Introduction to Metabolomics

Introduction to chemometrics

How to analyse many metabolites

UNIVERSITY OF
CAMBRIDGE

Department of
Plant Sciences

Introduction to chemometrics
- Aims

- Align and Bin raw spectral output files
- Merge files for statistical analysis
- Identify many metabolites
- Multivariate data analysis – PCA, Cluster analysis
- ANOVA, false-positive corrections
- Metadata

•Align and “Bin” raw spectral output files

Technical rep 1

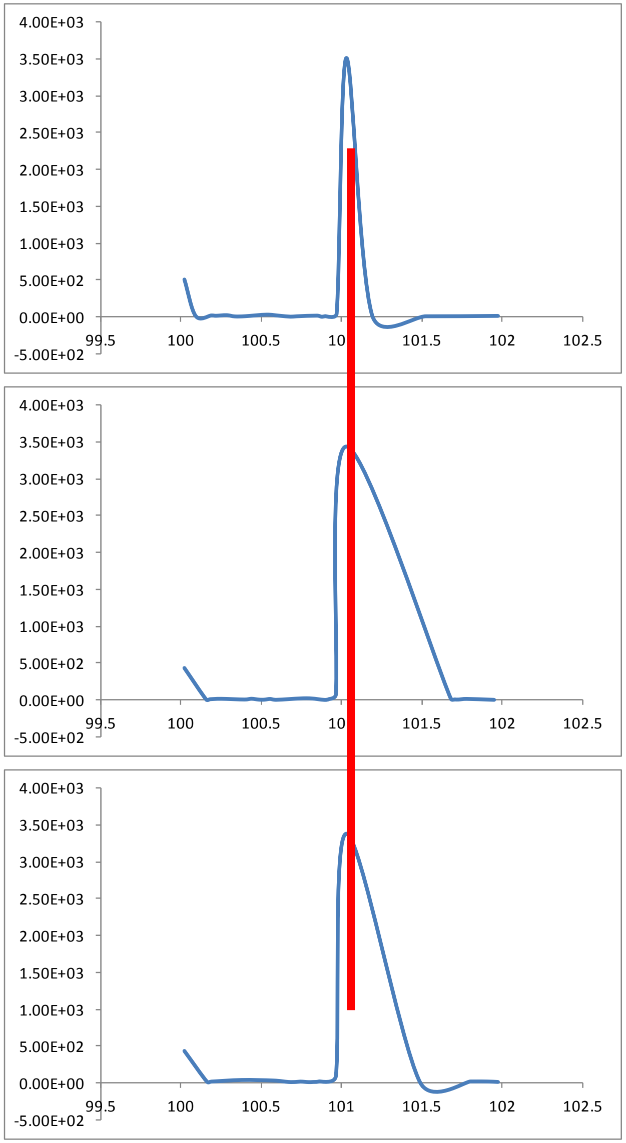
100.0181	4.27E+02
100.1587	8.33E+00
100.1869	9.78E+00
100.3746	3.13E+01
100.5719	2.41E+01
100.666	4.44E+00
100.713	4.56E+00
100.7412	9.67E+00
100.7694	5.78E+00
100.793	2.44E+00
100.8212	4.22E+00
100.8589	1.20E+01
100.9624	8.59E+01
101.0284	3.39E+03
101.4859	4.78E+00
101.8025	1.13E+01
101.9065	1.04E+01
101.9728	4.67E+00

Technical rep 2

100.0174	4.36E+02
100.1534	7.56E+00
100.1769	1.18E+01
100.2472	1.76E+01
100.3928	8.00E+00
100.435	1.81E+01
100.482	6.67E+00
100.5149	5.78E+00
100.5525	1.57E+01
100.5901	4.00E+00
100.7782	2.48E+01
100.8912	2.78E+00
100.9194	1.39E+01
100.9618	7.48E+01
101.0277	3.44E+03
101.6789	1.70E+01
101.7025	9.44E+00
101.7262	7.56E+00
101.7782	1.61E+01
101.9484	3.67E+00

Technical rep 3

100.0179	5.06E+02
100.0882	1.09E+01
100.1867	1.81E+01
100.2148	1.38E+01
100.2805	2.18E+01
100.3322	5.11E+00
100.3979	7.89E+00
100.5294	2.89E+01
100.6093	1.44E+01
100.6799	1.78E+00
100.7175	6.67E+00
100.8445	1.56E+01
100.8728	7.78E-01
100.8963	8.67E+00
100.9669	4.27E+01
101.0282	3.51E+03
101.1884	1.90E+01
101.5234	6.89E+00
101.9726	1.20E+01

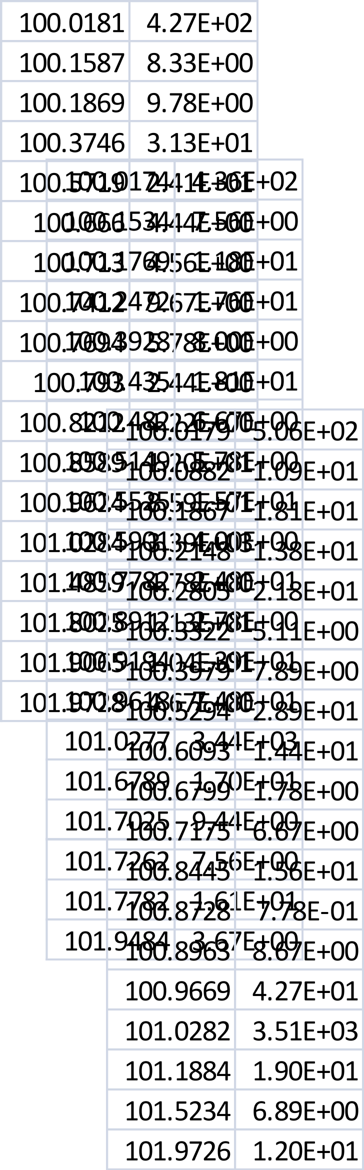


Technical rep 1

Technical rep 2

Technical rep 3

•Align and “Bin” raw spectral output files

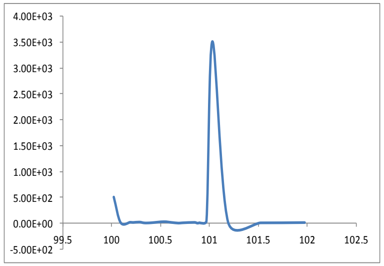


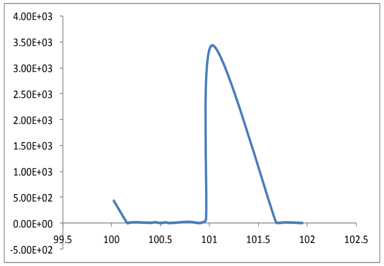
100.0181	4.27E+02
100.1587	8.33E+00
100.1869	9.78E+00
100.3746	3.13E+01
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

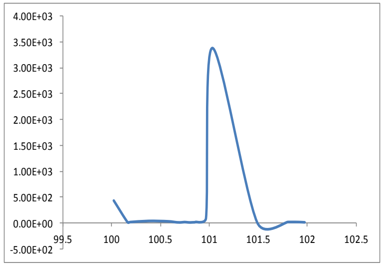
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
101.0277	3.44E+03
101.6789	1.70E+01
101.7025	9.44E+00
101.7262	7.56E+00
101.7782	1.61E+01
101.9484	3.67E+00

100.0179	5.06E+02
100.0882	1.09E+01
100.1867	1.81E+01
100.2148	1.38E+01
100.2805	2.18E+01
100.3322	5.11E+00
100.3579	7.89E+00
100.5294	2.89E+01
100.6093	1.44E+01
100.6799	1.78E+00
100.7175	6.67E+00
100.8443	1.56E+01
100.8728	7.78E-01
100.8963	8.67E+00
100.9669	4.27E+01
101.0282	3.51E+03
101.1884	1.90E+01
101.5234	6.89E+00
101.9726	1.20E+01

“Bin” 101 = contains all data between 100.95 to 101.05

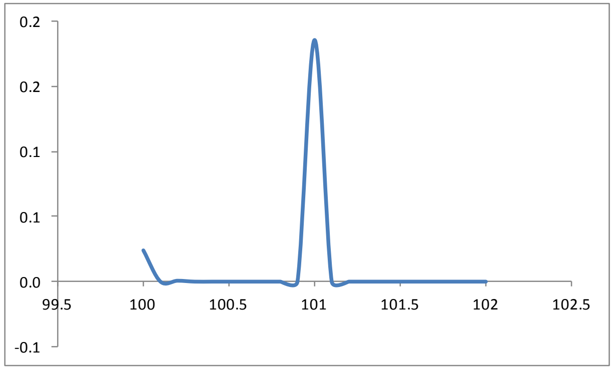


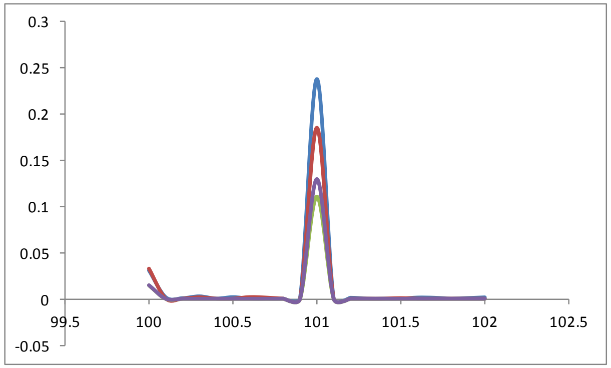




100	0.024
100.1	0
100.2	0.0007
100.3	0
100.4	0
100.5	0
100.6	0
100.7	0
100.8	0
100.9	0
101	0.1855
101.1	0
101.2	0
101.3	0
101.4	0
101.5	0
101.6	0
101.7	0
101.8	0
101.9	0
102	0

Peak alignment





Compare peak height for that bin across many samples

Normalise height to %TIC (TIC/sum total TIC * 100)

Why do we merge? - help with identifying many metabolites

- Identify many compounds – earlier we looked at identifying one metabolite
- Very few sites allows the searching for many metabolites

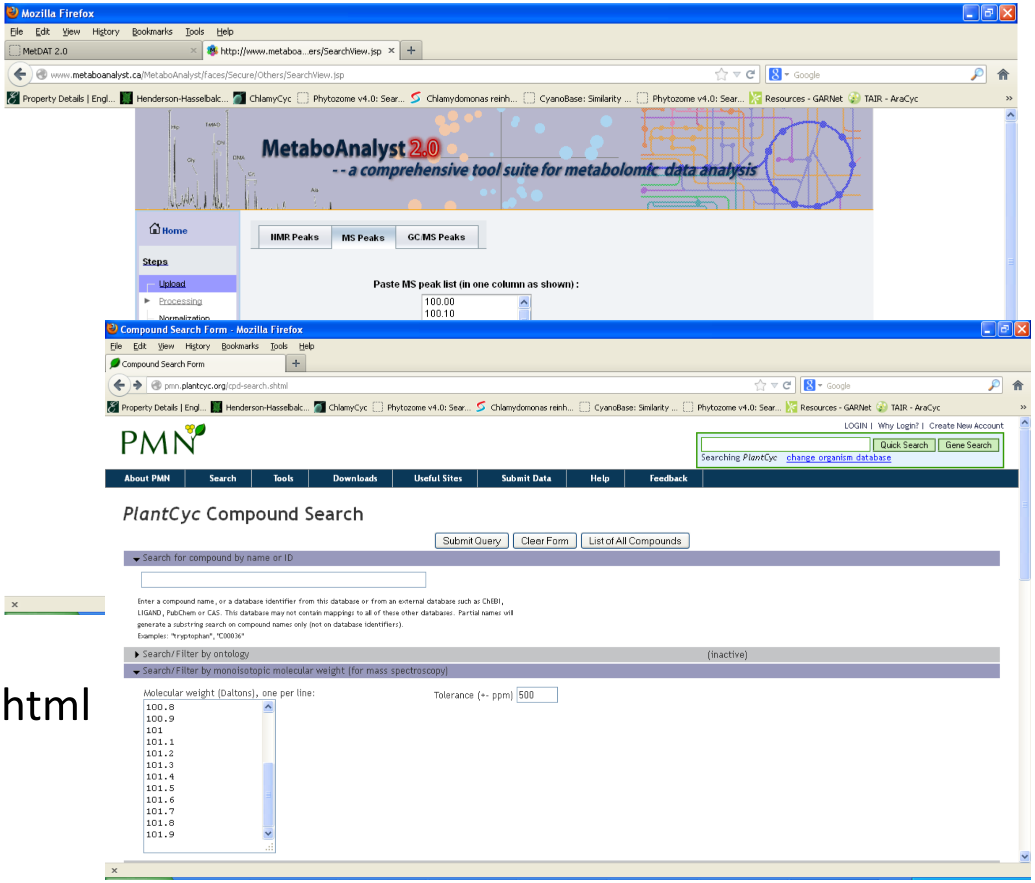
Equipment software

Usually lab specific software
– excel macros

METABOANALYST
http://www.metaboanalyst.ca/MetaboAnalyst/faces/Secure/Others/SearchView.jsp

PlantCyc
http://pmn.plantcyc.org/cpd-search.shtml

http://smbl.nus.edu.sg/METDAT2/



Mozilla Firefox
File Edit View History Bookmarks Tools Help
MetDAT 2.0
http://www.metaboa...ers/SearchView.jsp
www.metaboanalyst.ca/MetaboAnalyst/faces/Secure/Others/SearchView.jsp
Google
Property Details | Engl... Henderson-Hasselbalc... ChlamyCyc Phytozome v4.0: Sear... Chlamydomonas reinh... CyanoBase: Similarity ... Phytozome v4.0: Sear... Resources - GARNet TAIR - AraCyc
MetaboAnalyst 2.0
-- a comprehensive tool suite for metabolomic data analysis
Home
Steps
Upload
Processing
Normalization
NMR Peaks
MS Peaks
GC/MS Peaks
Compound Search Form - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Compound Search Form
pmn.plantcyc.org/cpd-search.shtml
Google
Property Details | Engl... Henderson-Hasselbalc... ChlamyCyc Phytozome v4.0: Sear... Chlamydomonas reinh... CyanoBase: Similarity ... Phytozome v4.0: Sear... Resources - GARNet TAIR - AraCyc
PMN
LOGIN | Why Login? | Create New Account
Quick Search
Gene Search
Searching PlantCyc change organism database
About PMN
Search
Tools
Downloads
Useful Sites
Submit Data
Help
Feedback
Submit Query
Clear Form
List of All Compounds

Paste MS peak list (in one column as shown) :

100.00
100.10

PlantCyc Compound Search

Search for compound by name or ID

Enter a compound name, or a database identifier from this database or from an external database such as ChEBI, LIGAND, PubChem or CAS. This database may not contain mappings to all of these other databases. Partial names will generate a substring search on compound names only (not on database identifiers).
Examples: "tryptophan", "C00036"

Search/Filter by ontology (inactive)

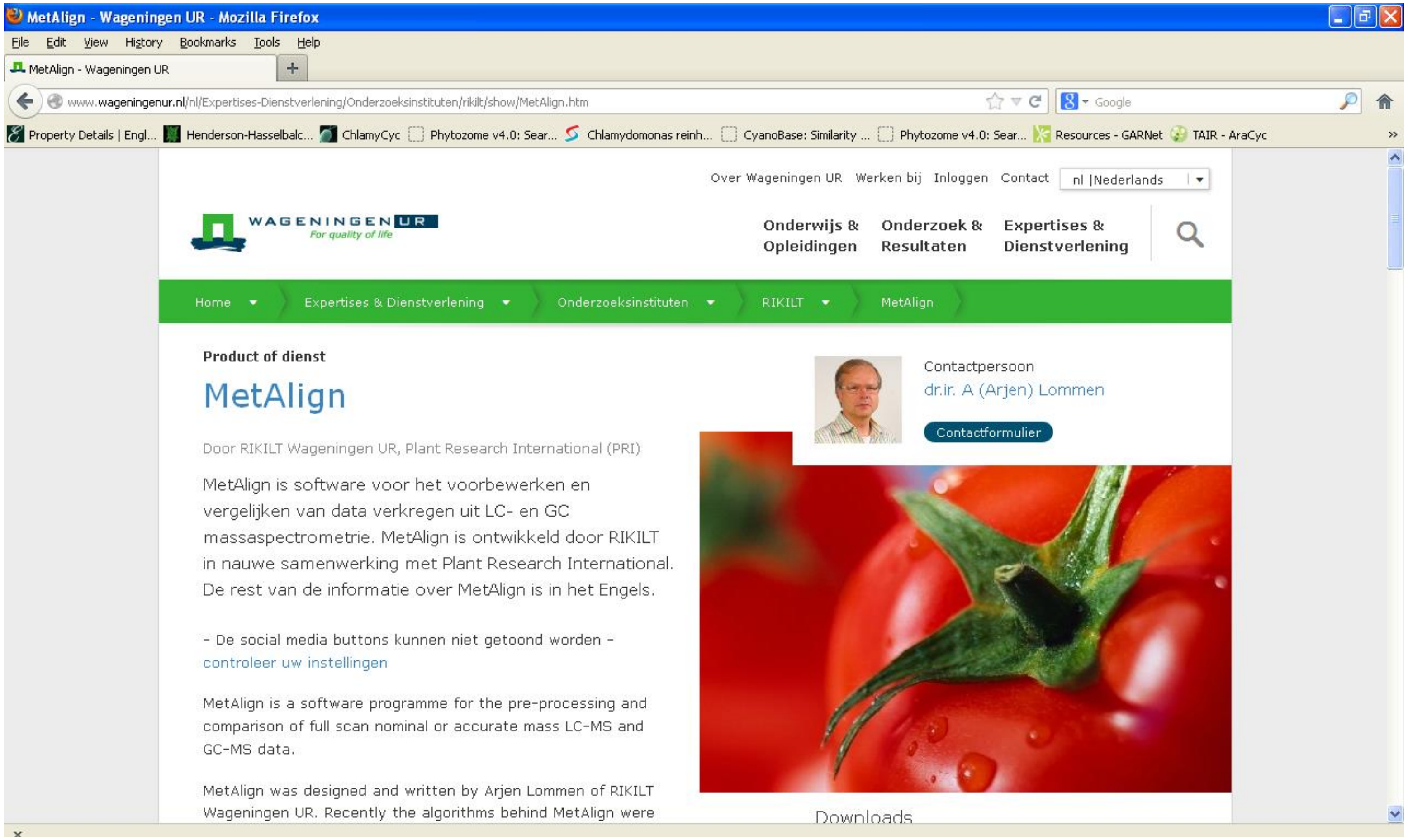
Search/Filter by monoisotopic molecular weight (for mass spectroscopy)

Molecular weight (Daltons), one per line:

100.8
100.9
101
101.1
101.2
101.3
101.4
101.5
101.6
101.7
101.8
101.9

Tolerance (+- ppm) 500

•Align and “Bin” raw spectral output files
•METALIGN


MetAlign - Wageningen UR - Mozilla Firefox
File Edit View History Bookmarks Tools Help
MetAlign - Wageningen UR
www.wageningenur.nl/nl/Expertises-Dienstverlening/Onderzoeksinstituten/rikilt/show/MetAlign.htm
Google
Property Details | Engl... Henderson-Hasselbalc... ChlamyCyc Phytozome v4.0: Sear... Chlamydomonas reinh... CyanoBase: Similarity ... Phytozome v4.0: Sear... Resources - GARNet TAIR - AraCyc
Over Wageningen UR Werken bij Inloggen Contact nl |Nederlands
WAGENINGEN UR For quality of life
Onderwijs & Opleidingen
Onderzoek & Resultaten
Expertises & Dienstverlening
Home
Expertises & Dienstverlening
Onderzoeksinstituten
RIKILT
MetAlign
Contactformulier
Downloads

Product of dienst

MetAlign

Door RIKILT Wageningen UR, Plant Research International (PRI)

MetAlign is software voor het voorbewerken en vergelijken van data verkregen uit LC- en GC massaspectrometrie. MetAlign is ontwikkeld door RIKILT in nauwe samenwerking met Plant Research International. De rest van de informatie over MetAlign is in het Engels.

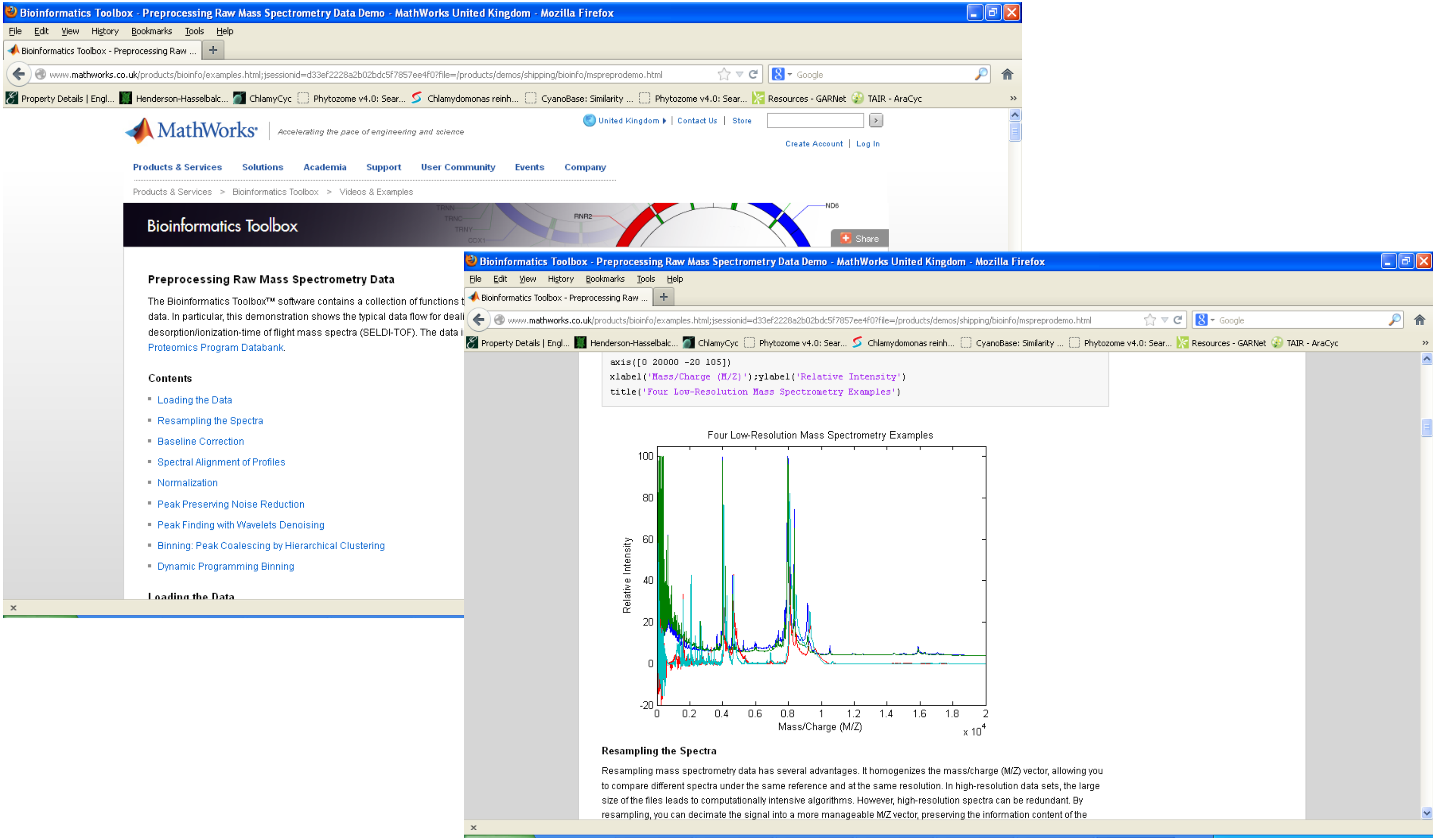
- De social media buttons kunnen niet getoond worden - controleer uw instellingen

MetAlign is a software programme for the pre-processing and comparison of full scan nominal or accurate mass LC-MS and GC-MS data.

MetAlign was designed and written by Arjen Lommen of RIKILT Wageningen UR. Recently the algorithms behind MetAlign were

Contactpersoon
dr.ir. A (Arjen) Lommen

Matlab – MathWorks - commercial



Bioinformatics Toolbox - Preprocessing Raw Mass Spectrometry Data Demo - MathWorks United Kingdom - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Bioinformatics Toolbox - Preprocessing Raw ...
www.mathworks.co.uk/products/bioinfo/examples.html;jsessionid=d33ef2228a2b02bdc5f7857ee4f0?file=/products/demos/shipping/bioinfo/mspreprodemo.html
Google
Property Details | Engl... Henderson-Hasselbalc... ChlamyCyc Phytozome v4.0: Sear... Chlamydomonas reinh... CyanoBase: Similarity ... Phytozome v4.0: Sear... Resources - GARNet TAIR - AraCyc
MathWorks
Accelerating the pace of engineering and science
United Kingdom | Contact Us | Store
Create Account | Log In
Products & Services Solutions Academia Support User Community Events Company
Products & Services > Bioinformatics Toolbox > Videos & Examples
Bioinformatics Toolbox
Share

Preprocessing Raw Mass Spectrometry Data

The Bioinformatics Toolbox™ software contains a collection of functions t data. In particular, this demonstration shows the typical data flow for deali desorption/ionization-time of flight mass spectra (SELDI-TOF). The data i Proteomics Program Databank.

Contents

- Loading the Data
- Resampling the Spectra
- Baseline Correction
- Spectral Alignment of Profiles
- Normalization
- Peak Preserving Noise Reduction
- Peak Finding with Wavelets Denoising
- Binning: Peak Coalescing by Hierarchical Clustering
- Dynamic Programming Binning

Loading the Data

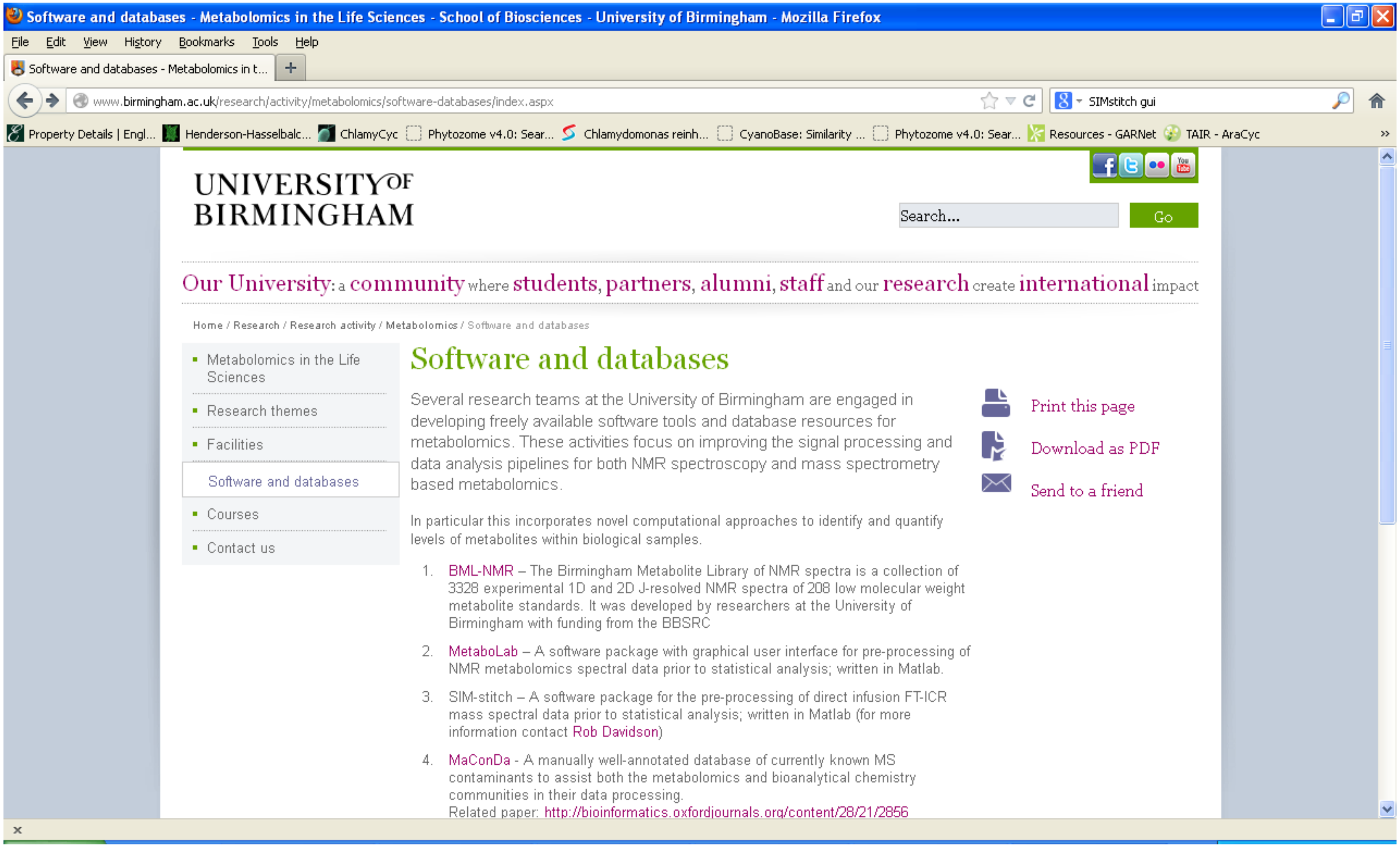
```
axis([0 20000 -20 105])
xlabel('Mass/Charge (M/Z)');ylabel('Relative Intensity')
title('Four Low-Resolution Mass Spectrometry Examples')
```

Resampling the Spectra

Resampling mass spectrometry data has several advantages. It homogenizes the mass/charge (M/Z) vector, allowing you to compare different spectra under the same reference and at the same resolution. In high-resolution data sets, the large size of the files leads to computationally intensive algorithms. However, high-resolution spectra can be redundant. By resampling, you can decimate the signal into a more manageable M/Z vector, preserving the information content of the

http://www.mathworks.co.uk/products/bioinfo/examples.html?file=/products/demos/shipping/bioinfo/mspreprodemo.html

SIM-Stitch University of Birmingham


Software and databases - Metabolomics in the Life Sciences - School of Biosciences - University of Birmingham - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
Software and databases - Metabolomics in t...
www.birmingham.ac.uk/research/activity/metabolomics/software-databases/index.aspx
SIMstitch gui
Property Details | Engl...
Henderson-Hasselbalc...
ChlamyCyc
Phytozome v4.0: Sear...
Chlamydomonas reinh...
CyanoBase: Similarity ...
Phytozome v4.0: Sear...
Resources - GARNet
TAIR - AraCyc
UNIVERSITY OF BIRMINGHAM
Search...
Go

Our University: a community where students, partners, alumni, staff and our research create international impact

Home / Research / Research activity / Metabolomics / Software and databases

- Metabolomics in the Life Sciences
- Research themes
- Facilities
- Software and databases
- Courses
- Contact us

Software and databases

Several research teams at the University of Birmingham are engaged in developing freely available software tools and database resources for metabolomics. These activities focus on improving the signal processing and data analysis pipelines for both NMR spectroscopy and mass spectrometry based metabolomics.

In particular this incorporates novel computational approaches to identify and quantify levels of metabolites within biological samples.

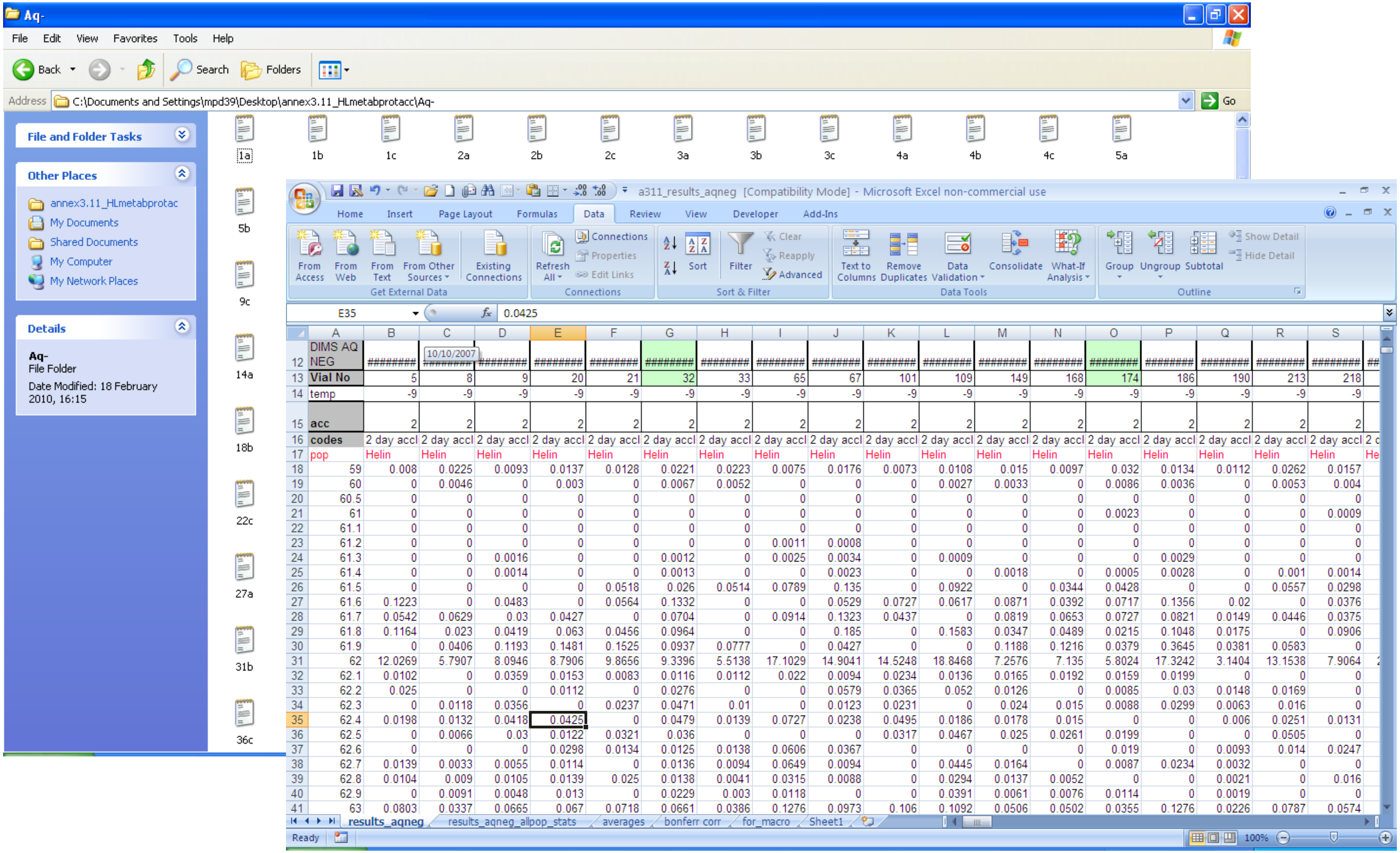
1. BML-NMR – The Birmingham Metabolite Library of NMR spectra is a collection of 3328 experimental 1D and 2D J-resolved NMR spectra of 208 low molecular weight metabolite standards. It was developed by researchers at the University of Birmingham with funding from the BBSRC
2. MetaboLab – A software package with graphical user interface for pre-processing of NMR metabolomics spectral data prior to statistical analysis; written in Matlab.
3. SIM-stitch – A software package for the pre-processing of direct infusion FT-ICR mass spectral data prior to statistical analysis; written in Matlab (for more information contact Rob Davidson)
4. MaConDa - A manually well-annotated database of currently known MS contaminants to assist both the metabolomics and bioanalytical chemistry communities in their data processing.
Related paper: http://bioinformatics.oxfordjournals.org/content/28/21/2856

Print this page

Download as PDF

Send to a friend

•Most people merge files for statistical analysis


Aq-
File Edit View Favorites Tools Help
Back Search Folders
Go
File and Folder Tasks
Other Places
Details
Home Insert Page Layout Formulas Data Review View Developer Add-Ins
From Access From Web From Text From Other Sources Existing Connections
Get External Data
Refresh All Connections Properties Edit Links
Connections
Sort Filter Clear Reapply Advanced
Sort & Filter
Text to Columns Remove Duplicates Data Validation Consolidate What-If Analysis
Data Tools
Group Ungroup Subtotal Show Detail Hide Detail
Outline
Ready 100%

Address C:\Documents and Settings\mpd39\Desktop\annex3.11_HLmetabprotacc\Aq-

Other Places: annex3.11_HLmetabprotac, My Documents, Shared Documents, My Computer, My Network Places

Details: **Aq-** File Folder Date Modified: 18 February 2010, 16:15

Files: 1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b, 3c, 4a, 4b, 4c, 5a, 5b, 9c, 14a, 18b, 22c, 27a, 31b, 36c

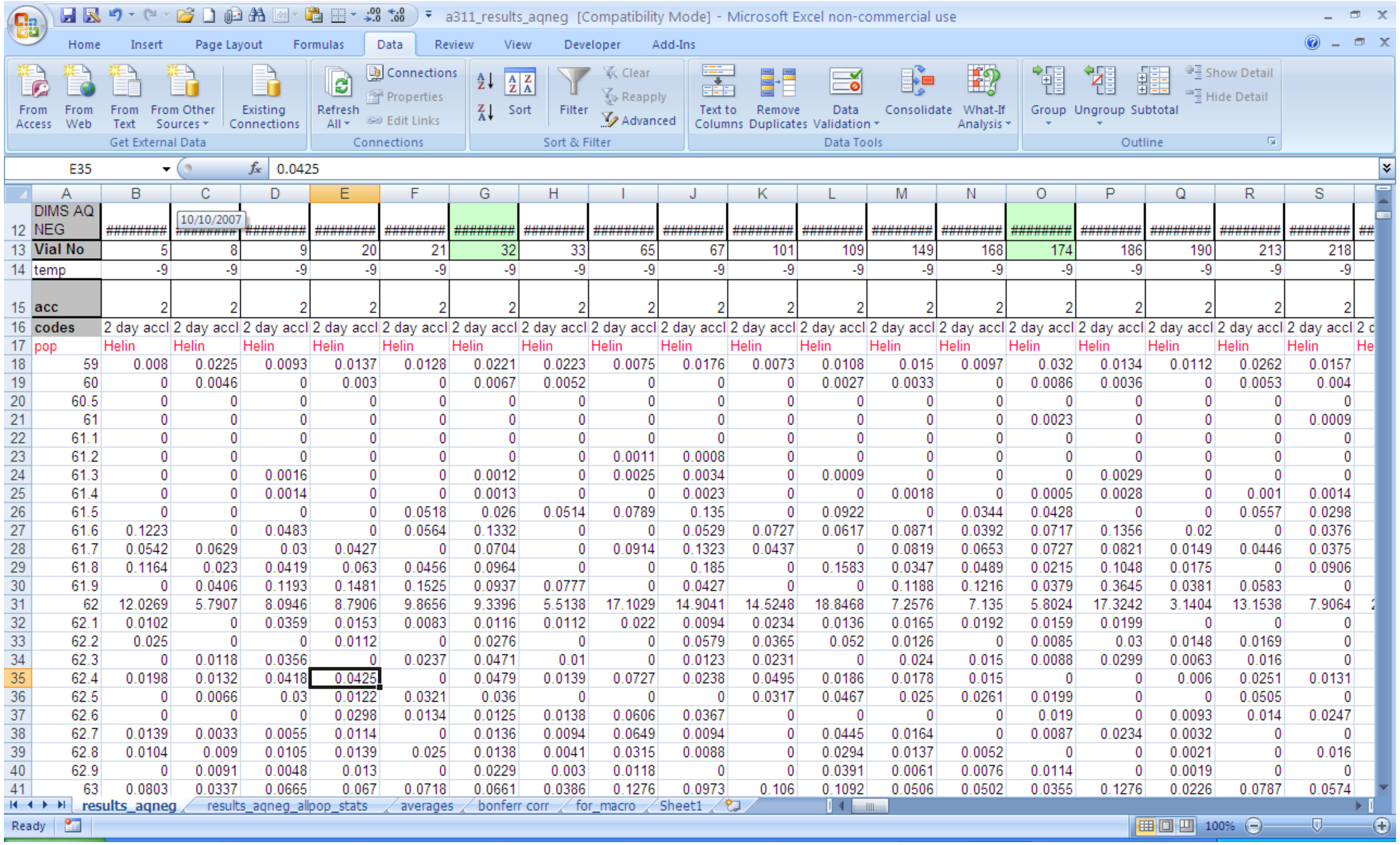
a311_results_aqneg [Compatibility Mode] - Microsoft Excel non-commercial use

E35 0.0425

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
12	DIMS AQ NEG	########	######## 10/10/2007	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########
13	**Vial No**	5	8	9	20	21	32	33	65	67	101	109	149	168	174	186	190	213	218
14	temp	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9
15	**acc**	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
16	**codes**	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl
17	pop	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin
18	59	0.008	0.0225	0.0093	0.0137	0.0128	0.0221	0.0223	0.0075	0.0176	0.0073	0.0108	0.015	0.0097	0.032	0.0134	0.0112	0.0262	0.0157
19	60	0	0.0046	0	0.003	0	0.0067	0.0052	0	0	0	0.0027	0.0033	0	0.0086	0.0036	0	0.0053	0.004
20	60.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0023	0	0	0	0.0009
22	61.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	61.2	0	0	0	0	0	0	0	0.0011	0.0008	0	0	0	0	0	0	0	0	0
24	61.3	0	0	0.0016	0	0	0.0012	0	0.0025	0.0034	0	0.0009	0	0	0	0.0029	0	0	0
25	61.4	0	0	0.0014	0	0	0.0013	0	0	0.0023	0	0	0.0018	0	0.0005	0.0028	0	0.001	0.0014
26	61.5	0	0	0	0	0.0518	0.026	0.0514	0.0789	0.135	0	0.0922	0	0.0344	0.0428	0	0	0.0557	0.0298
27	61.6	0.1223	0	0.0483	0	0.0564	0.1332	0	0	0.0529	0.0727	0.0617	0.0871	0.0392	0.0717	0.1356	0.02	0	0.0376
28	61.7	0.0542	0.0629	0.03	0.0427	0	0.0704	0	0.0914	0.1323	0.0437	0	0.0819	0.0653	0.0727	0.0821	0.0149	0.0446	0.0375
29	61.8	0.1164	0.023	0.0419	0.063	0.0456	0.0964	0	0	0.185	0	0.1583	0.0347	0.0489	0.0215	0.1048	0.0175	0	0.0906
30	61.9	0	0.0406	0.1193	0.1481	0.1525	0.0937	0.0777	0	0.0427	0	0	0.1188	0.1216	0.0379	0.3645	0.0381	0.0583	0
31	62	12.0269	5.7907	8.0946	8.7906	9.8656	9.3396	5.5138	17.1029	14.9041	14.5248	18.8468	7.2576	7.135	5.8024	17.3242	3.1404	13.1538	7.9064
32	62.1	0.0102	0	0.0359	0.0153	0.0083	0.0116	0.0112	0.022	0.0094	0.0234	0.0136	0.0165	0.0192	0.0159	0.0199	0	0	0
33	62.2	0.025	0	0	0.0112	0	0.0276	0	0	0.0579	0.0365	0.052	0.0126	0	0.0085	0.03	0.0148	0.0169	0
34	62.3	0	0.0118	0.0356	0	0.0237	0.0471	0.01	0	0.0123	0.0231	0	0.024	0.015	0.0088	0.0299	0.0063	0.016	0
35	62.4	0.0198	0.0132	0.0418	0.0425	0	0.0479	0.0139	0.0727	0.0238	0.0495	0.0186	0.0178	0.015	0	0	0.006	0.0251	0.0131
36	62.5	0	0.0066	0.03	0.0122	0.0321	0.036	0	0	0	0.0317	0.0467	0.025	0.0261	0.0199	0	0	0.0505	0
37	62.6	0	0	0	0.0298	0.0134	0.0125	0.0138	0.0606	0.0367	0	0	0	0	0.019	0	0.0093	0.014	0.0247
38	62.7	0.0139	0.0033	0.0055	0.0114	0	0.0136	0.0094	0.0649	0.0094	0	0.0445	0.0164	0	0.0087	0.0234	0.0032	0	0
39	62.8	0.0104	0.009	0.0105	0.0139	0.025	0.0138	0.0041	0.0315	0.0088	0	0.0294	0.0137	0.0052	0	0	0.0021	0	0.016
40	62.9	0	0.0091	0.0048	0.013	0	0.0229	0.003	0.0118	0	0	0.0391	0.0061	0.0076	0.0114	0	0.0019	0	0
41	63	0.0803	0.0337	0.0665	0.067	0.0718	0.0661	0.0386	0.1276	0.0973	0.106	0.1092	0.0506	0.0502	0.0355	0.1276	0.0226	0.0787	0.0574

Sheets: results_aqneg, results_aqneg_allpop_stats, averages, bonferr corr, for_macro, Sheet1

- Multivariate data analysis – PCA, Cluster analysis



E35 0.0425

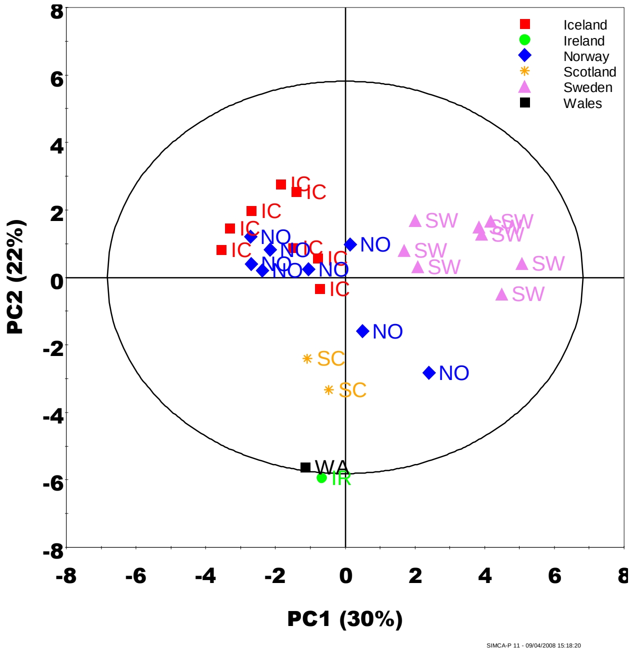
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
12	DIMS AQ NEG	########	10/10/2007 ########	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########	##
13	Vial No	5	8	9	20	21	32	33	65	67	101	109	149	168	174	186	190	213	218	
14	temp	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	
15	acc	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
16	codes	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 c
17	pop	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	He
18	59	0.008	0.0225	0.0093	0.0137	0.0128	0.0221	0.0223	0.0075	0.0176	0.0073	0.0108	0.015	0.0097	0.032	0.0134	0.0112	0.0262	0.0157	
19	60	0	0.0046	0	0.003	0	0.0067	0.0052	0	0	0	0.0027	0.0033	0	0.0086	0.0036	0	0.0053	0.004	
20	60.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0023	0	0	0	0.0009	
22	61.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	61.2	0	0	0	0	0	0	0	0.0011	0.0008	0	0	0	0	0	0	0	0	0	
24	61.3	0	0	0.0016	0	0	0.0012	0	0.0025	0.0034	0	0.0009	0	0	0	0.0029	0	0	0	
25	61.4	0	0	0.0014	0	0	0.0013	0	0	0.0023	0	0	0.0018	0	0.0005	0.0028	0	0.001	0.0014	
26	61.5	0	0	0	0	0.0518	0.026	0.0514	0.0789	0.135	0	0.0922	0	0.0344	0.0428	0	0	0.0557	0.0298	
27	61.6	0.1223	0	0.0483	0	0.0564	0.1332	0	0	0.0529	0.0727	0.0617	0.0871	0.0392	0.0717	0.1356	0.02	0	0.0376	
28	61.7	0.0542	0.0629	0.03	0.0427	0	0.0704	0	0.0914	0.1323	0.0437	0	0.0819	0.0653	0.0727	0.0821	0.0149	0.0446	0.0375	
29	61.8	0.1164	0.023	0.0419	0.063	0.0456	0.0964	0	0	0.185	0	0.1583	0.0347	0.0489	0.0215	0.1048	0.0175	0	0.0906	
30	61.9	0	0.0406	0.1193	0.1481	0.1525	0.0937	0.0777	0	0.0427	0	0	0.1188	0.1216	0.0379	0.3645	0.0381	0.0583	0	
31	62	12.0269	5.7907	8.0946	8.7906	9.8656	9.3396	5.5138	17.1029	14.9041	14.5248	18.8468	7.2576	7.135	5.8024	17.3242	3.1404	13.1538	7.9064	2
32	62.1	0.0102	0	0.0359	0.0153	0.0083	0.0116	0.0112	0.022	0.0094	0.0234	0.0136	0.0165	0.0192	0.0159	0.0199	0	0	0	
33	62.2	0.025	0	0	0.0112	0	0.0276	0	0	0.0579	0.0365	0.052	0.0126	0	0.0085	0.03	0.0148	0.0169	0	
34	62.3	0	0.0118	0.0356	0	0.0237	0.0471	0.01	0	0.0123	0.0231	0	0.024	0.015	0.0088	0.0299	0.0063	0.016	0	
35	62.4	0.0198	0.0132	0.0418	0.0425	0	0.0479	0.0139	0.0727	0.0238	0.0495	0.0186	0.0178	0.015	0	0	0.006	0.0251	0.0131	
36	62.5	0	0.0066	0.03	0.0122	0.0321	0.036	0	0	0	0.0317	0.0467	0.025	0.0261	0.0199	0	0	0.0505	0	
37	62.6	0	0	0	0.0298	0.0134	0.0125	0.0138	0.0606	0.0367	0	0	0	0	0.019	0	0.0093	0.014	0.0247	
38	62.7	0.0139	0.0033	0.0055	0.0114	0	0.0136	0.0094	0.0649	0.0094	0	0.0445	0.0164	0	0.0087	0.0234	0.0032	0	0	
39	62.8	0.0104	0.009	0.0105	0.0139	0.025	0.0138	0.0041	0.0315	0.0088	0	0.0294	0.0137	0.0052	0	0	0.0021	0	0.016	
40	62.9	0	0.0091	0.0048	0.013	0	0.0229	0.003	0.0118	0	0	0.0391	0.0061	0.0076	0.0114	0	0.0019	0	0	
41	63	0.0803	0.0337	0.0665	0.067	0.0718	0.0661	0.0386	0.1276	0.0973	0.106	0.1092	0.0506	0.0502	0.0355	0.1276	0.0226	0.0787	0.0574	

results_aqneg | results_aqneg_allpop_stats | averages | bonferr corr | for_macro | Sheet1

Multivariate Data Analysis

Unsupervised

Principal Component Analysis (PCA)

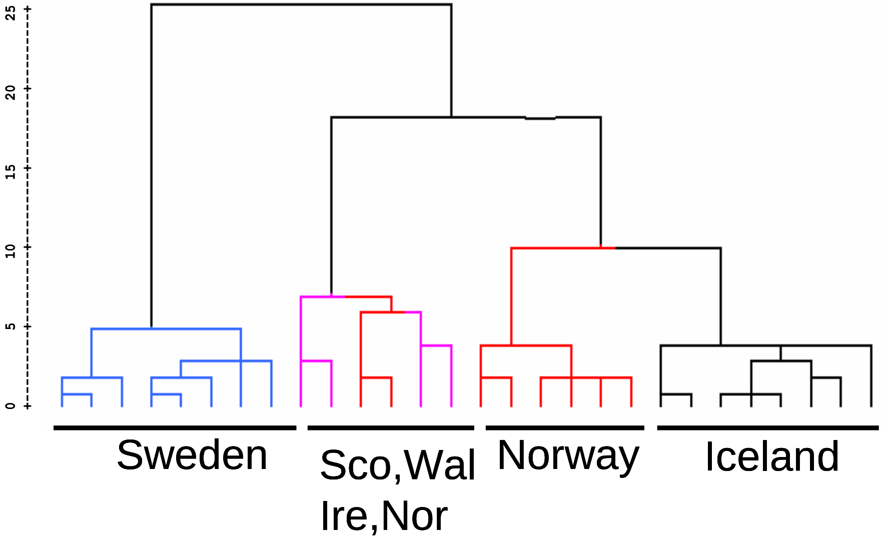


Supervised

Partial Least Squares

-Discriminant Analysis (PLS-DA)

Hierarchical Cluster Analysis (HCA)



Trygg et al. 2007

PCA – Principal Component Analysis

Objectives of PCA

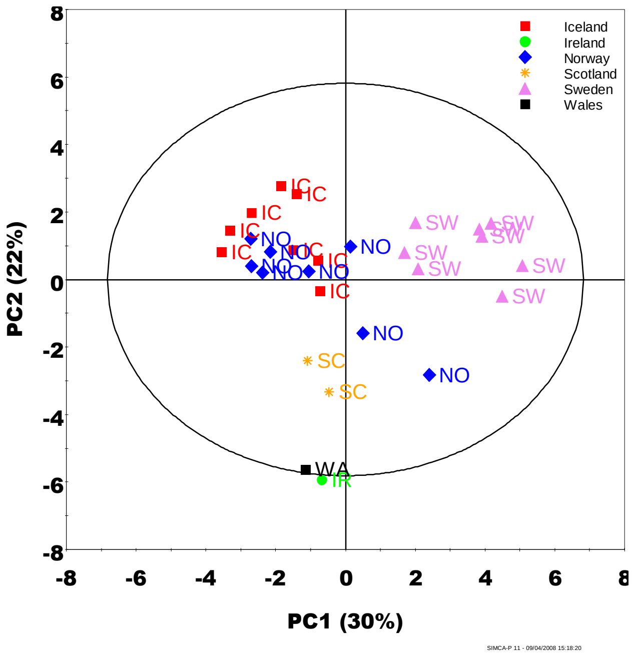
•reduce number of variables

•identify outliers

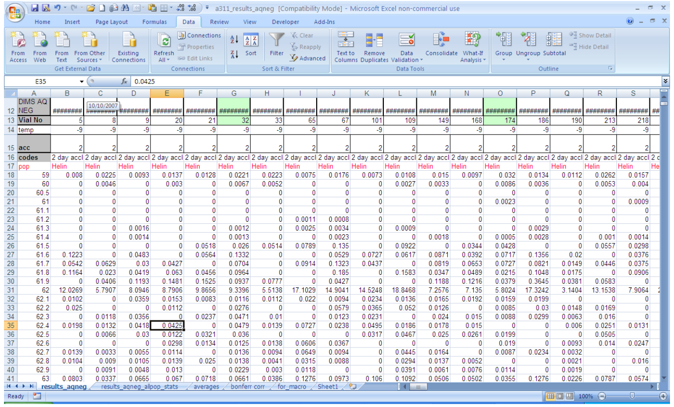
•identify any splits within the data

•discriminate between samples (eg, cold stress, GM, disease, control) or separation by other unplanned means (eg, analytical error)

How does it work?

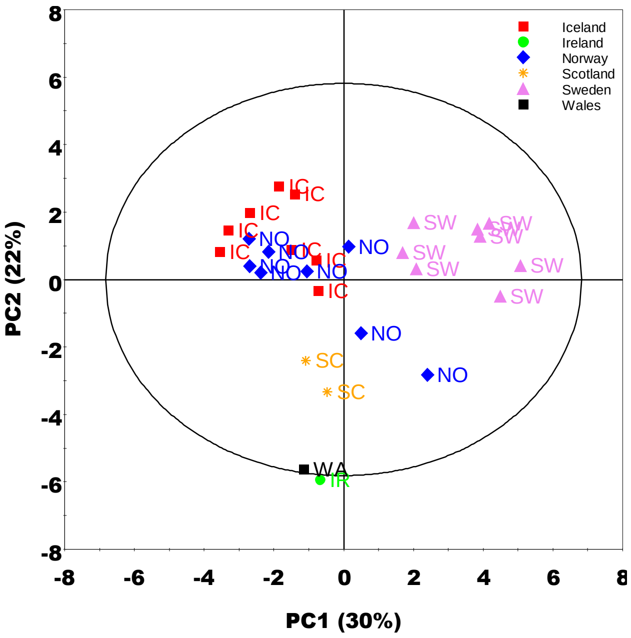


PCA – Principal Component Analysis



E35 0.0425

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
12	DIMS AQ NEG	########	10/10/2007 ########	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########	########
13	Vial No	5	8	9	20	21	32	33	65	67	101	109	149	168	174	186	190	213	218
14	temp	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9
15	acc	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
16	codes	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl
17	pop	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin
18	59	0.008	0.0225	0.0093	0.0137	0.0128	0.0221	0.0223	0.0075	0.0176	0.0073	0.0108	0.015	0.0097	0.032	0.0134	0.0112	0.0262	0.0157
19	60	0	0.0046	0	0.003	0	0.0067	0.0052	0	0	0	0.0027	0.0033	0	0.0086	0.0036	0	0.0053	0.004
20	60.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0023	0	0	0	0.0009
22	61.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	61.2	0	0	0	0	0	0	0	0.0011	0.0008	0	0	0	0	0	0	0	0	0
24	61.3	0	0	0.0016	0	0	0.0012	0	0.0025	0.0034	0	0.0009	0	0	0	0.0029	0	0	0
25	61.4	0	0	0.0014	0	0	0.0013	0	0	0.0023	0	0	0.0018	0	0.0005	0.0028	0	0.001	0.0014
26	61.5	0	0	0	0	0.0518	0.026	0.0514	0.0789	0.135	0	0.0922	0	0.0344	0.0428	0	0	0.0557	0.0298
27	61.6	0.1223	0	0.0483	0	0.0564	0.1332	0	0	0.0529	0.0727	0.0617	0.0871	0.0392	0.0717	0.1356	0.02	0	0.0376
28	61.7	0.0542	0.0629	0.03	0.0427	0	0.0704	0	0.0914	0.1323	0.0437	0	0.0819	0.0653	0.0727	0.0821	0.0149	0.0446	0.0375
29	61.8	0.1164	0.023	0.0419	0.063	0.0456	0.0964	0	0	0.185	0	0.1583	0.0347	0.0489	0.0215	0.1048	0.0175	0	0.0906
30	61.9	0	0.0406	0.1193	0.1481	0.1525	0.0937	0.0777	0	0.0427	0	0	0.1188	0.1216	0.0379	0.3645	0.0381	0.0583	0
31	62	12.0269	5.7907	8.0946	8.7906	9.8656	9.3396	5.5138	17.1029	14.9041	14.5248	18.8468	7.2576	7.135	5.8024	17.3242	3.1404	13.1538	7.9064
32	62.1	0.0102	0	0.0359	0.0153	0.0083	0.0116	0.0112	0.022	0.0094	0.0234	0.0136	0.0165	0.0192	0.0159	0.0199	0	0	0
33	62.2	0.025	0	0	0.0112	0	0.0276	0	0	0.0579	0.0365	0.052	0.0126	0	0.0085	0.03	0.0148	0.0169	0
34	62.3	0	0.0118	0.0356	0	0.0237	0.0471	0.01	0	0.0123	0.0231	0	0.024	0.015	0.0088	0.0299	0.0063	0.016	0
35	62.4	0.0198	0.0132	0.0418	0.0425	0	0.0479	0.0139	0.0727	0.0238	0.0495	0.0186	0.0178	0.015	0	0	0.006	0.0251	0.0131
36	62.5	0	0.0066	0.03	0.0122	0.0321	0.036	0	0	0	0.0317	0.0467	0.025	0.0261	0.0199	0	0	0.0505	0
37	62.6	0	0	0	0.0298	0.0134	0.0125	0.0138	0.0606	0.0367	0	0	0	0	0.019	0	0.0093	0.014	0.0247
38	62.7	0.0139	0.0033	0.0055	0.0114	0	0.0136	0.0094	0.0649	0.0094	0	0.0445	0.0164	0	0.0087	0.0234	0.0032	0	0
39	62.8	0.0104	0.009	0.0105	0.0139	0.025	0.0138	0.0041	0.0315	0.0088	0	0.0294	0.0137	0.0052	0	0	0.0021	0	0.016
40	62.9	0	0.0091	0.0048	0.013	0	0.0229	0.003	0.0118	0	0	0.0391	0.0061	0.0076	0.0114	0	0.0019	0	0
41	63	0.0803	0.0337	0.0665	0.067	0.0718	0.0661	0.0386	0.1276	0.0973	0.106	0.1092	0.0506	0.0502	0.0355	0.1276	0.0226	0.0787	0.0574



SIMCA-P 11 - 09/04/2008 15:18:20

If each variable (ie. metabolite ion intensity or concentration) is thought of as a dimension,

and there are *n* variables,

every sample is at a unique position in the *n*-dimensional space defined by these *n* variables.

Very difficult for people to visualise – aim is to reduce this dimensional space by summarising the data using relatively few parameters

Ie. Make a 2D plot!

Will attempt to explain PCA…

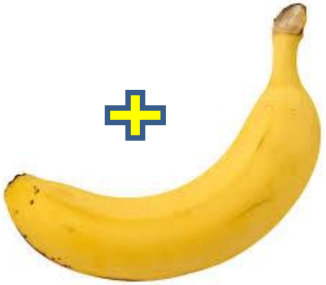
PCA – Principal Component Analysis

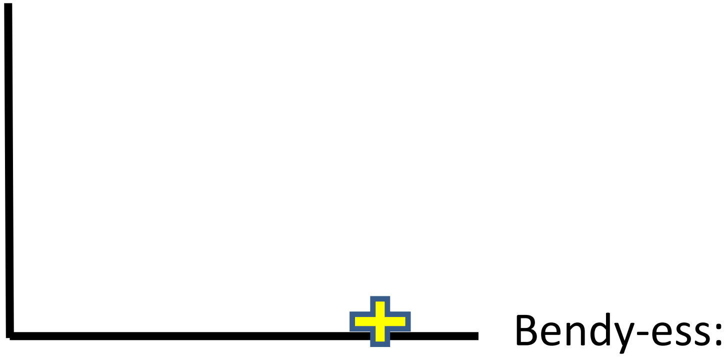
Scores out of 10

Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

Plot these data in K-space

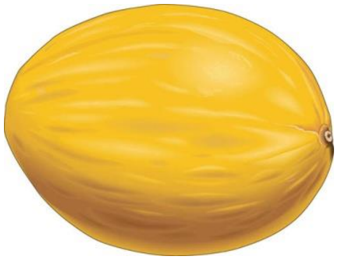
PCA – Principal Component Analysis

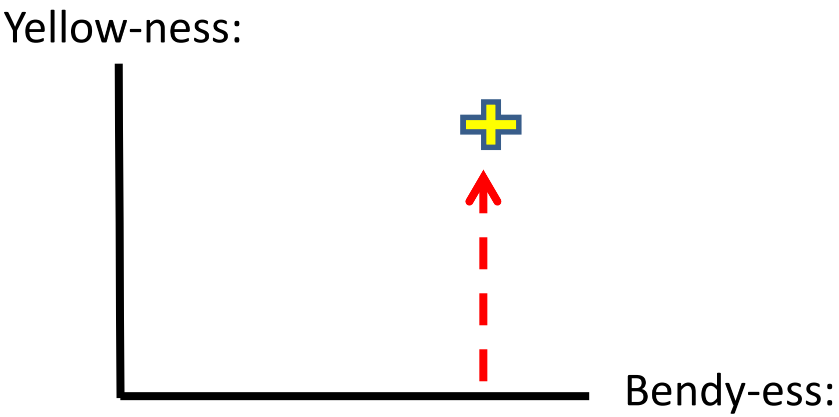




Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

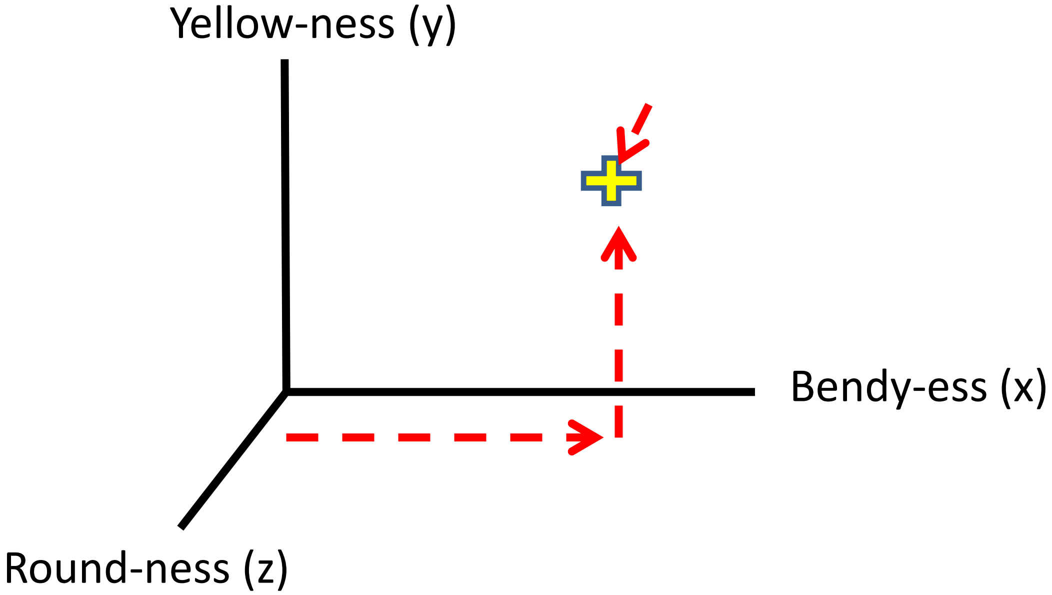
PCA – Principal Component Analysis





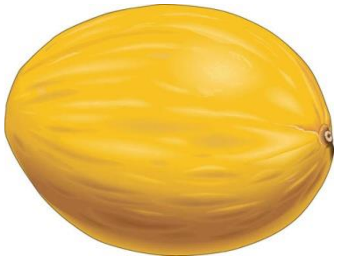
Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

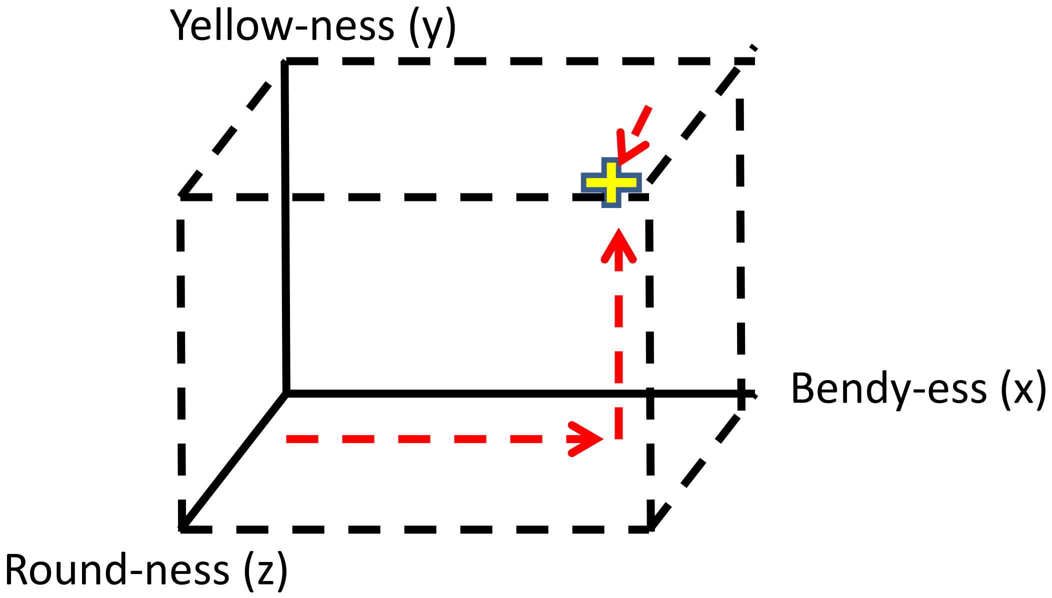
PCA – Principal Component Analysis



Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

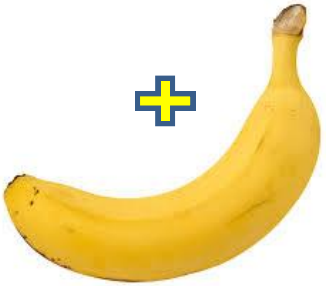
PCA – Principal Component Analysis

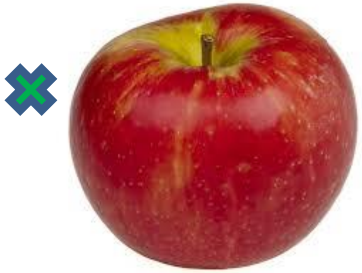


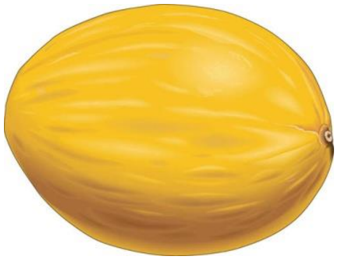


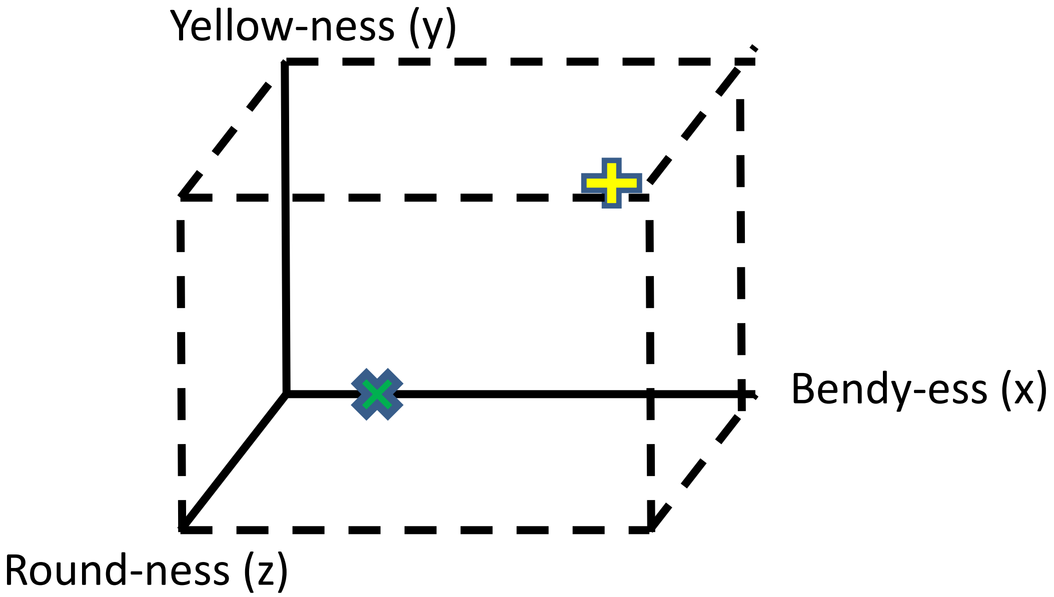
Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis



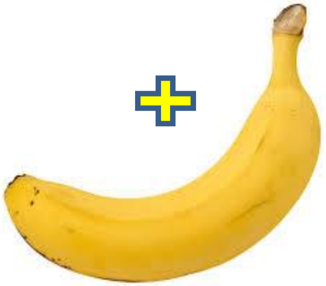


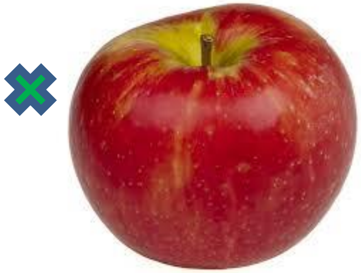


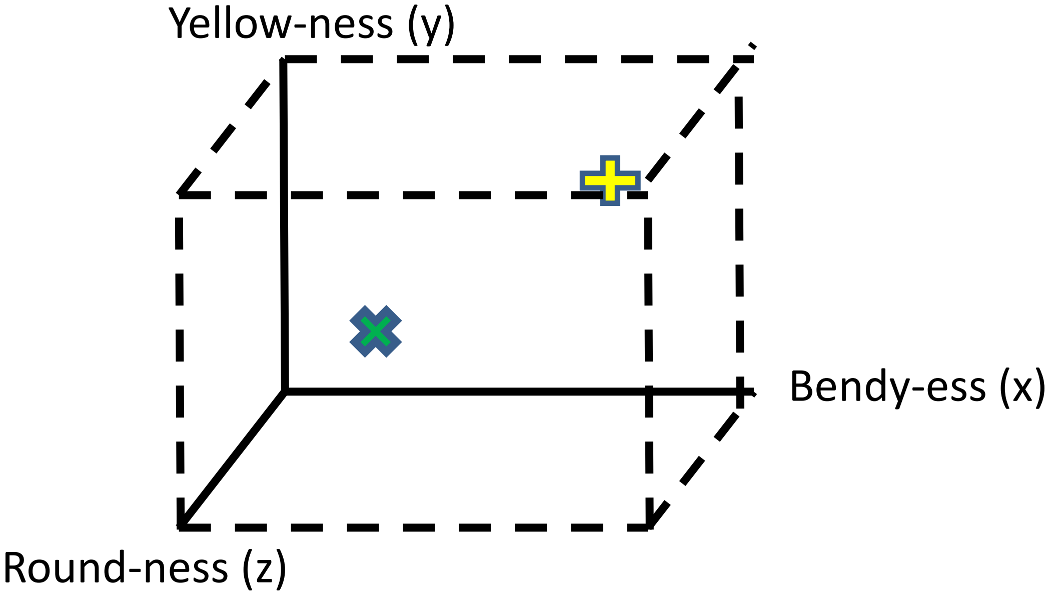


Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis

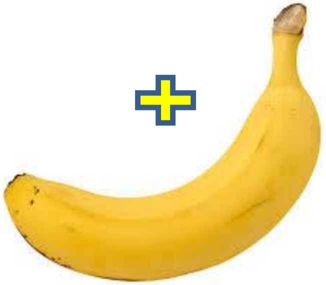


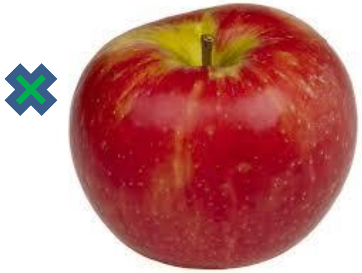


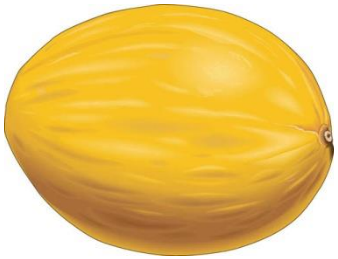


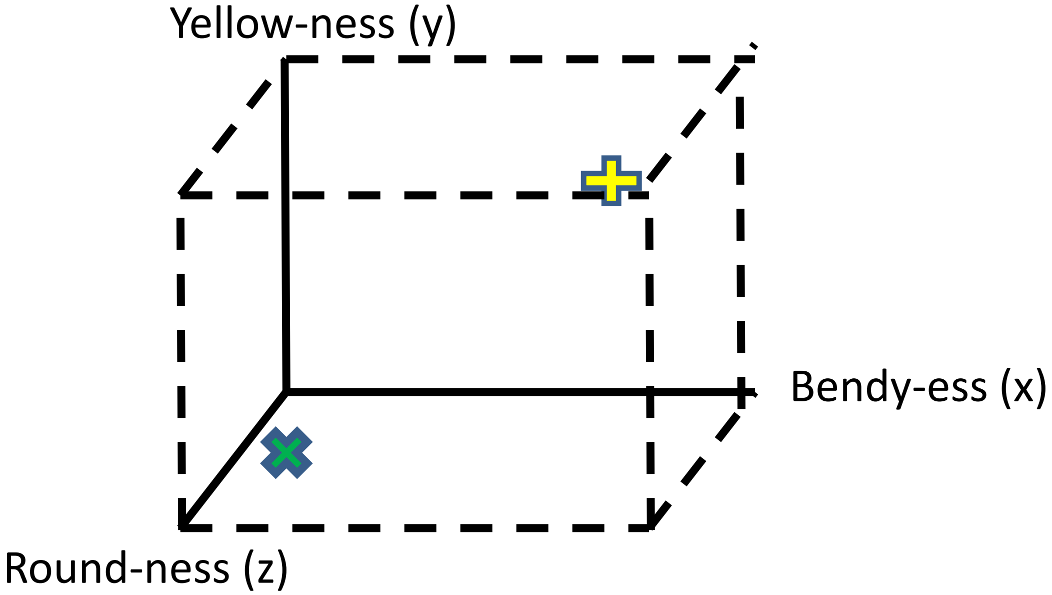
Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis



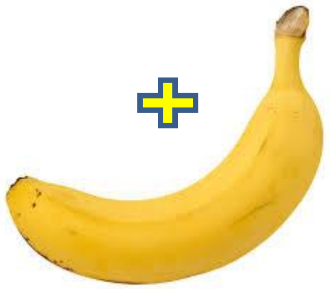


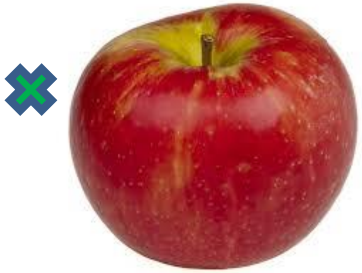


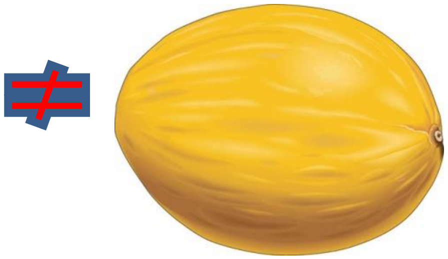


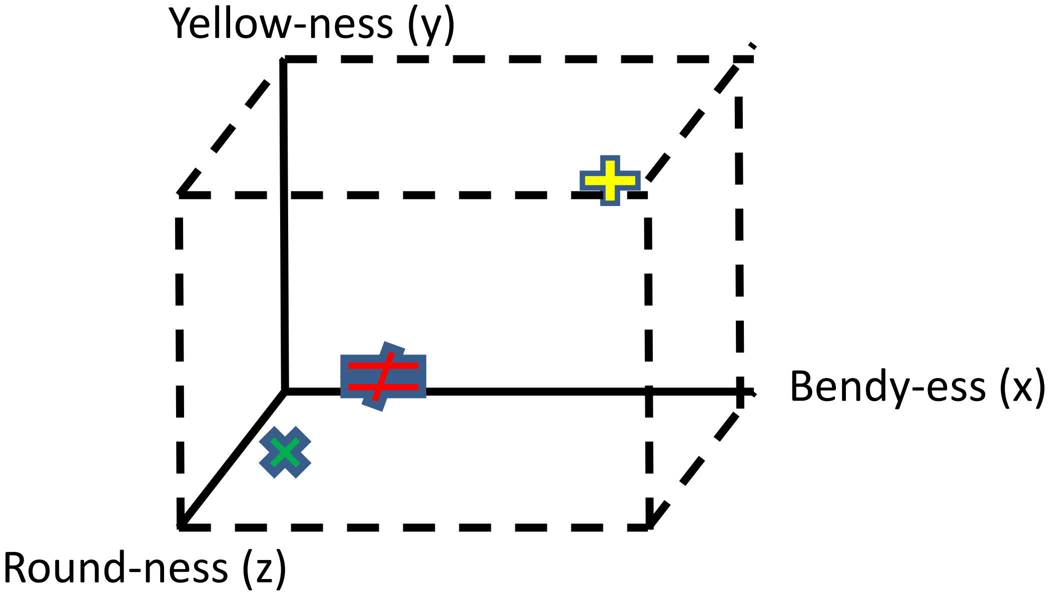
Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis



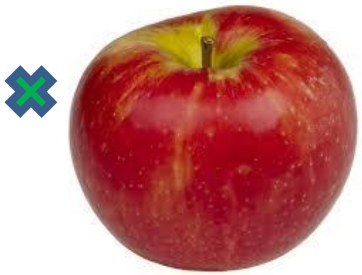


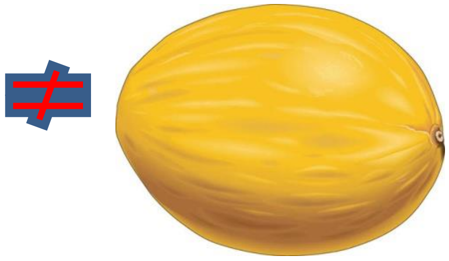


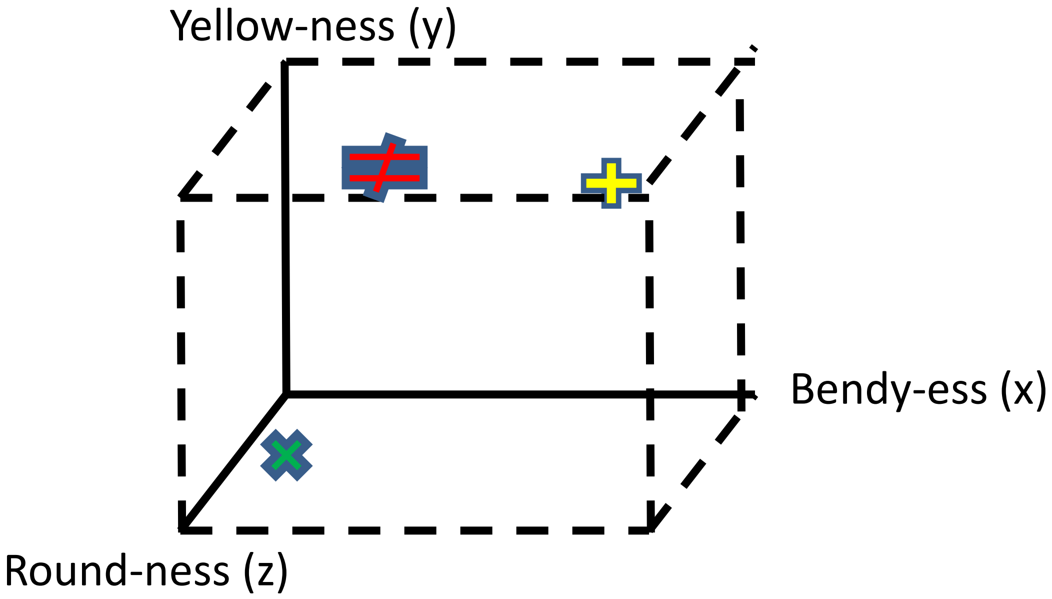


Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis

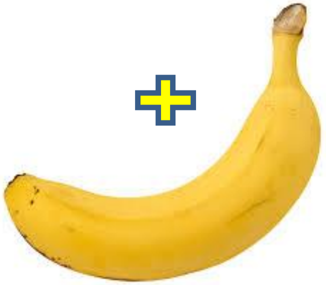


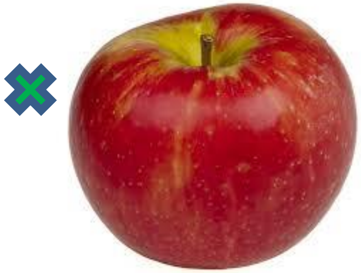


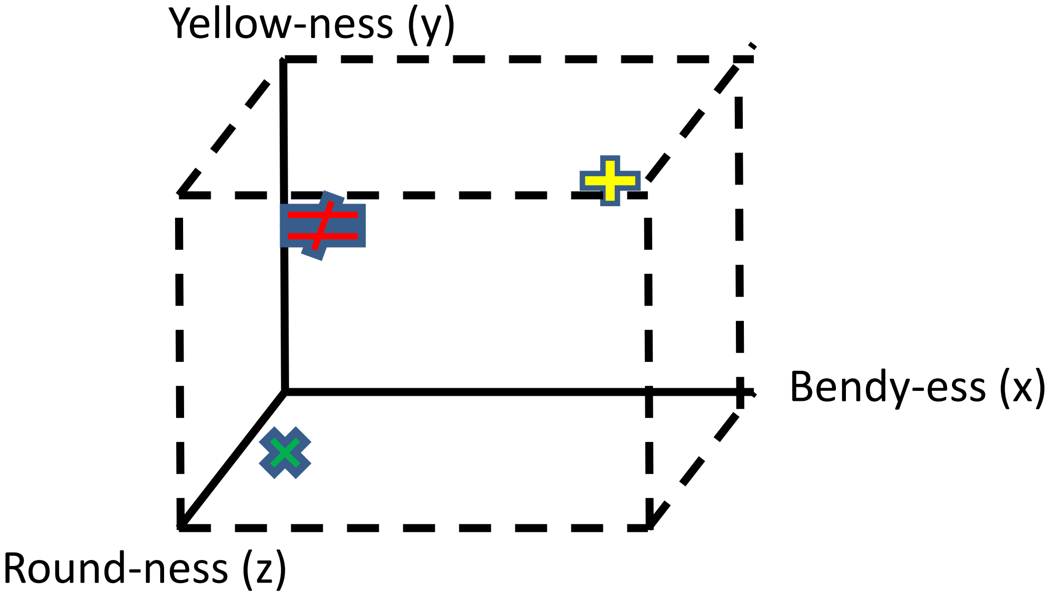


Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis

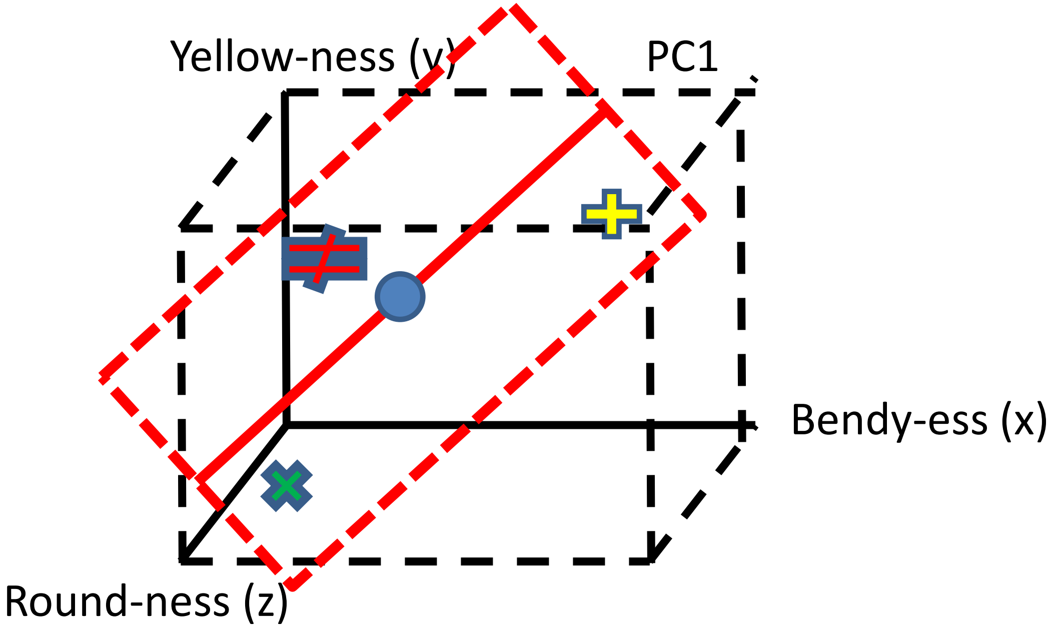






Bendy-ess:	8	2	3
Yellow-ness:	9	2	7
Round-ness:	3	8	5

PCA – Principal Component Analysis

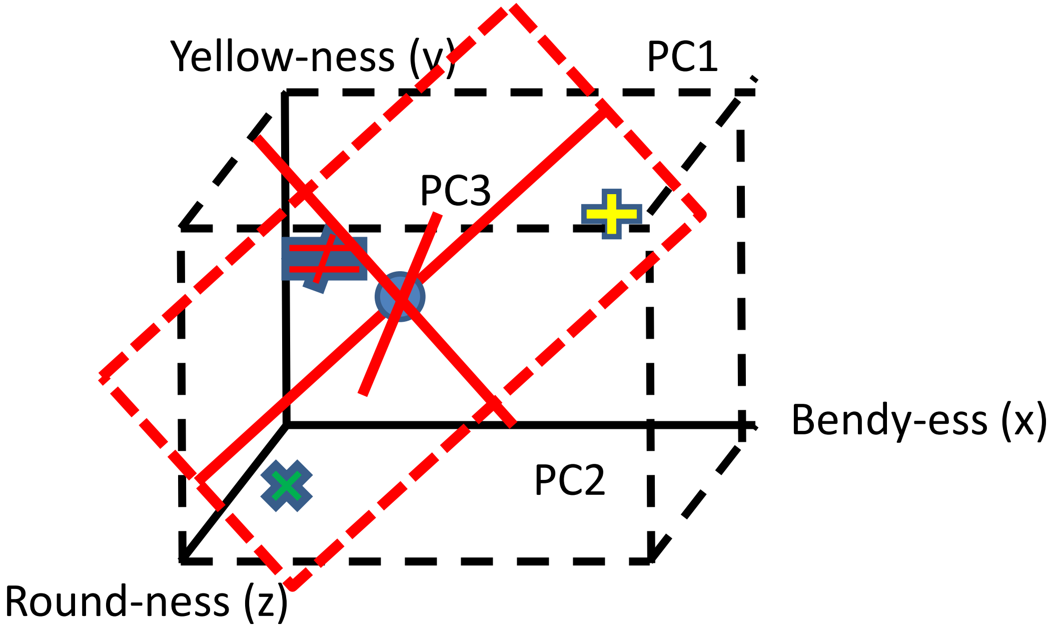


A 2D window is inserted over this 3D space (or in real PCA datasets over all k-dimensions) that covers the most variation of the data set

A line is then placed through this window and a dot is placed in the **centre (mean centred)**

This is what we call PC (principal component) 1 !

PCA – Principal Component Analysis

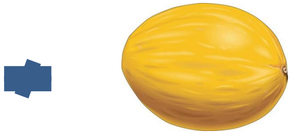


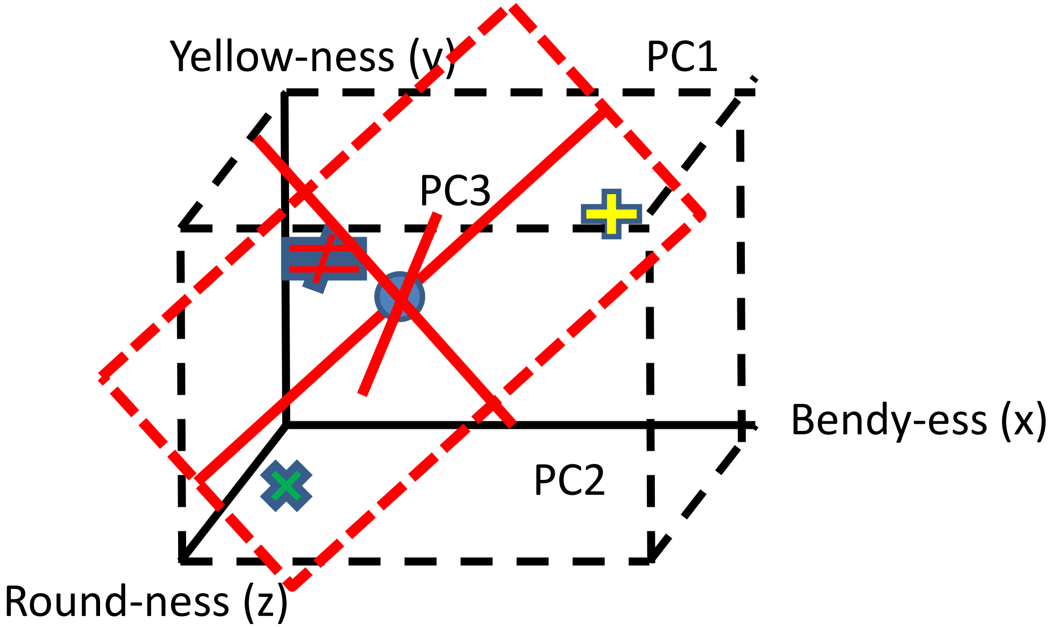
The second PC is calculated by looking at the variation at 90 ° to the first PC

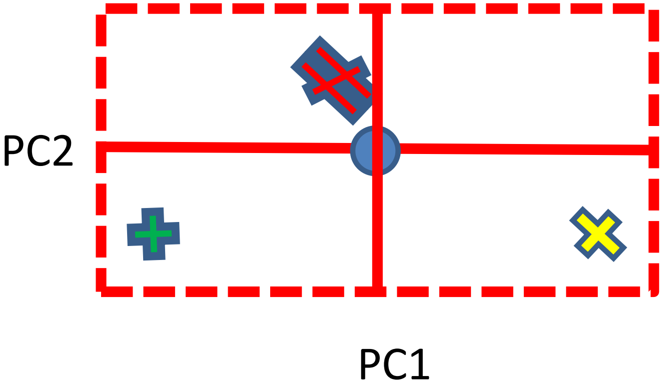
The third PC is calculated by looking at the variation at 90 ° to the second PC

Each subsequent PC lies in an orthogonal direction of maximum variance that has not been considered by the former components.

PCA – Principal Component Analysis







Rotating the window converts the multidimension data to a 2D PCA plot
(this is called a score scatter plot)

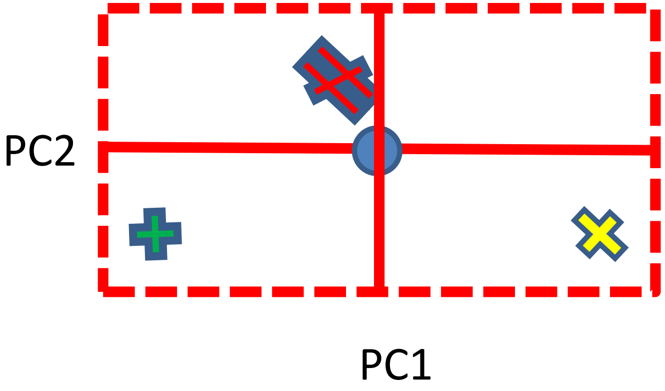
How do we know what causes each point to be in each position? – loadings plot

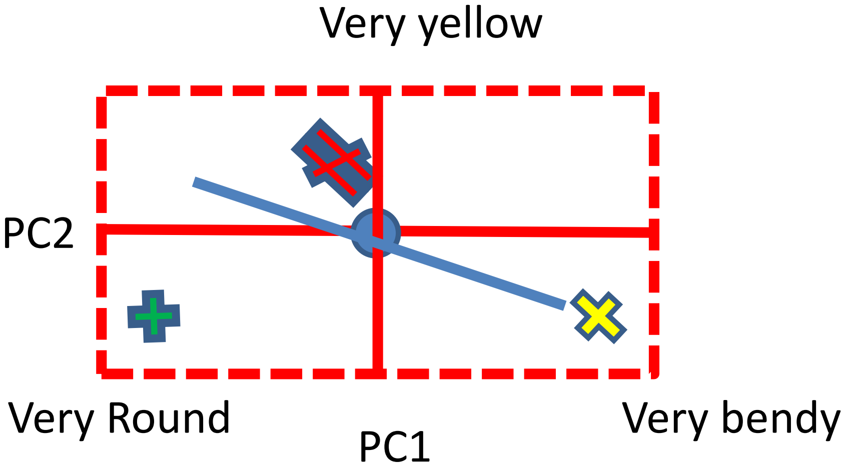
PCA – Principal Component Analysis

How do we see what causes each point to be in each position? – loadings plot
Displays the relationships among the variables

Loading score for each measurement is given between 0-1 (essentially an R2)

Are inversely related measurement – what is high in bendy is low in opposite measurement

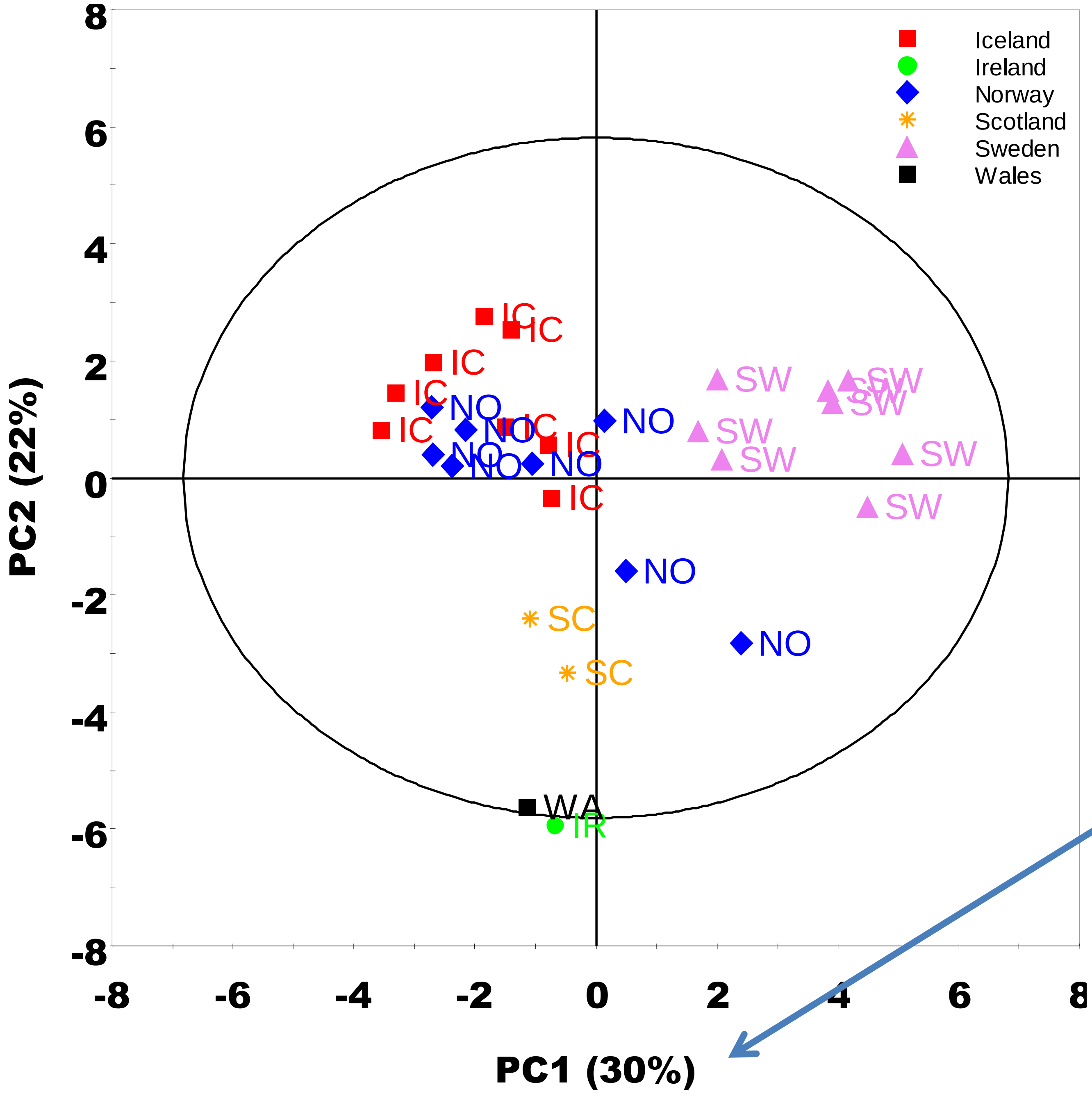




PCA Score Scatter Plot

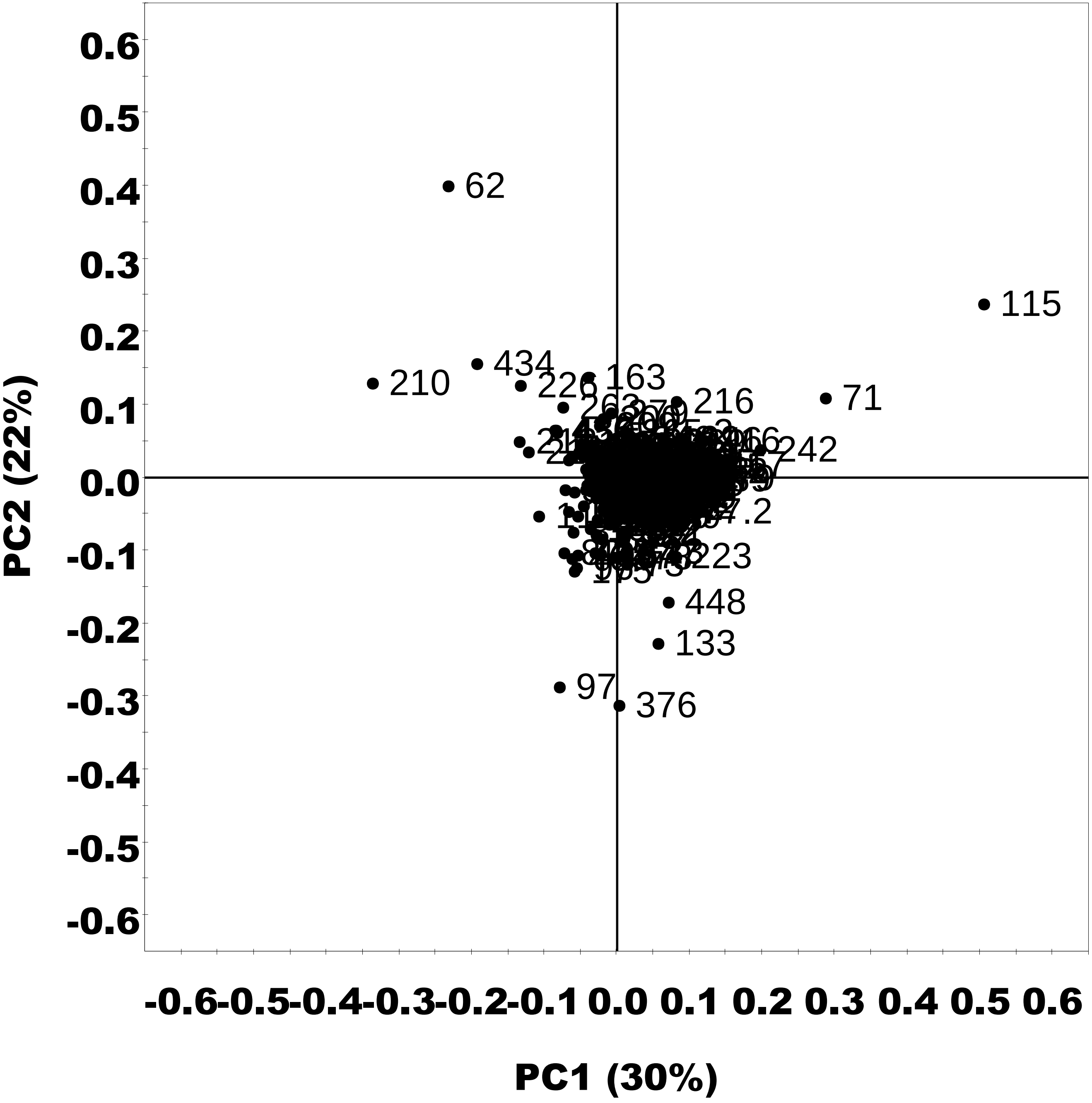
Aqueous phase
Negative ionisation

Direct Injection Mass Spectrometry (DIMS)



30% of ALL the variation in the masses detected ALL samples is explained in PC1

Davey et al. 2008/2009

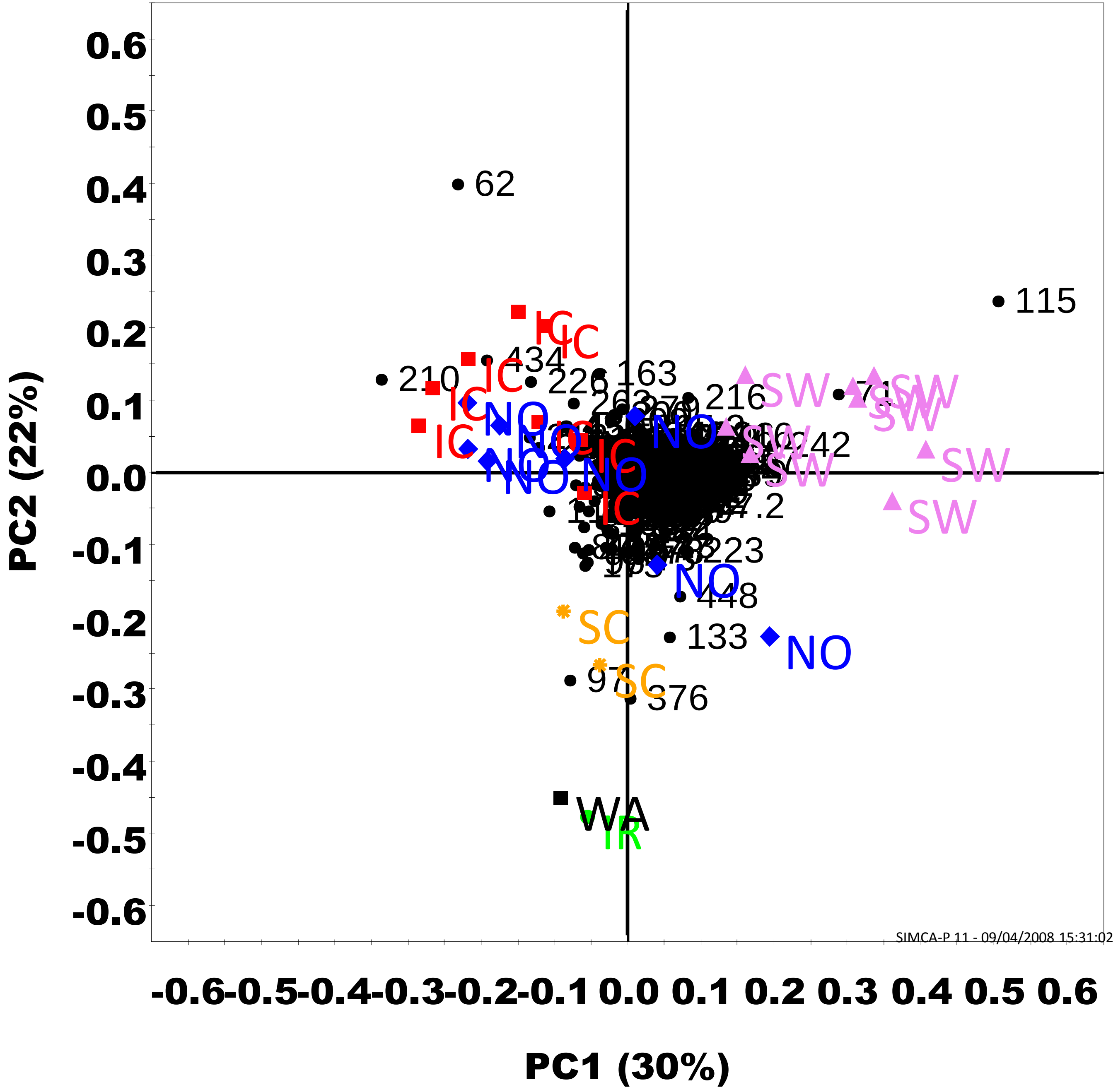


PCA Score Loadings Plot

0.2Da bins

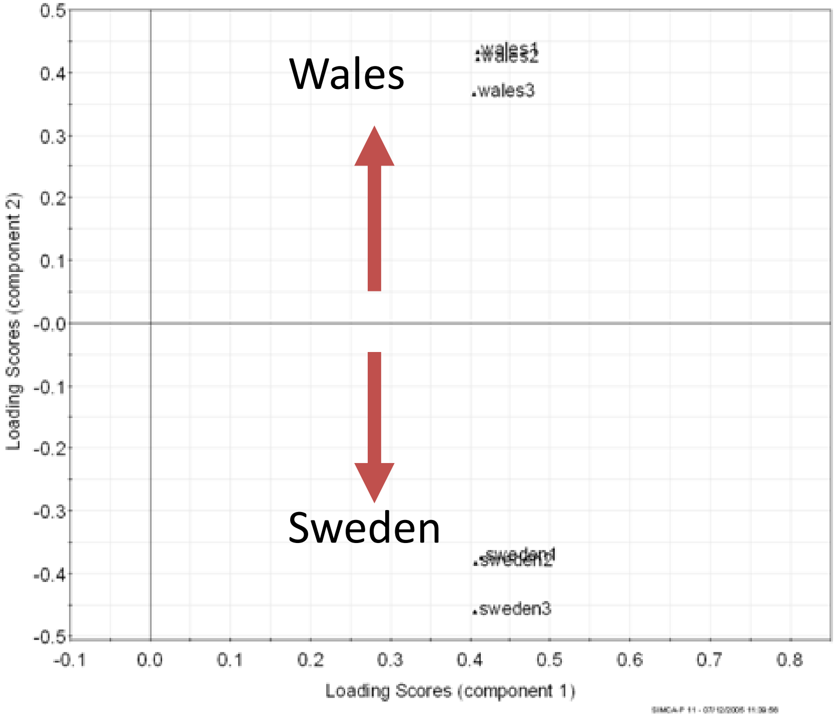
SIMCA-P 11 - 09/04/2008 15:26:59

Overlay PCA Score Scatter and Loadings Plot



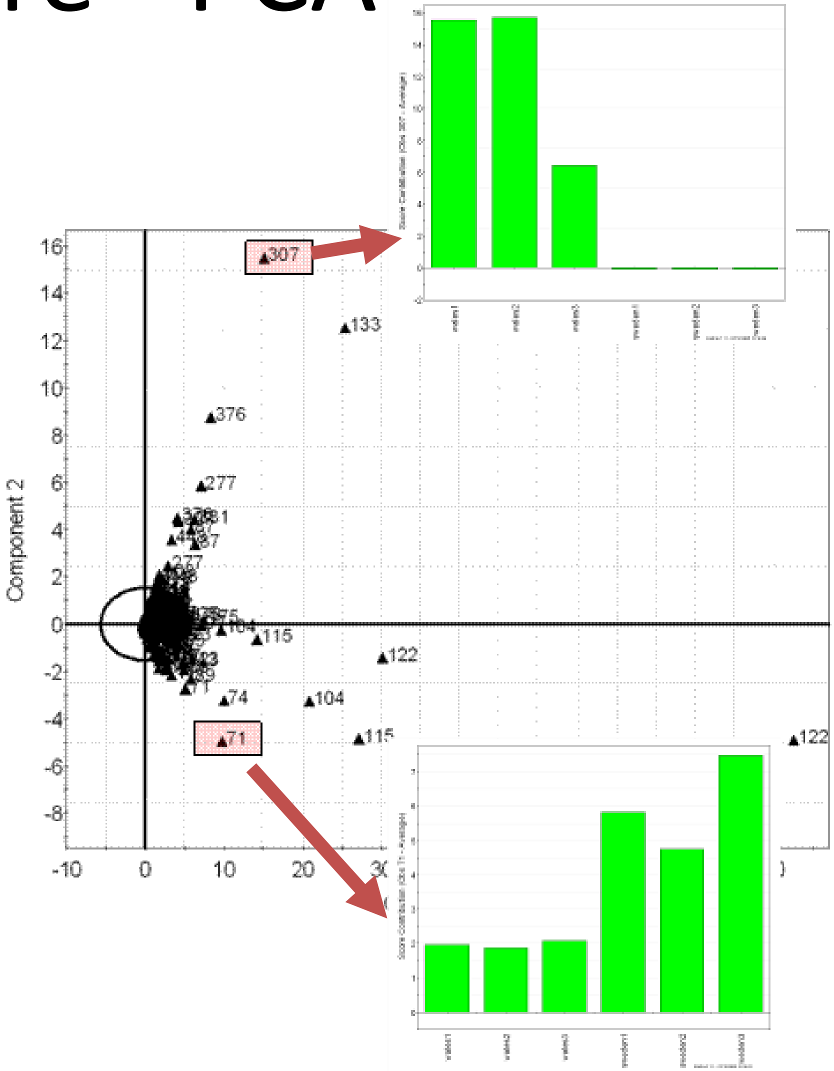
SIMCA-P 11 - 09/04/2008 15:26:59

Simca-P software - PCA



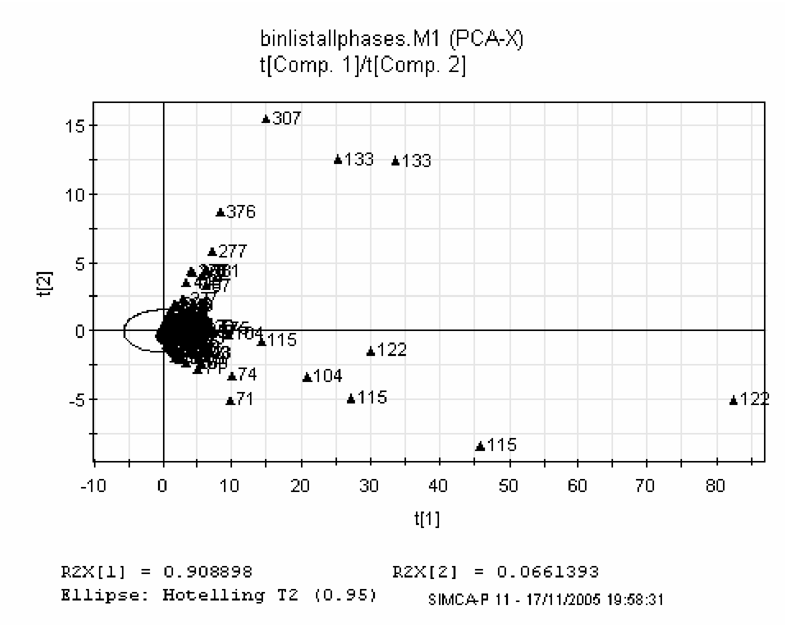
Wales

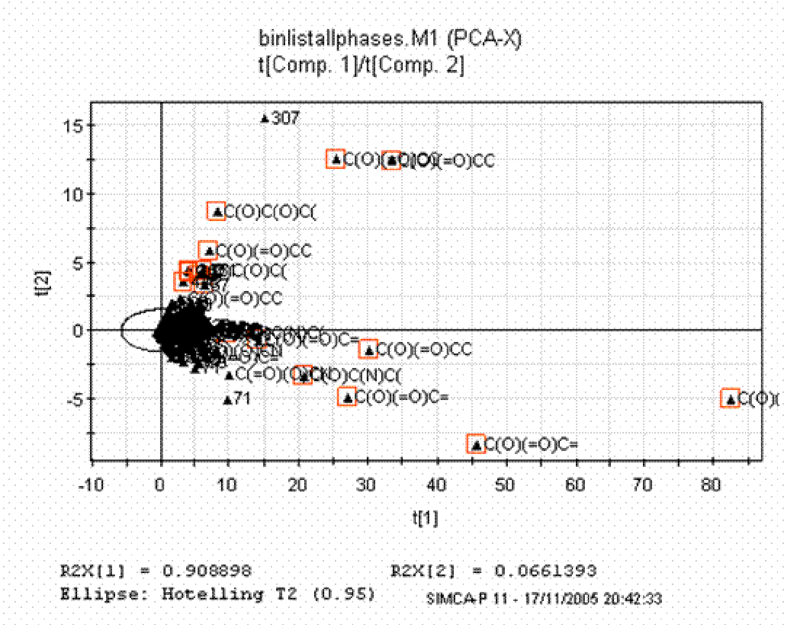
Sweden



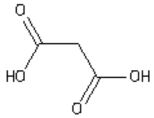
Data mining – search masses in
our metabolite database

SMILE STRUCTURE REPRESENTATION

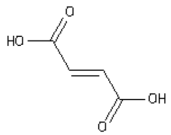




M1.t[1] = 82.5556
M1.t[2] = -4.89724
Primary ID = 2431
MASS 122 – Malonate

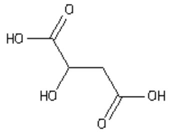


M1.t[1] = 45.7915
M1.t[2] = -8.28212
Primary ID = 448
MASS 115 – Fumaric acid



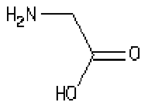
M1.t[1] = 25.4171
M1.t[2] = 12.544
Primary ID = 1786

MASS 133 – Malic acid



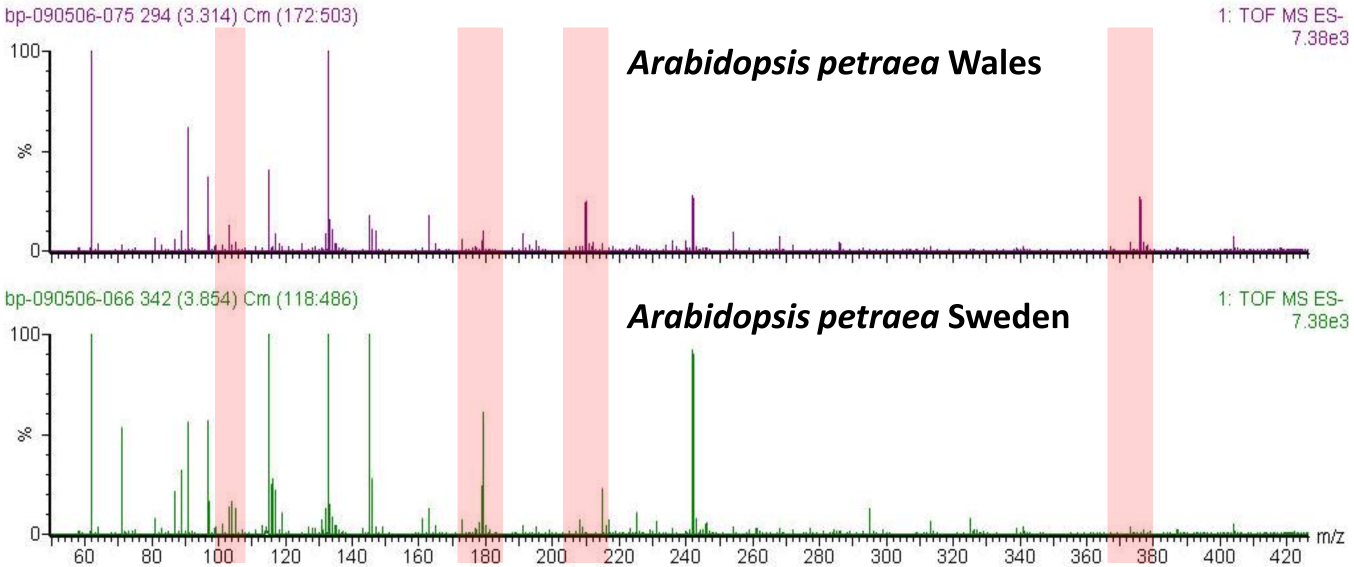
M1.t[1] = 9.94449
M1.t[2] = -3.22985
Primary ID = 1730

MASS 74 - Glycine



Scaling data

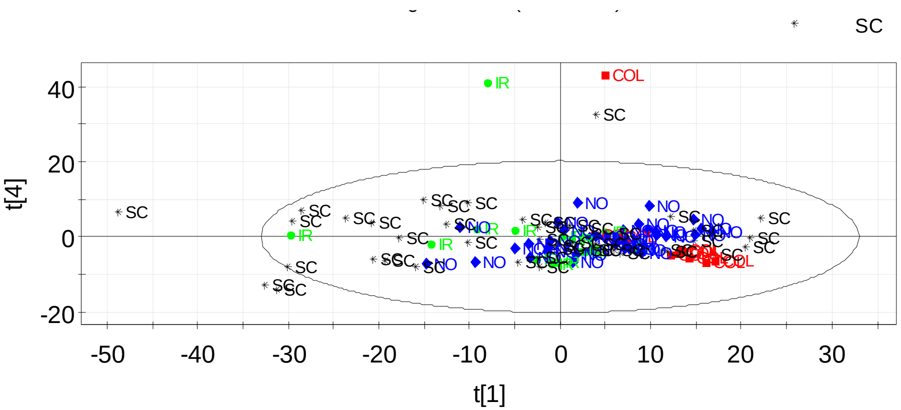
Reducing bias against very intense peaks
-need to highlight fold/relative intensities among samples of the same peak



Scaling data

Reducing bias against very intense peaks

-need to highlight fold/relative intensities among samples of the same peak



Unit variance scaling

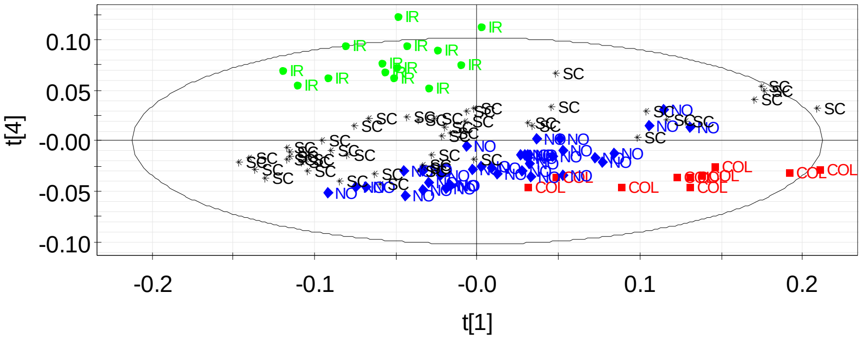
Most objective – takes data as face value

Pareto Scaling

Best for MS, NMR

Decreases importance of high intensity peaks

Depends on question…



How many components should I look at?

How many components should be included in the model?

Degree of fit and the predictive ability

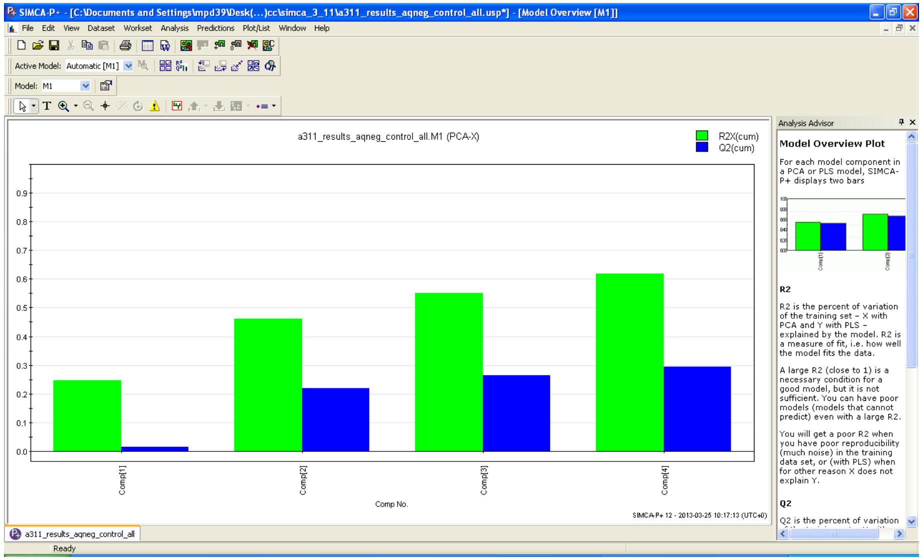
Fit = how well we are able to mathematically reproduce the data of a training set (goodness of fit) ***R^2X*** = the explained variation (0-1)

Predictive ability = how accurately can we predict the raw X data? (goodness of predictability) ***Q^2X***

Take out 25% of samples, does the PCA still look the same?
(a good reason to have many samples!)

Does adding more components make the model better?

Poor *R^2X* and *Q^2X if very noisy dataset or too few replicates*


SIMCA-P+ - [C:\Documents and Settings\mpd39\Desk(...)cc\simca_3_11\a311_results_aqneg_control_all.usp*] - [Model Overview [M1]]

Active Model: Automatic [M1]

Model: M1

a311_results_aqneg_control_all.M1 (PCA-X)

SIMCA-P+ 12 - 2013-03-25 10:17:13 (UTC+0)

Analysis Advisor

Model Overview Plot

For each model component in a PCA or PLS model, SIMCA-P+ displays two bars

R2

R2 is the percent of variation of the training set – X with PCA and Y with PLS – explained by the model. R2 is a measure of fit, i.e. how well the model fits the data.

A large R2 (close to 1) is a necessary condition for a good model, but it is not sufficient. You can have poor models (models that cannot predict) even with a large R2.

You will get a poor R2 when you have poor reproducibility (much noise) in the training data set, or (with PLS) when for other reason X does not explain Y.

Q2

Q2 is the percent of variation

a311_results_aqneg_control_all

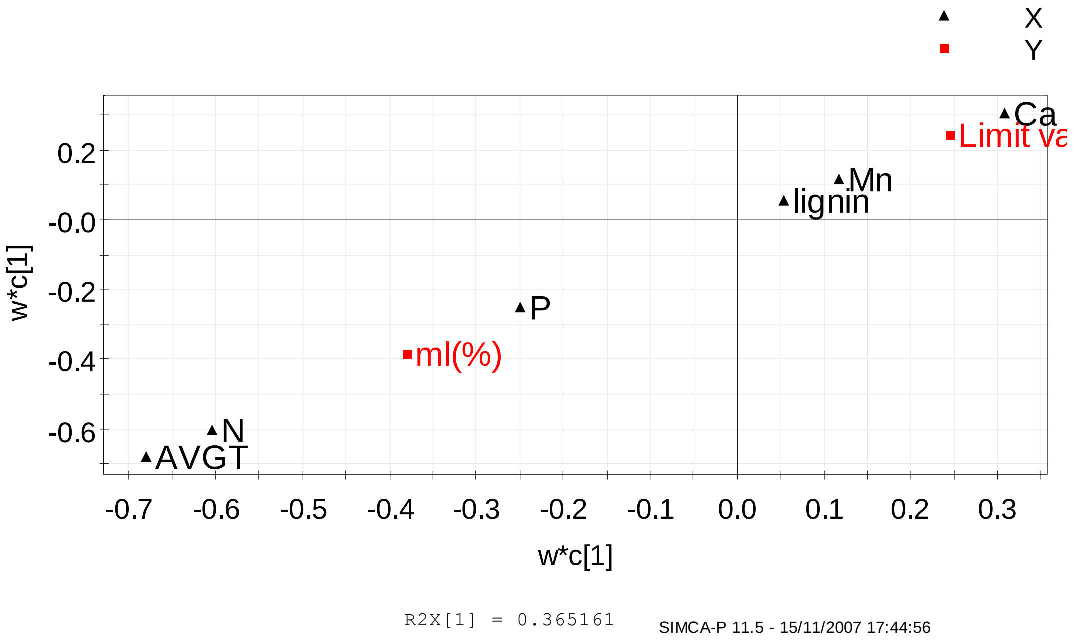
Ready

•PLS-DA – projections to latent structures by partial least squares – discriminate analysis

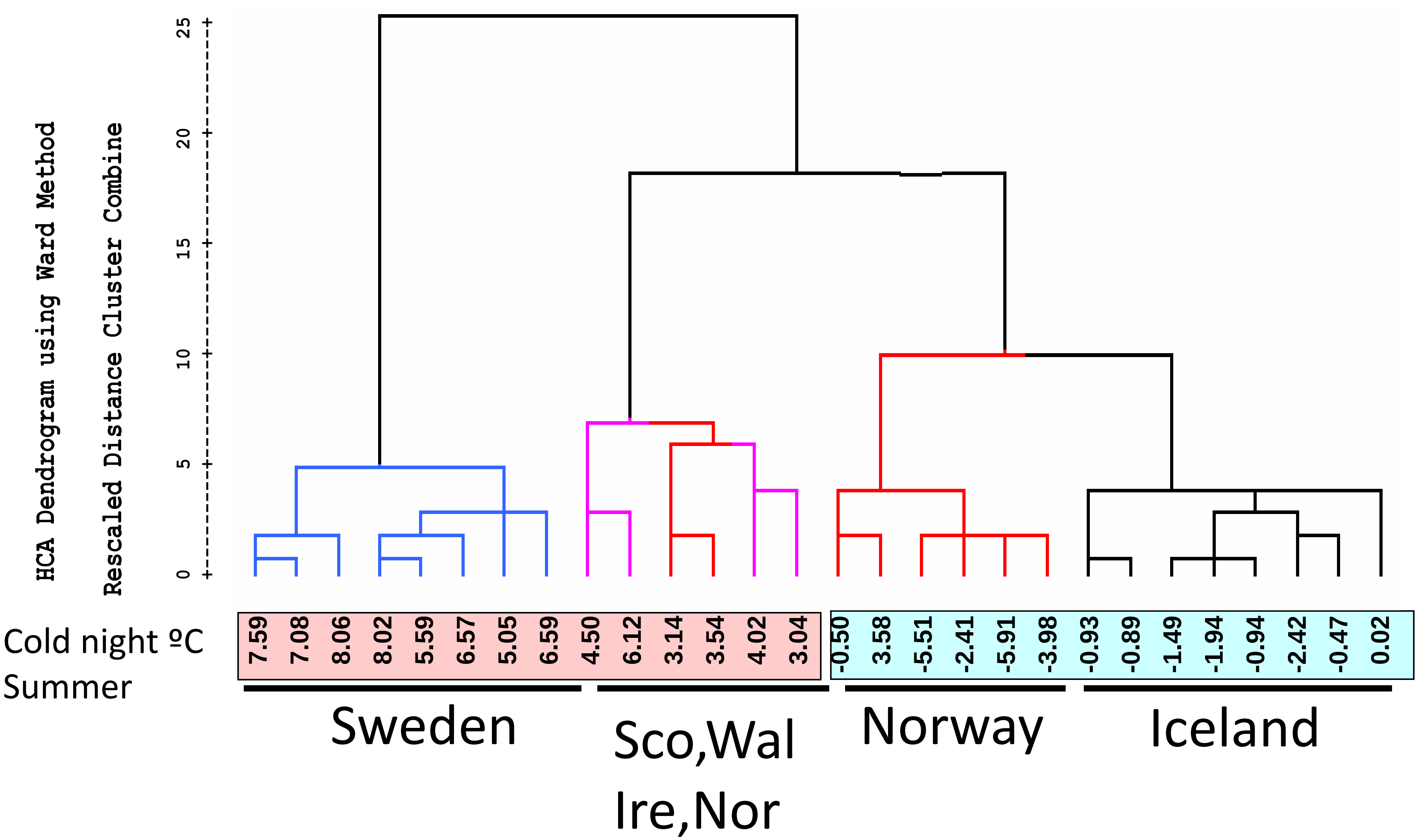
•Similar to PCA

•Supervised method – good for discovering biomarkers

•Can connect other data (Y) with your MS/NMR dataset (X)
– you tell the model that some samples are from control, others treated etc.



HCA – Hierarchical Cluster Analysis



•ANOVA, false-positive corrections

Which masses are statistically significantly different between our samples?

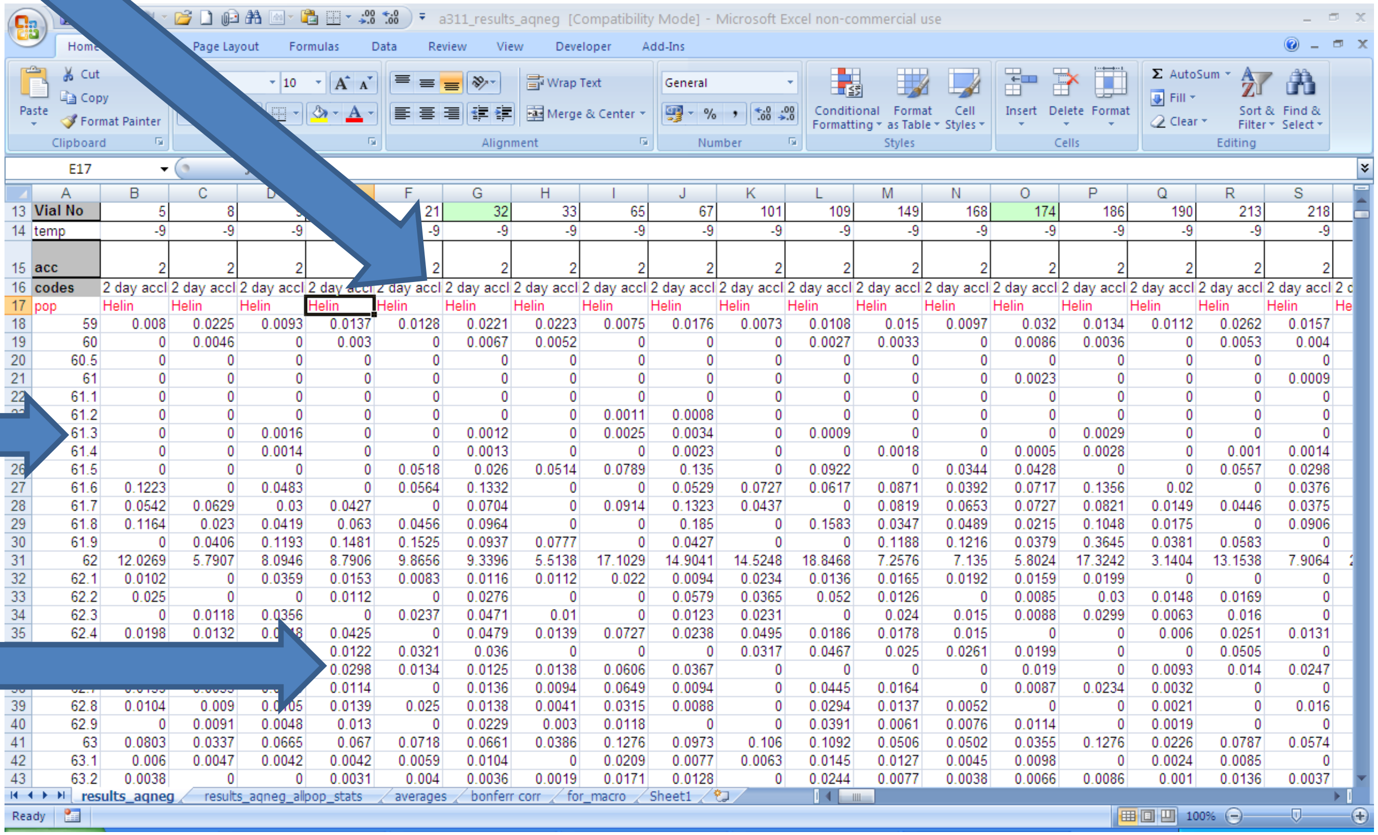
T-test

ANOVA

Many samples (10’s or 100’s)

Many bins (MW) 1000’s

Many intensities 10,000’s



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
13	Vial No	5	8	[illegible]	[illegible]	21	32	33	65	67	101	109	149	168	174	186	190	213	218
14	temp	-9	-9	-9	[illegible]	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9	-9
15	acc	2	2	2	[illegible]	2	2	2	2	2	2	2	2	2	2	2	2	2	2
16	codes	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl	2 day accl
17	pop	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin	Helin
18	59	0.008	0.0225	0.0093	0.0137	0.0128	0.0221	0.0223	0.0075	0.0176	0.0073	0.0108	0.015	0.0097	0.032	0.0134	0.0112	0.0262	0.0157
19	60	0	0.0046	0	0.003	0	0.0067	0.0052	0	0	0	0.0027	0.0033	0	0.0086	0.0036	0	0.0053	0.004
20	60.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	61	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0023	0	0	0	0.0009
22	61.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[illegible]	61.2	0	0	0	0	0	0	0	0.0011	0.0008	0	0	0	0	0	0	0	0	0
[illegible]	61.3	0	0	0.0016	0	0	0.0012	0	0.0025	0.0034	0	0.0009	0	0	0	0.0029	0	0	0
[illegible]	61.4	0	0	0.0014	0	0	0.0013	0	0	0.0023	0	0	0.0018	0	0.0005	0.0028	0	0.001	0.0014
26	61.5	0	0	0	0	0.0518	0.026	0.0514	0.0789	0.135	0	0.0922	0	0.0344	0.0428	0	0	0.0557	0.0298
27	61.6	0.1223	0	0.0483	0	0.0564	0.1332	0	0	0.0529	0.0727	0.0617	0.0871	0.0392	0.0717	0.1356	0.02	0	0.0376
28	61.7	0.0542	0.0629	0.03	0.0427	0	0.0704	0	0.0914	0.1323	0.0437	0	0.0819	0.0653	0.0727	0.0821	0.0149	0.0446	0.0375
29	61.8	0.1164	0.023	0.0419	0.063	0.0456	0.0964	0	0	0.185	0	0.1583	0.0347	0.0489	0.0215	0.1048	0.0175	0	0.0906
30	61.9	0	0.0406	0.1193	0.1481	0.1525	0.0937	0.0777	0	0.0427	0	0	0.1188	0.1216	0.0379	0.3645	0.0381	0.0583	0
31	62	12.0269	5.7907	8.0946	8.7906	9.8656	9.3396	5.5138	17.1029	14.9041	14.5248	18.8468	7.2576	7.135	5.8024	17.3242	3.1404	13.1538	7.9064
32	62.1	0.0102	0	0.0359	0.0153	0.0083	0.0116	0.0112	0.022	0.0094	0.0234	0.0136	0.0165	0.0192	0.0159	0.0199	0	0	0
33	62.2	0.025	0	0	0.0112	0	0.0276	0	0	0.0579	0.0365	0.052	0.0126	0	0.0085	0.03	0.0148	0.0169	0
34	62.3	0	0.0118	0.0356	0	0.0237	0.0471	0.01	0	0.0123	0.0231	0	0.024	0.015	0.0088	0.0299	0.0063	0.016	0
35	62.4	0.0198	0.0132	[illegible]	0.0425	0	0.0479	0.0139	0.0727	0.0238	0.0495	0.0186	0.0178	0.015	0	0	0.006	0.0251	0.0131
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0122	0.0321	0.036	0	0	0	0.0317	0.0467	0.025	0.0261	0.0199	0	0	0.0505	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0298	0.0134	0.0125	0.0138	0.0606	0.0367	0	0	0	0	0.019	0	0.0093	0.014	0.0247
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0114	0	0.0136	0.0094	0.0649	0.0094	0	0.0445	0.0164	0	0.0087	0.0234	0.0032	0	0
39	62.8	0.0104	0.009	[illegible]	0.0139	0.025	0.0138	0.0041	0.0315	0.0088	0	0.0294	0.0137	0.0052	0	0	0.0021	0	0.016
40	62.9	0	0.0091	0.0048	0.013	0	0.0229	0.003	0.0118	0	0	0.0391	0.0061	0.0076	0.0114	0	0.0019	0	0
41	63	0.0803	0.0337	0.0665	0.067	0.0718	0.0661	0.0386	0.1276	0.0973	0.106	0.1092	0.0506	0.0502	0.0355	0.1276	0.0226	0.0787	0.0574
42	63.1	0.006	0.0047	0.0042	0.0042	0.0059	0.0104	0	0.0209	0.0077	0.0063	0.0145	0.0127	0.0045	0.0098	0	0.0024	0.0085	0
43	63.2	0.0038	0	0	0.0031	0.004	0.0036	0.0019	0.0171	0.0128	0	0.0244	0.0077	0.0038	0.0066	0.0086	0.001	0.0136	0.0037

•ANOVA, false-positive corrections

Testing so many samples runs the risk of significant values occurring by chance
Change the usual p <0.05 to take into account the number of samples

p<0.05 is usually considered a significant result - error rate of 5 %
(**ie: 1 in every 20 tests gives a false result**)

If a dataset has 1000 metabolites (so 1000 univariate tests…)
50 metabolites are significantly different when they are not

•ANOVA, false-positive corrections

Testing so many samples runs the risk of significant values occurring by chance

Change the usual p <0.05 to take into account the number of samples

p<0.05 is usually considered a significant result - error rate of 5 %

(**ie: 1 in every 20 tests gives a false result**)

If a dataset has 1000 metabolites (so 1000 univariate tests…)

50 metabolites are significantly different when they are not

Two main False-positive corrections

Bonferroni adjusted P-value is: 0.05/number of samples (very strict)

=0.05/4254 P_{adj} = 0.00001175

Benjamini and Hochberg – rank the p-values highest to lowest

P_{adj} = (number of samples/p-value)*position in ranked p-value table

Keep going until you get to P_{adj} of 0.05

Original p-value = 0.011718

P_{adj} = 0.051755

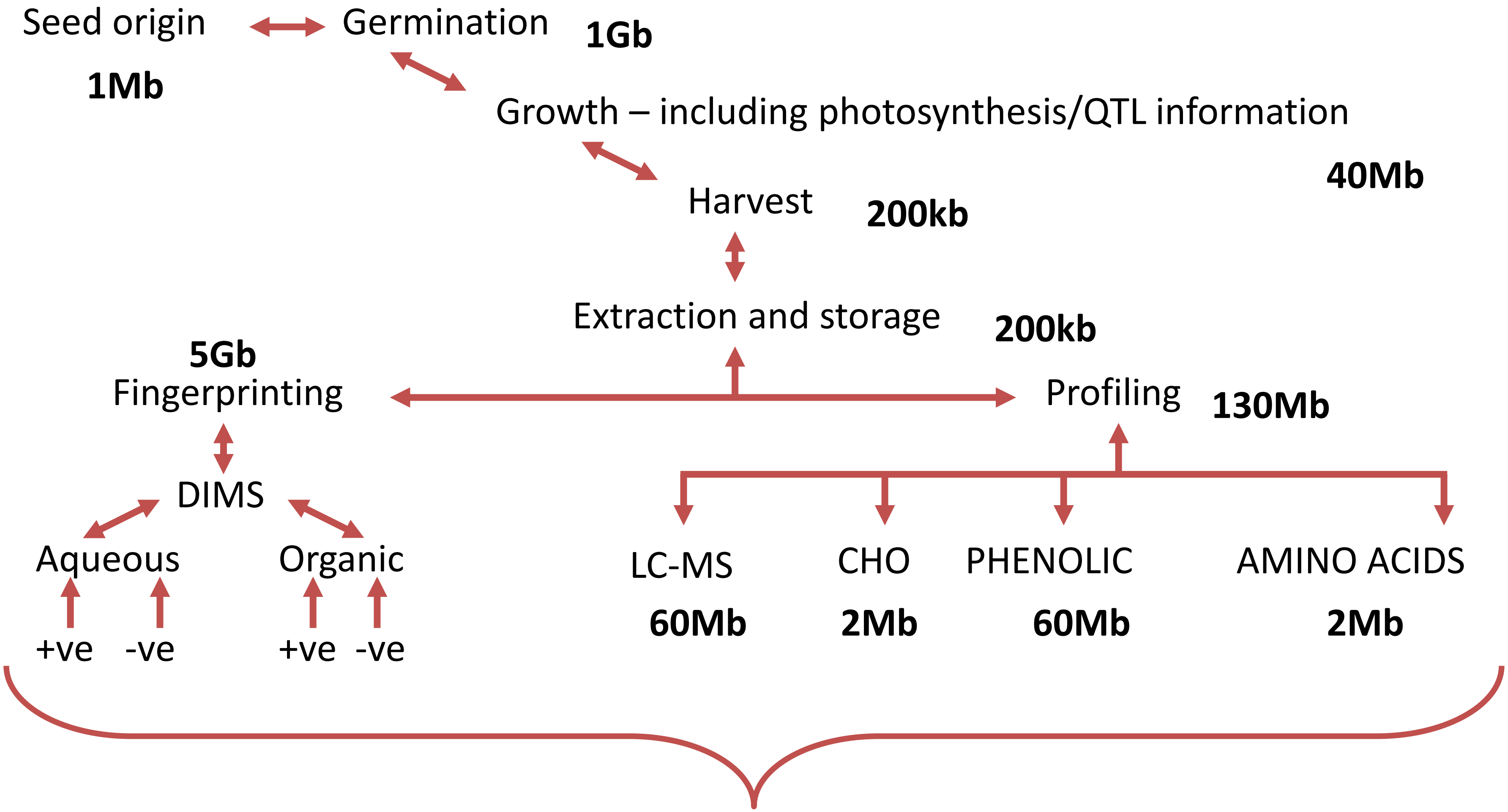
Benjamini and Hochberg. 1995. Journal of the Royal Statistical Society, Series B 57: 289–300.

Introduction to chemometrics

- Metadata – useful for interpreting PCA, Y-var for PLS-DA

	META-DATA
Experiment set-up	Seed origin
	Germination conditions – soil, light, humidity, CO2, day length, temp
	Growth conditions – randonisation etc
	Harvest conditions, time, weight of plant
Gas exchange	Time, growth conditions, IRGA conditions saved per raw sample run
Growth	Plant and cabinet data

Metadata – data storage



Seed origin

1Mb

Germination

1Gb

Growth – including photosynthesis/QTL information

40Mb

Harvest

200kb

Extraction and storage

200kb

5Gb

Fingerprinting

DIMS

Aqueous

+ve -ve

Organic

+ve -ve

Profiling

130Mb

LC-MS

60Mb

CHO

2Mb

PHENOLIC

60Mb

AMINO ACIDS

2Mb

List of masses; compounds; quantity
Chemometrics – PCA, PLS-DA

1Gb

Online practical involving multivariate statistics – PCA etc