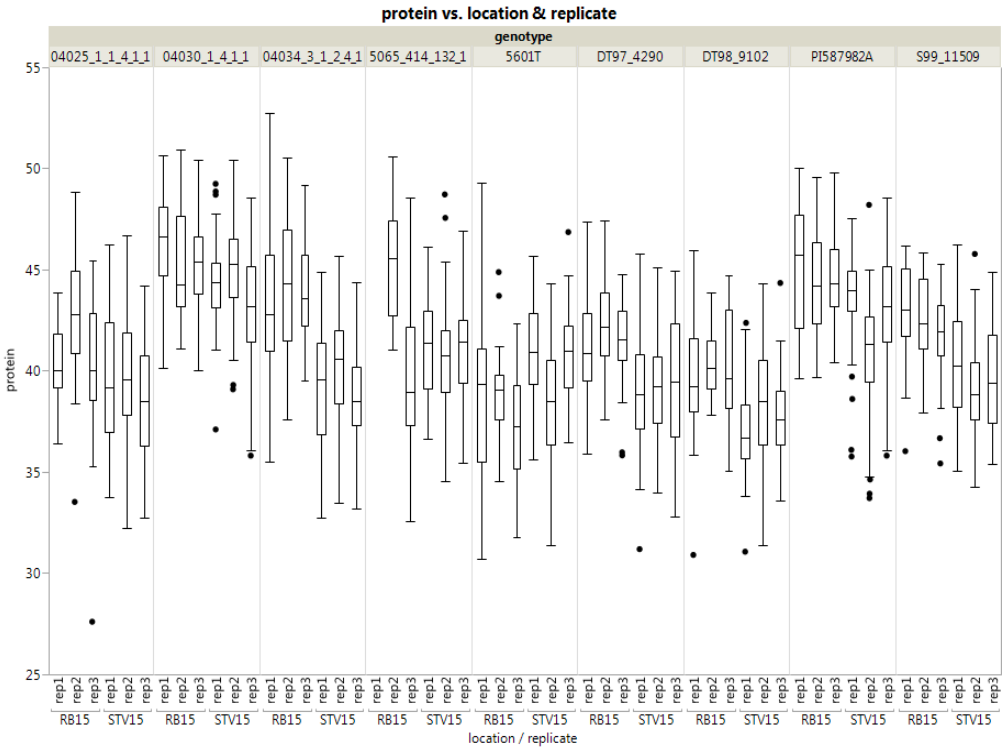
Paul
Armstrong
USDA-ARS

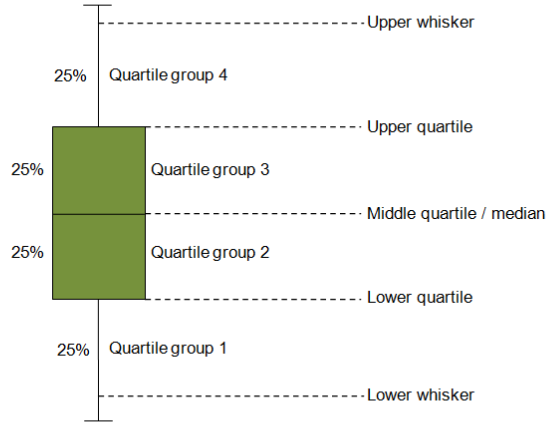
P. R. Armstrong, J. Tallada, C. Hurburgh, D. Hildebrand, J. Specht
(2011) *ASABE* **54**, 1529-1535

DOI: 10.1021/acs.jafc.5b05508
J. Agric. Food Chem. 2016, 64, 1079−1086

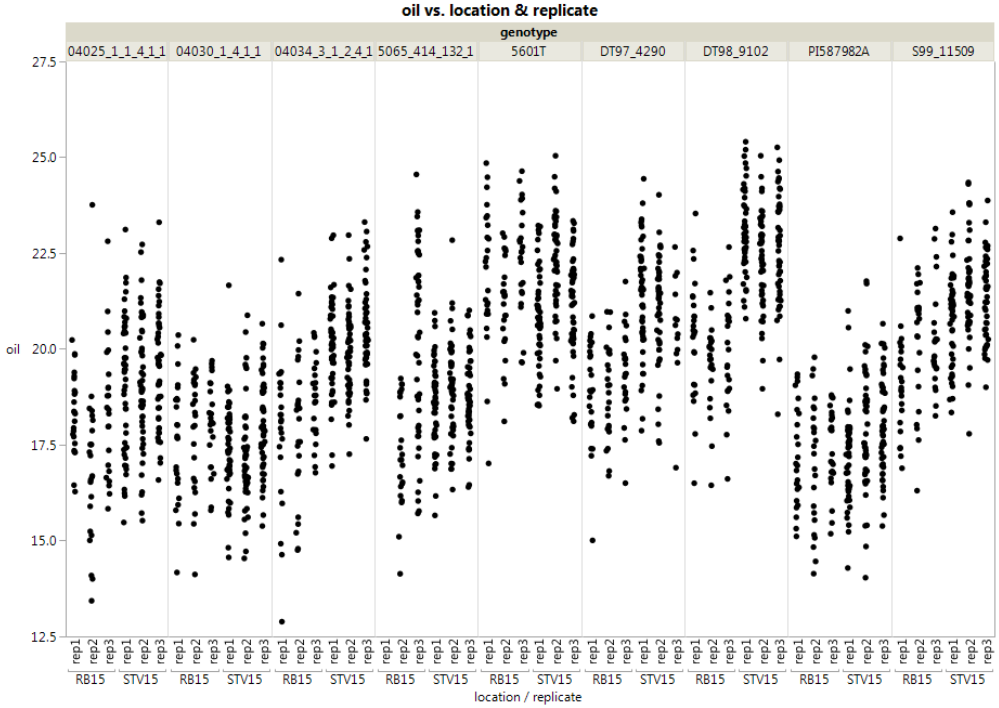
There is considerable within-plot variation in soybean for % seed protein

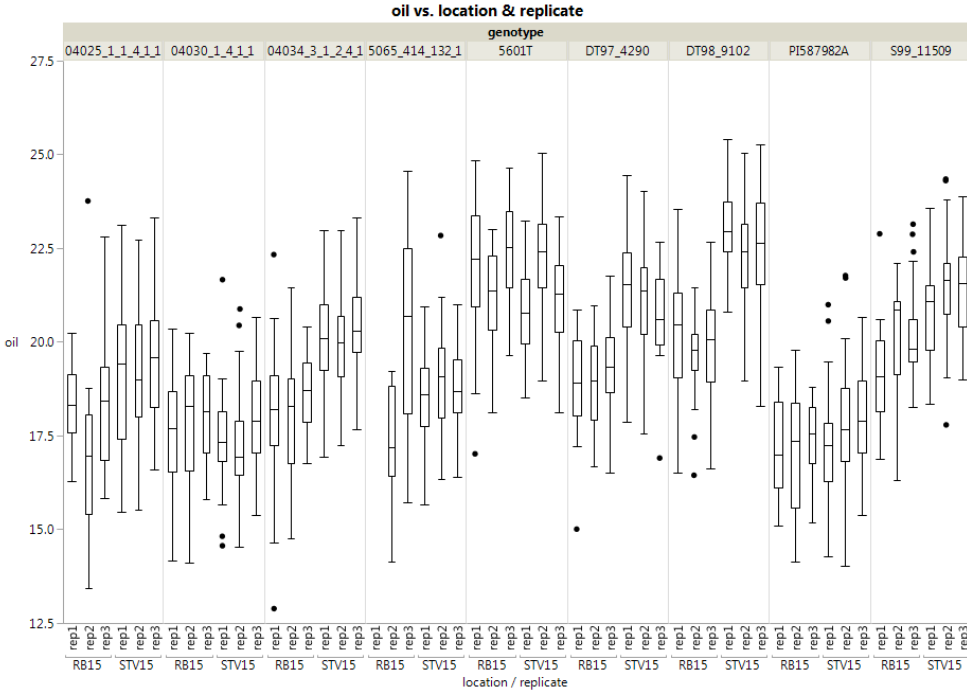


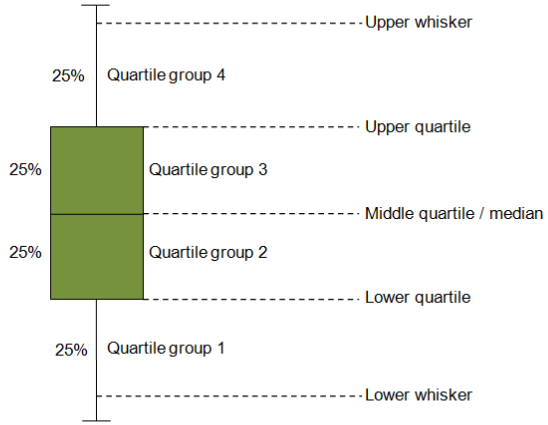




There is considerable within-plot variation in soybean for % seed oil

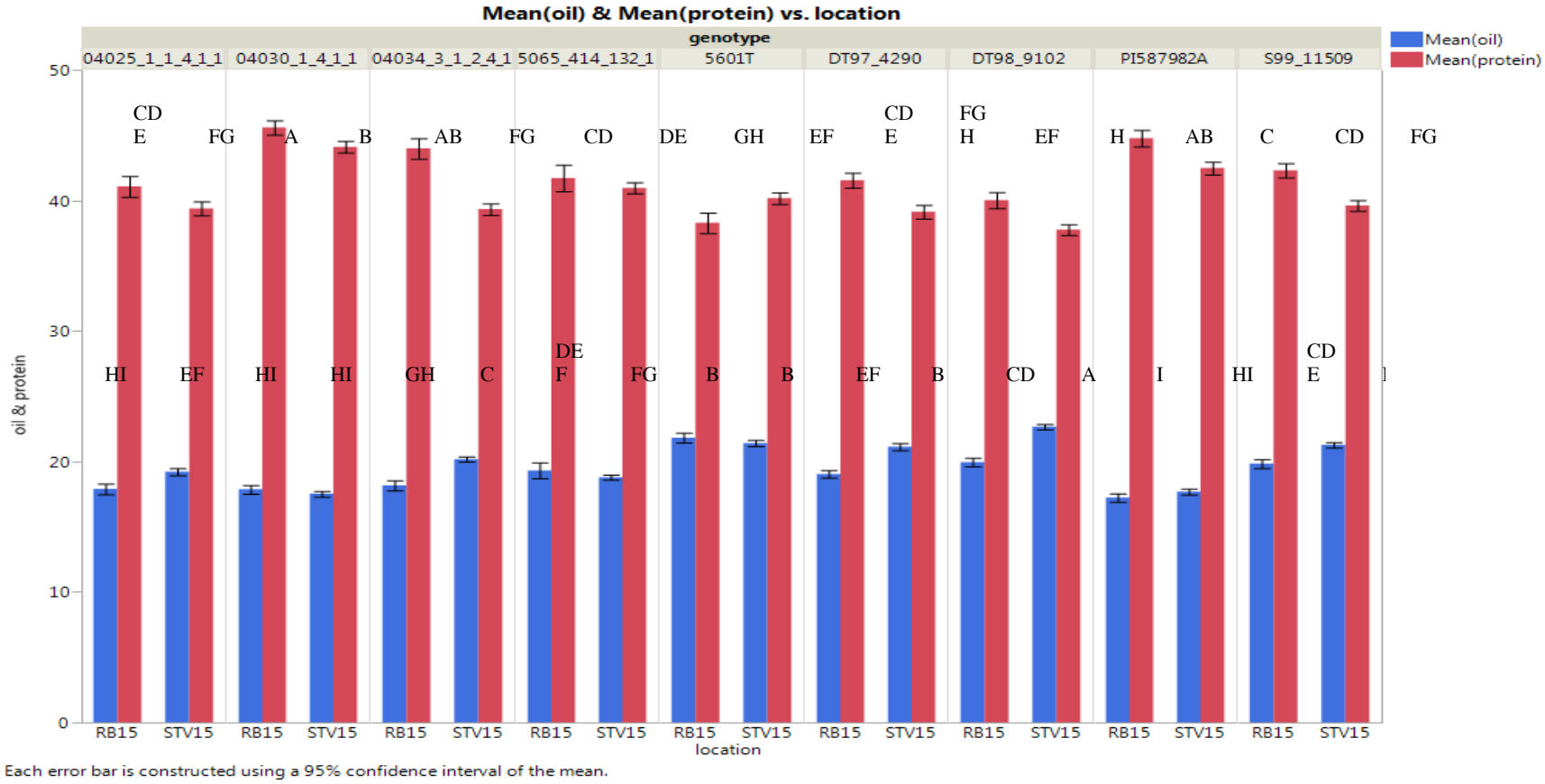






But with enough seeds it's possible to detect entry, location and (entry x location) differences

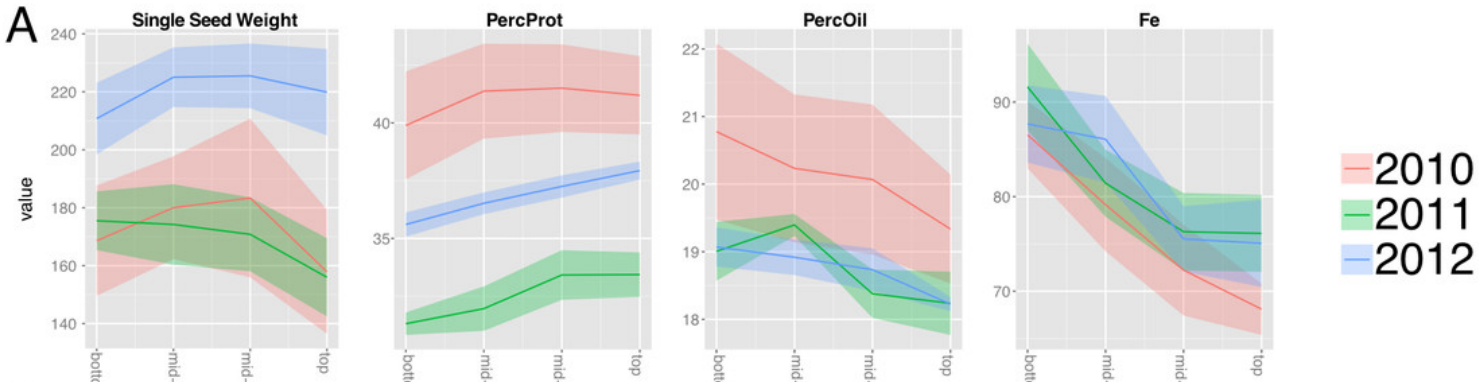
- (n=24 for RB2015, n=48 for STV2015)



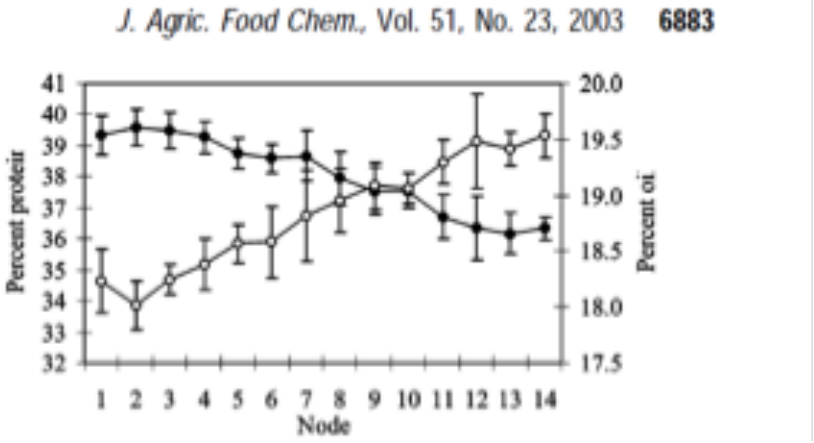
Anova/HSD overlapping letters indicate insignificantly different means (α=0.05)

Within-plot variance is most likely driven by canopy position based variation

Huber et al. (2016) PeerJ **4**:e2452



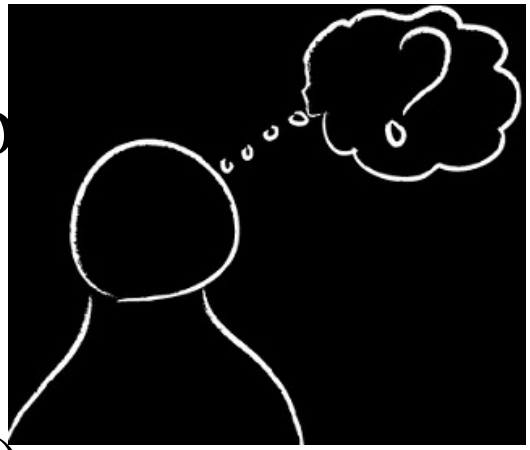
Bennet et al. (2003) JAFC **51**:6882-6886



- Protein is higher closer to the base of the plant
- Oil is higher closer to the top of the canopy
- Ionomic components are also affected

Acknowledgements and questions?

- Crystal Buerke Heim (grad student, Univ. of Missouri)
- Avinash Karn (grad student, Univ. o
- Germplasm/cultivars
  - Dr. Kristin Bilyeu
  - Dr. Walter Fehr (Iowa State, emeritus)
  - Dr. Andrea Cardinal (formerly NCSU)
  - Dr. Toyoaki Anai (Saga University, Japan)
  - Dr. David Sleper (Univ. Missouri, emeritus)
  - USDA GRIN



Seed oil modification	Gene	Mutant Allele	Mutant (cultivar)	Reference
Elevated oleic acid C18:1↑ Range (16.1 - 89.4%)	FAD2-1A	S117N	17D (W82)	(Dierking and Bilyeu 2009)
	FAD2-1A	indel	PI603452	(Pham, Lee et al. 2010)
	FAD2-1B	P137R	PI283327	(Pham, Lee et al. 2010)
	Unknown	Unknown	FA8077	(Graef, Miller et al. 1985)
Reduced linolenic acid C18:3↓↑ Range (1.75 - 9.5%)	FAD3A	splice	CX1512-44	(Bilyeu, Palavalli et al. 2005)
	FAD3A	W266*	C1640 (Century)	(Chappell and Bilyeu 2006)
	FAD3A	indel	PI361088B	(Chappell and Bilyeu 2007)
	FAD3C	G128E	CX1512-44	(Bilyeu, Palavalli et al. 2005)
Reduced palmitic acid C16:0↓(2.78 - 12.62%)	FATB1A	W231L	A22	(De Vries, Fehr et al. 2011)
	KAS3	Splice defect	C1726 (Century)	(Cardinal, Whetten et al. 2013)
Elevated stearic acid C18:0↑ Range (1.85 - 28.04%)	SACPD-C	P286L	RG8 (C1640/Century)	(Boersma, Gillman et al. 2012)
	SACPD-C	V211E	194D (W82)	(Gillman, Stacey et al. 2014)
	SACPD-C	Indel	M25 (Bay)	(Mizanur, Takagi et al. 1995, Gillman, Stacey et al. 2014)
	SACPD-C	deletion	A6 (unknown)	(Hammond and Fehr 1983, Gillman, Stacey et al. 2014)
	SACPD-C	deletion	MM106 (Bay)	(Mizanur, Takagi et al. 1995, Rahman, Takagi et al. 1997)
	SACPD-B	Deletion(*)	KK2 (Bay)	(Rahman, Takagi et al. 1997)
	N/A	N/A	'Williams 82'	(Bernard and Cremeens 1988)
	N/A	N/A	'Bay'	(Buss, Smith et al. 1979)
	N/A	N/A	'5601T'	(Pantalone, Allen et al. 2003)

*unpublished - Gillman