Databases for the Future: Where Can Technology Take Us

James Harnly
Food Composition & Methods Development Laboratory
Beltsville Human Nutrition Research Center
Agricultural Research Service
U.S. Department of Agriculture
Beltsville, MD, 20705 USA

More Information in Databases - Why?

- Concentration variation of commodity food components
  - wrt: species, variety, location, harvest season, processing
- Interest in secondary metabolites (bioactive compounds)
- Genetic information: taxonomy, DNA barcoding, microDNAs
- Growing information: location, season, soil, weather, etc.
- New technology
- Authentication/quality control
- Research support
  - Nutritional epidemiology
  - Nutrigenomics
  - Nutrigenetics
  - Personalized nutrition

New Technology

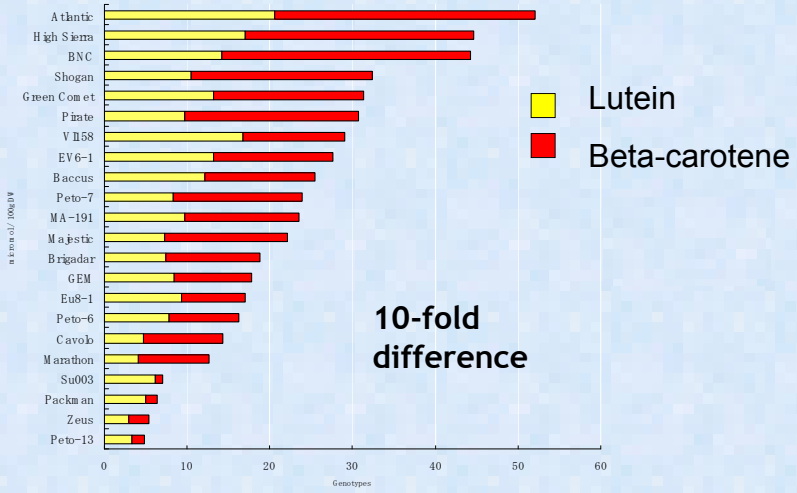
Computer Science (solid state electronics)
Instrumentation:
- **High performance liquid chromatography (HPLC)**
- **Ultra-high performance chromatography (U-HPLC)**
- **Tandem mass spectrometry (MS^n)**
- **High resolution mass spectrometry (HRMS)**

Biology:
- **PCR amplification**
- **Genetic profiling**
- **DNA Bar-Coding**

Systems Biology:
- **Genomics**
- **Transcriptomics**
- **Proteomics**
- **Metabolomics**

Concentration Variation

Carotenoid Levels in Different Varieties of Broccoli



Means, 22 different broccoli genotypes (µmol/100g DW)

Kurilich et al 1999

Concentration Variation

Vitamin Levels in Different Varieties of Broccoli

Mean Vitamin Levels in 50 Broccoli Genotypes (mg/ 100 g fresh weight)

β-carotene	0.89
range	0.37 - 2.42
α -tocopherol	1.62
range	0.46 - 4.29
Ascorbate	74.7
range	54.0 - 119.8

Kurilich et al, 1999

Secondary Metabolites

Phenolic Compounds: phenolic acids, polyphenols (flavonoids), proanthocyanidins (flavonoid polymers), lignans (phytoestrogens), stilbenes (resveratrol)

Terpenes: monoterpenes (limonene), sesquiterpenes (pre-essential oils), diterpenes (retinol, retinal, pre-taxol), triterpenes (pre-steroids, saponins), tetraterpenes (carotenoids, xanthophyls)

Sulfur compounds: isothiocyanates, allyls, polysulfides, indoles, glucosinolates

Organic acids

Steroids: cholesterol, sterols, hormones

Alkaloids: caffeine, nicotine, berberine, atropine

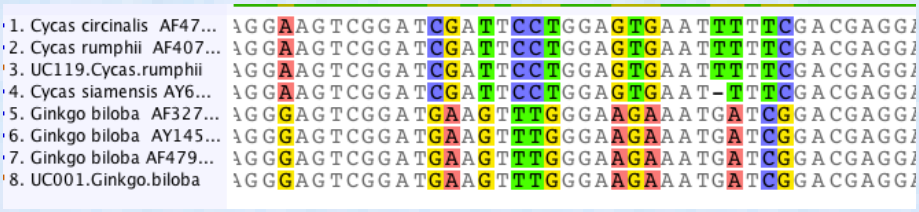
Micro RNAs

A **microRNA** (abbreviated **miRNA**) is a short ribonucleic acid (RNA) molecule found in eukaryotic cells. miRNA molecules have very few nucleotides (an average of 22) compared with other RNAs. miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA (mRNAs), usually resulting in gene silencing (some gene activation has also been reported). The human genome may encode over 1000 miRNAs. To date over 900 miRNAs have been identified in 71 plant species. Zhang et al.* have reported that exogenous plant miRNAs in food can regulate the expression of target genes in mammals.

*Cell Research (2012) 22:107–126.

Genetic Identification

DNA Bar-Coding: *Ginkgo biloba* vs Several Cycas species*



***Harbaugh-Reynaud et al. – Value added barcodes for NIST reference botanicals.**

FDA has reported:

Using DNA barcodes for identification of fish species.

Complete gene profiling of bacterial species in 4 hours.

Computer Science
Cost of hardware

1977	2008	2012
32 KB Core Memory $30,000	32 MB Flash Memory $30	32 GB Flash Memory $30
DEC 1134 Minicomputer $45,000	Lap Top ~$2,000	Lap Top ~$1,000

Major expense is software, not hardware!

Metabolomics

Inventory of small molecules:
In food materials = nutrition

Based on new instrumentation: U-HPLC and HRMS

U-HPLC → hundreds of peaks and each peak has a mass

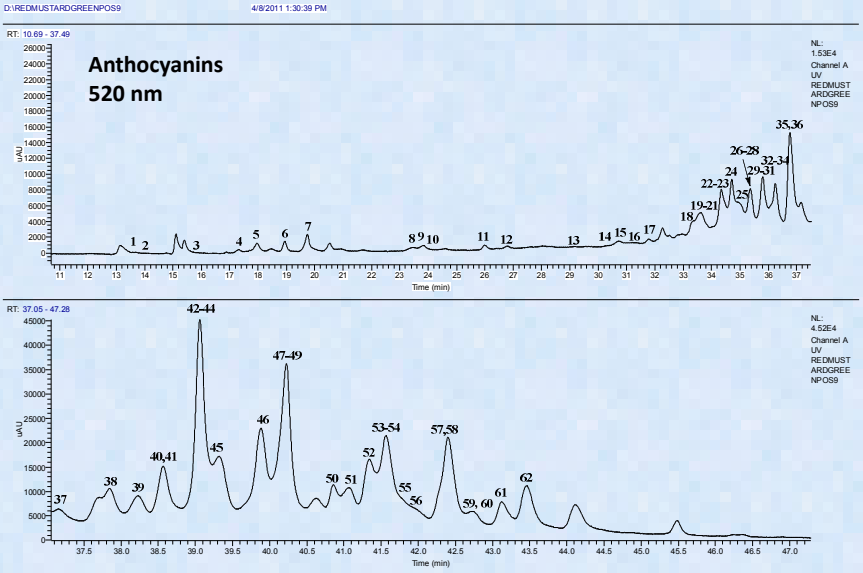
HRMS → m/z 645.8923
→ exact molecular formula
→ only a few possible compounds

High potential for identifying all the peaks in a chromatogram

More biological samples → better databases for identifying compounds

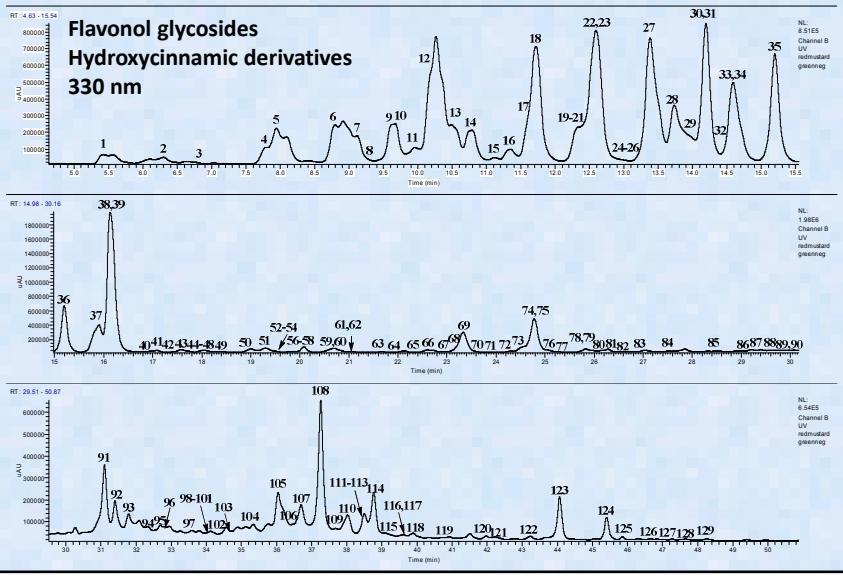
U-HPLC with HRMS

Red Mustard Greens - more than 60 anthocyanins identified



U-HPLC with HRMS

Red Mustard Greens - more than 140 glycosylated flavonols identified



Red Mustard Greens

more than 140 glycosylated flavonols identified

Peak No	t_R (min)	$[M]^+$ weight	$[M]^+$ formula	Error (ppm)	major and important MS^2 ions (m/z) (%)	MS^3 ion (m/z) (%)	UV/Vis λ_{max} (nm)	Tentative identification
27 Acylated cyanidin 3-sophoroside-5-diglucosides								
3	15.89	1141.3234	$C_{50}H_{61}O_{30}$	-0.72	817(100), 611(48)	287(100)	nd	Cy 3-sinapoylsophoroside-5-diglucoside
28	35.41	1287.3604	$C_{59}H_{67}O_{32}$	-0.46	963(100), 611(10)	287(100)	328,536	Cy 3-*p*-coumaroylsinapoylsophoroside-5-diglucoside#
14	30.37	1289.3394	$C_{58}H_{65}O_{33}$	-0.67	965(100), 611(15)	287(100)	nd	Cy 3-caffeoylhydroxyferuloylsophoroside-5-diglucoside
17	31.84	1303.3560	$C_{59}H_{67}O_{33}$	0.07	1141(8), 979(100), 611(26)	287(100)	328,536	Cy 3-caffeoylsinapoylsophoroside-5-diglucoside
34	36.32	1317.3724	$C_{60}H_{69}O_{33}$	0.64	993(100), 611(9), 593(5)	287(100)	328, 536	Cy 3-sinfersophoroside-5-diglucoside
ad-5	34.86	1183.2782	$C_{54}H_{55}O_{30}$	0.79	773(69), 697(39),	287(100)	328,534	Cy 3-caffeoylsophoroside-5-malonyldiglucoside
20	33.56	1197.3137	$C_{57}H_{51}O_{32}$	-0.29	949(17), 787(61), 697(100),	287(100)	326, 538	Cy 3-feruloylsophoroside-5-malonyldiglucoside
6	19.01	1227.3226	$C_{53}H_{63}O_{33}$	-1.64	817(16), 697(100), 653(7)	287(100)	328,536	Cy 3-caffeoylsinapoylsophoroside-5-malonyldiglucoside
45	39.75	1343.3512	$C_{61}H_{67}O_{34}$	0.35	933(100), 697(61), 455(11)	287(100)	328,536	Cy 3-*p*-coumaroylferuloylsophoroside-5-malonyldiglucoside
24	34.56	1345.3313	$C_{60}H_{65}O_{35}$	0.90	935(91), 697(100), 679(2), 653(11)	287(100)	326,538	Cy 3- dicaffeoylsophoroside-5-malonyldiglucoside
29	35.70	1359.3455	$C_{61}H_{67}O_{35}$	-0.18	1315(20), 949(100), 697(90), 653(6)	287(100)	328,536	Cy 3-caffeoylferuloylsophoroside-5-malonyldiglucoside
43	39.12	1359.3472	$C_{61}H_{67}O_{35}$	1.07	1195(2), 1067(2), 949(85), 697(100)	287(100)	328,536	Cy 3-caffeoylferuloylsoporoside-5-malonyldiglucoside
38	37.88	1359.3469	$C_{61}H_{67}O_{35}$	0.85	1315(38), 949(100), 697(85)	287(100)	328,536	Cy 3-caffeoylferuloylsophoroside-5-malonyldiglucoside
47	40.23	1373.3639	$C_{62}H_{69}O_{35}$	1.83	963(100), 962(3), 697(88), 653(3)	287(100)	326, 538	Cy 3-*p*-coumaroylsinapoylsophoroside-5-malonyldiglucoside#
48	40.23	1373.3628	$C_{62}H_{69}O_{35}$	1.03	963(54), 697(100), 653(12)	287(100)	328,536	Cy 3-*p*-coumaroylsinapoylsophoroside-5-malonyldiglucoside#
19	33.44	1375.3407	$C_{61}H_{67}O_{36}$	0.03	965(100), 697(93), 611(4), 541(2)	287(100)	326,538	Cy 3-hydroxyferuloylcaffeoylsophoroside-5-malonyldiglucoside
22	34.40	1389.3557	$C_{62}H_{69}O_{36}$	-0.44	979(100), 697(67), 653(22)	287(100)	330, 534	Cy 3-caffeoylsinapoylsophoroside-5-malonyldiglucoside
39	38.32	1389.3566	$C_{62}H_{69}O_{36}$	0.21	1077(27), 979(17), 734(17), 697(57)	287(100)	328, 538	Cy 3-caffeoylsinapoylsophoroside-5-malonyldiglucoside
42	39.12	1403.3719	$C_{63}H_{71}O_{36}$	-0.04	1359(6), 993(81), 697(100), 653(5)	287(100)	326, 538	Cy 3-sinapoylferuloylsophoroside-5-succinoyldiglucoside
10	24.74	1211.3290	$C_{53}H_{63}O_{32}$	-0.57	787(28), 711(100)	287(100)	nd	Cy 3-feruloylsophoroside-5-succinoyldiglucoside
9	23.89	1241.3389	$C_{54}H_{65}O_{33}$	-1.10	817(20), 711(100)	287(100)	328, 536	Cy 3-sinapoylsophoroside-5-succinoyldiglucoside
53	41.63	1357.3668	$C_{62}H_{69}O_{34}$	0.24	1325(2), 933(100), 711(97)	287(100)	326,538	Cy 3-*p*-coumaroylferuloylsophoroside-5-succinoyldiglucoside
40	38.58	1373.3608	$C_{62}H_{69}O_{35}$	-0.43	1111(5), 963(4), 949(98), 711(100)	287(100)	328,536	Cy 3-caffeoylferuloylsophoroside-5-succinoyldiglucoside
50	40.93	1387.3773	$C_{63}H_{71}O_{35}$	0.19	1355(2), 1287(2), 963(100), 711(89)	287(100)	328,536	Cy 3-*p*-coumarolysinapoylsophoroside-5-succinoyldiglucoside#
57	42.36	1387.3767	$C_{63}H_{71}O_{35}$	-0.25	963(100), 711(68)	287(100)	328,536	Cy 3-*p*-coumarolysinapoylsophoroside-5-succinoyldiglucoside#
44	39.12	1403.3726	$C_{63}H_{71}O_{36}$	0.46	1347(83), 979(100), 711(92)	287(100)	328,536	Cy 3-feruloylhydroxyferuloylsophoroside-5-succinoyldiglucoside
54	41.66	1417.3879	$C_{64}H_{73}O_{36}$	0.21	993(100), 711(86)	287(100)	328,536	Cy 3-dihydroxyferuloylsophoroside-5-succinoyldiglucoside

Chemical Identification/Authentication

Chemical fingerprints can be used to discriminate between plant materials based on species, variety, location, harvest season, and processing.

Chromatographic fingerprints: chromatogram as an image.

Spectral fingerprints: solids or direct analysis (no separation) of extracts.

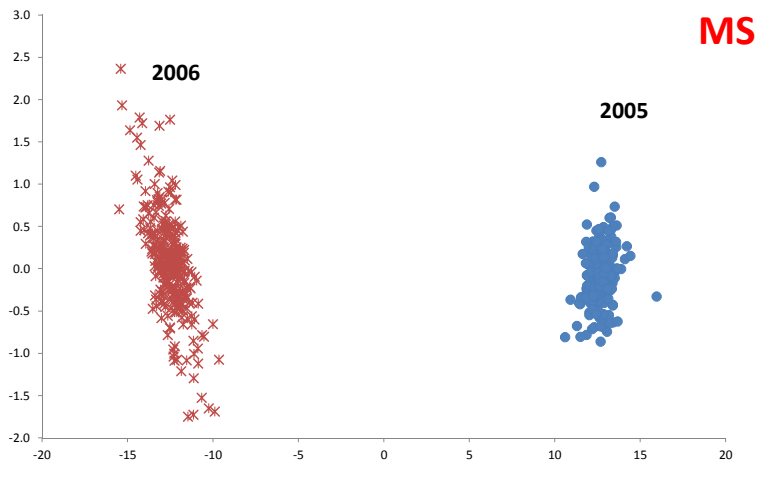
Methods: IR, MS, NIR, NMR, UV

- **Best stability: NIR, NMR, UV**
- **Best sensitivity: MS**
- **Most information: MS, NMR**

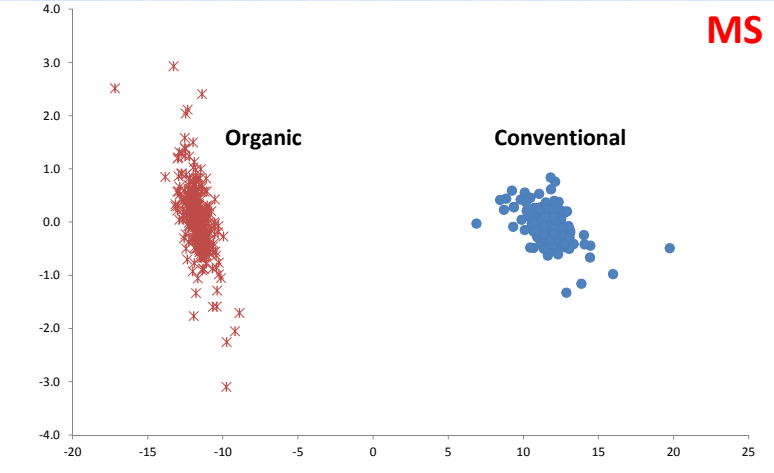
Data Processing:

- **Chemometrics: SIMCA (soft independent modeling of class analogy), i.e. models of authentic materials**

Identification Based on Harvest Year
Rio Red Grapefruit - 2005 vs 2006

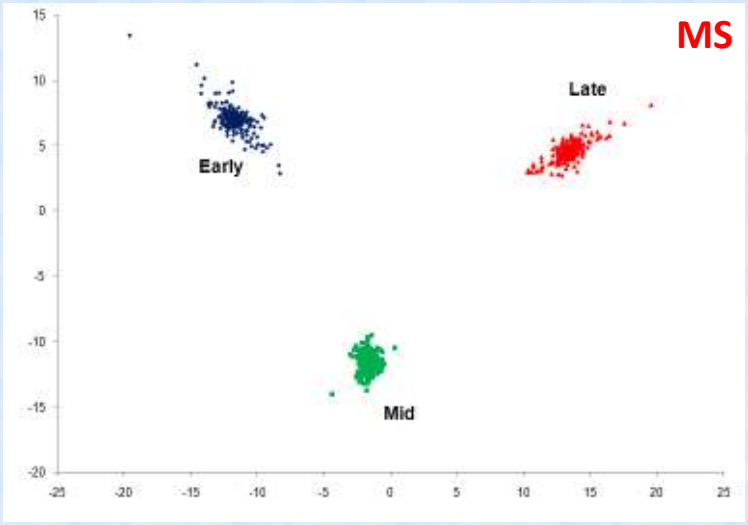


Identification Based on Farming Mode
conventionally vs organically grown grapefruit



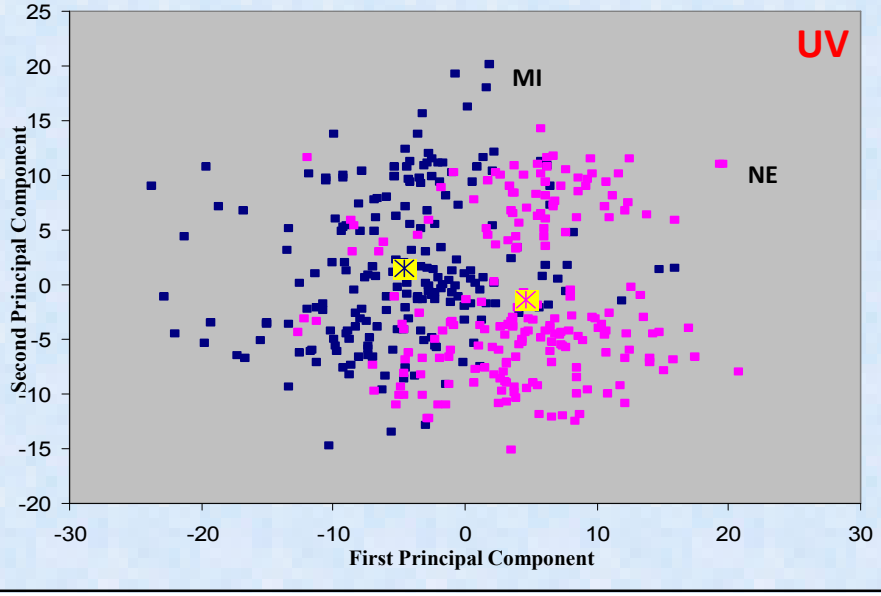
Identification Based on Time of Harvest

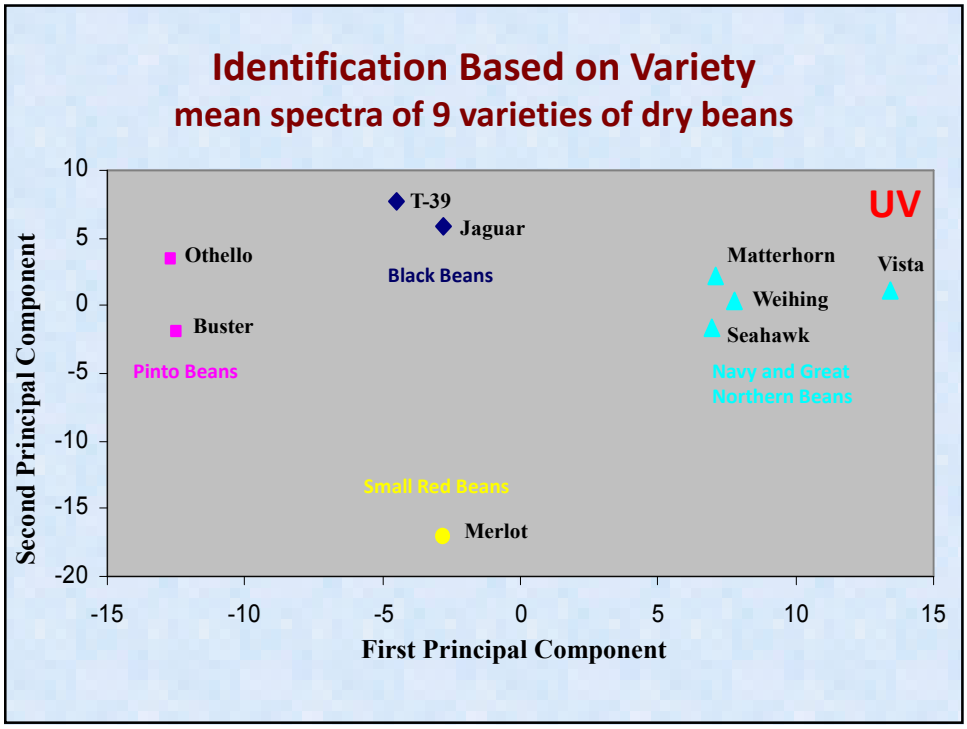
Rio Red Grapefruit - early, mid, & late season

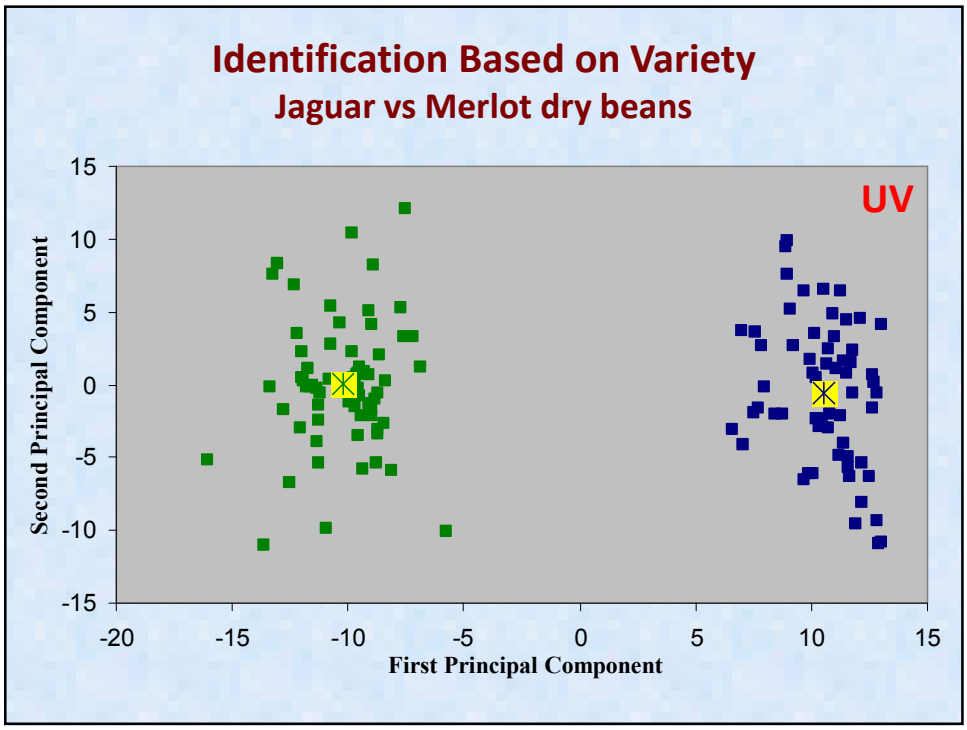


Identification Based on Location

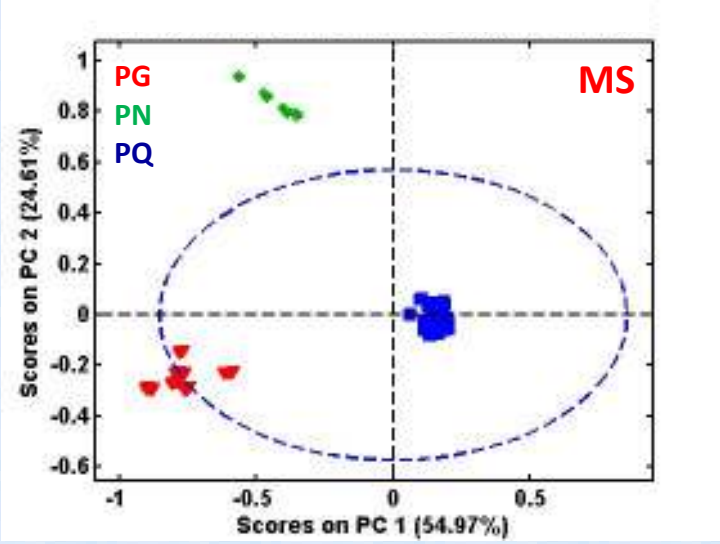
9 varieties of dry beans – MI vs NE



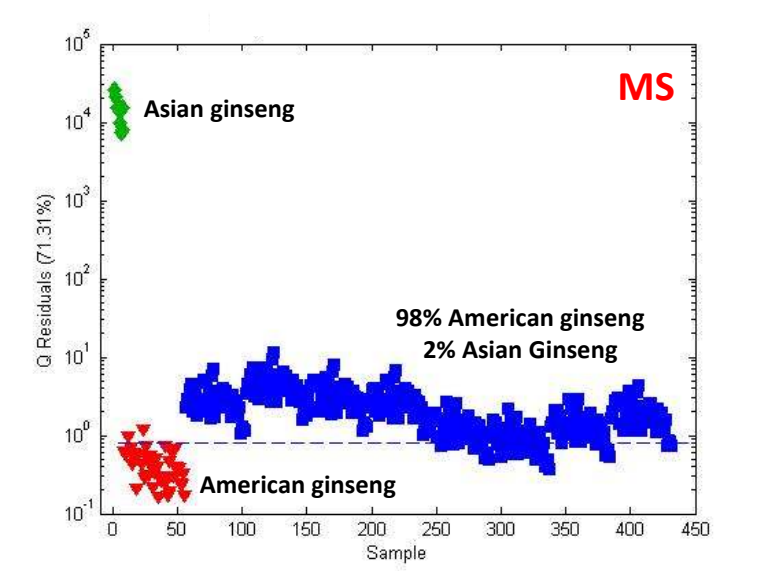

Identification Based on Variety
mean spectra of 9 varieties of dry beans
UV
T-39
Jaguar
Othello
Black Beans
Matterhorn
Vista
Weihing
Buster
Seahawk
Pinto Beans
Navy and Great
Northern Beans
Small Red Beans
Merlot
Second Principal Component
First Principal Component
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10
15


Identification Based on Variety
Jaguar vs Merlot dry beans
UV
Second Principal Component
First Principal Component
15
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10
15

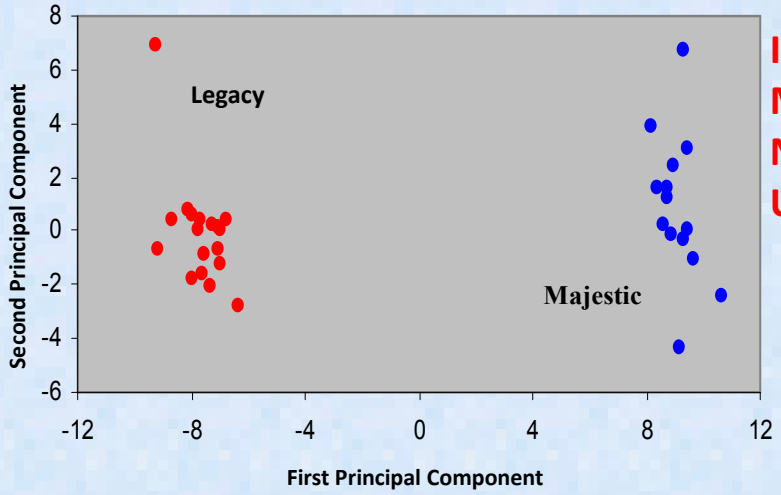
Identification Based on Species
American vs Asian ginseng vs Notoginseng



Identification Based on Authenticity
American vs Asian ginseng

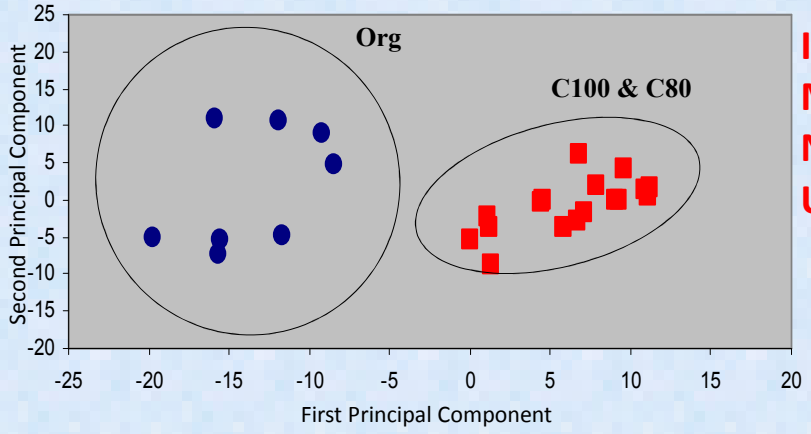


Identification Based on Variety
Legacy vs Majestic Broccoli



IR
MS
NIR
UV

Identification Based on Farming Mode
organically vs conventionally grown broccoli

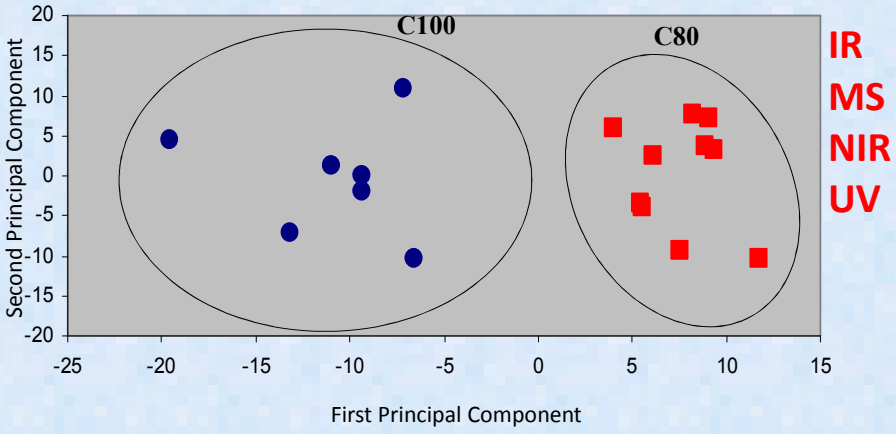


IR
MS
NIR
UV

Ovals represent 95% confidence limits

Identification Based on Farming Conditions

100% vs 80% irrigation for Majestic broccoli (based on transpiration rates)



IR
MS
NIR
UV

Ovals represent 95% confidence limits

Tentative Identification of Significant Masses

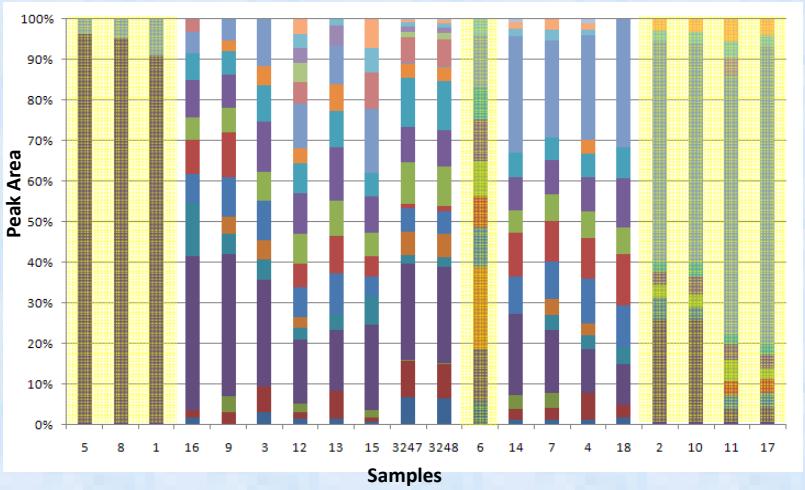
Positive Ions

serine (*m/z* 106)
proline (*m/z* 116)
valine (*m/z* 118)
leucine.isoleucine (*m/z* 132)
aspartic acid (*m/z* 134)
glutamine (*m/z* 147)
glutamic acid (*m/z* 148)
histidine (*m/z* 156)
phenylalanine (*m/z* 166)
arginine (*m/z* 175)
tyrosine (*m/z* 182)

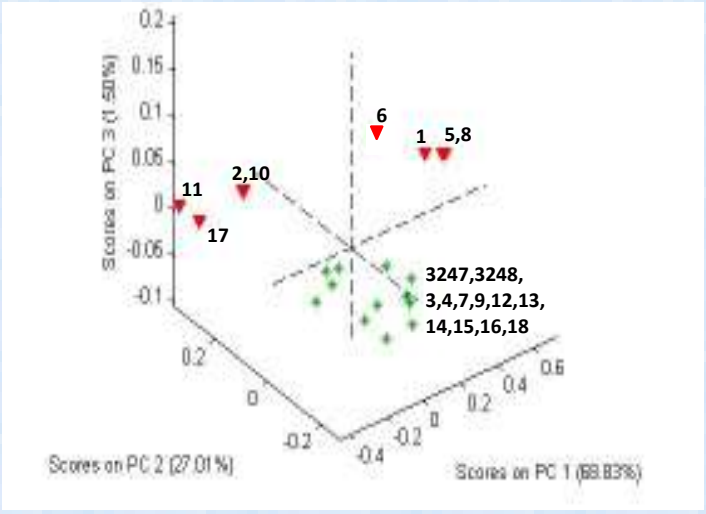
Negative Ions

fumarate (*m/z* 115)
malate (*m/z* 133)
2-oxoglutarate (*m/z* 145)
citrate (*m/z* 191)
hexoses (*m/z* 179)
sucrose & isomers (*m/z* 341)
glucoheptonate (*m/z* 225)

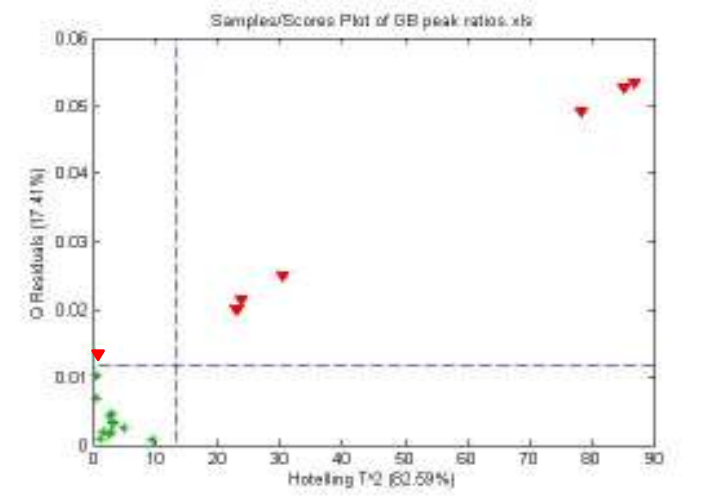
Authenticity of Dietary Supplements
commercially available *Ginkgo biloba*
(relative peak areas by LC-UV)



Authenticity of Dietary Supplements
commercially available *Ginkgo bil*oba
(direct analysis by UV)



Authenticity of Dietary Supplements
commercially available *Ginkgo bil*oba
(direct analysis by UV)

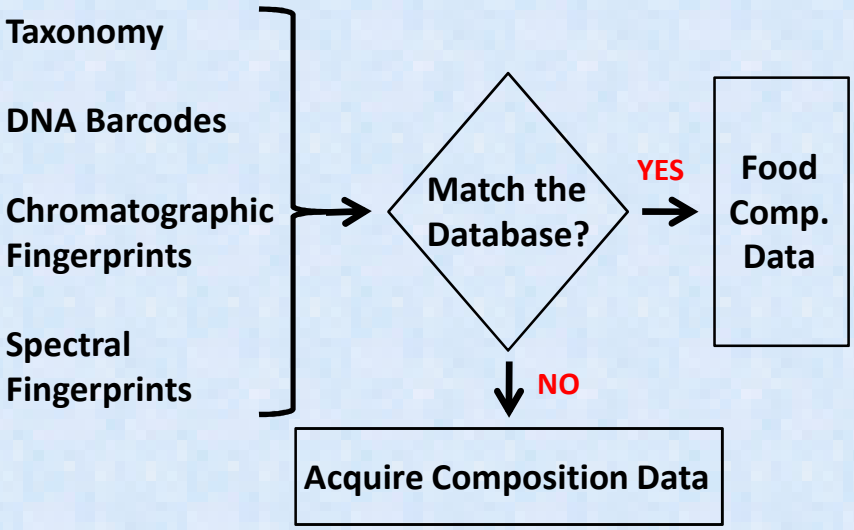


What Information Do We Want in a Database

Taxonomic description: genus, species, sub-species
Growing information
DNA bar-codes
Metabolomic data:
- **Nutrients and secondary metabolites**
- **Polar, semi-polar, and lipid compounds**
  - **Identification**
  - **Concentration**

Micro RNAs
Characteristic analytical features:
- **Chromatograms (HPLC, UHPLC, CE, TLC)**
- **Spectra (MS, IR, NIR, UV, NMR)**

How To Use a Detailed Database



Conclusions

Current and continuously advancing technology will allow tremendous amounts of information from many sources to be incorporated into databases.

The public and researchers need different databases.

Tiered databases, from the most detailed (for researchers) to the most practical/general (for the public) may be the solution.

The analytical perspective: we are acquiring lots of data and we don't want to throw any of it away. Data mining at a later date will be a valuable tool.

Acknowledgements

This research was supported by the USDA Agricultural Research Service and by an Interagency Agreement with the Office of Dietary Supplements at the National Institutes of Health.

Additional thanks to:

Pei Chen	**FCMDL, USDA**
Dave Luthria	**FCMDL, USDA**
Long-Ze Lin	**FCMDL, USDA**
Jianghao Sun	**FCMDL, USDA**
Gene Lester	**PQL, USDA**
Danica Harbaugh-Reynaud	**AuthenTechnologies**

DNA Bar-Coding:

Ginkgo biloba
vs
Cycas species