

Reflections on data analysis from a non-analyst

Jake Bundy

Introduction

I am not a card-carrying bioinformatician!

My background:

- First degree in chemistry
- Higher degrees in environmental and life sciences

I'm aiming to give a personal perspective only

- Share some experience – my own personal prejudices
- Certainly not an attempt to systematically cover a whole area

Metabolomics/metabonomics is multidisciplinary

- Laboratory/clinical scientist
- Analytical chemist
- Bioinformatician



*Works best when there is
genuine crossover between
the different areas*



Transforming data

Data transformations are often ignored or considered of low importance in metabolomic studies

- Many papers don't even fully report what transformation and/or scaling was done

What tends to be done in practice? (My unscientific impression)

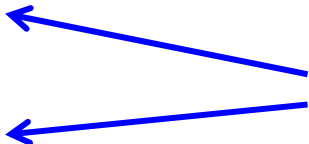
- No scaling (mean-centering only)
- Pareto scaling
- Autoscaling (to unit variance)

Transforming data

Data transformations are often ignored or considered of low importance in metabolomic studies

- Many papers don't even fully report what transformation and/or scaling was done

What tends to be done in practice? (My unscientific impression)

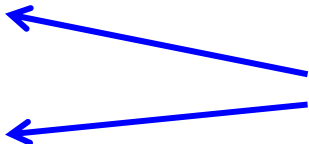
- No scaling (mean-centering only)
 - Pareto scaling
 - Autoscaling (to unit variance)
- Well-defined reason for picking this option*
- 

Transforming data

Data transformations are often ignored or considered of low importance in metabolomic studies

- Many papers don't even fully report what transformation and/or scaling was done

What tends to be done in practice? (My unscientific impression)

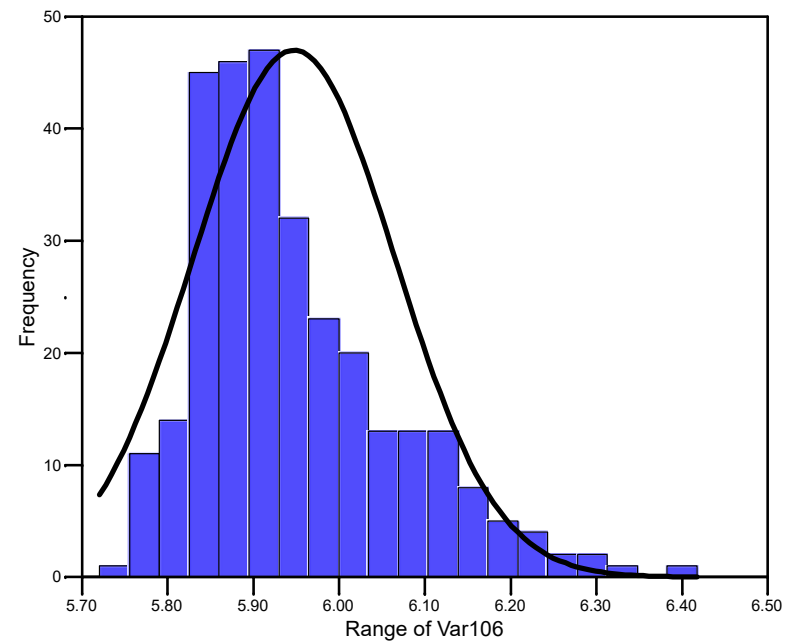
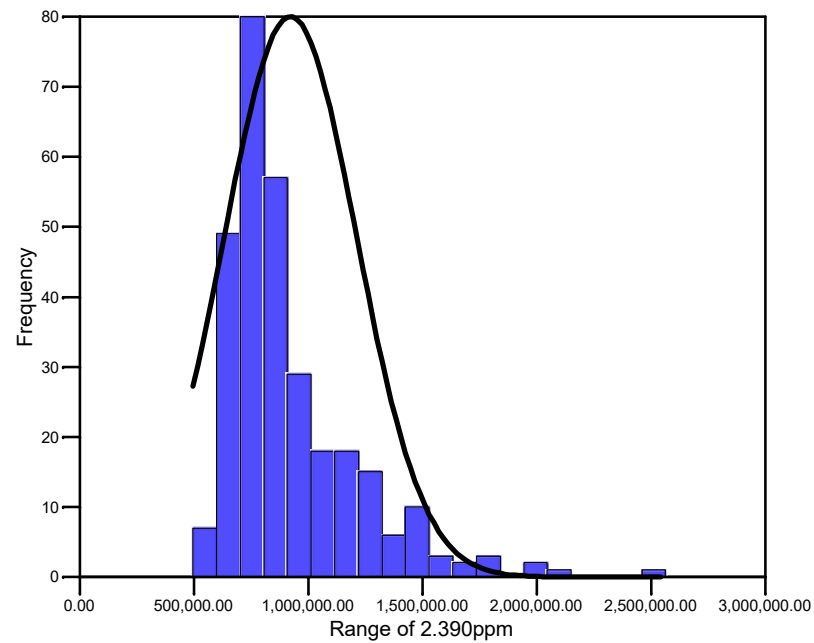
- No scaling (mean-centering only)
 - Pareto scaling
 - Autoscaling (to unit variance)
- 
- Well-defined reason
for picking this option*

What is often left out?

- **Log transformation not used as frequently as you might expect**

Log-transformation of data

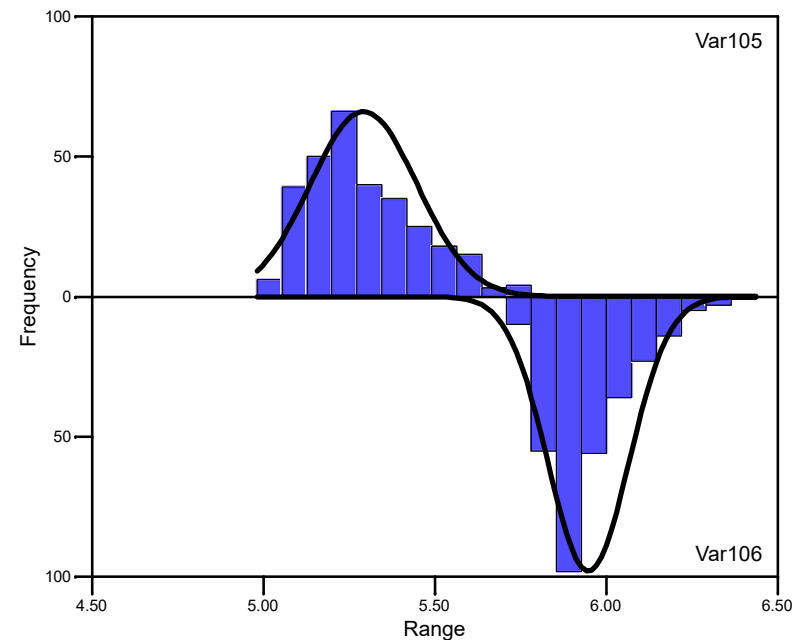
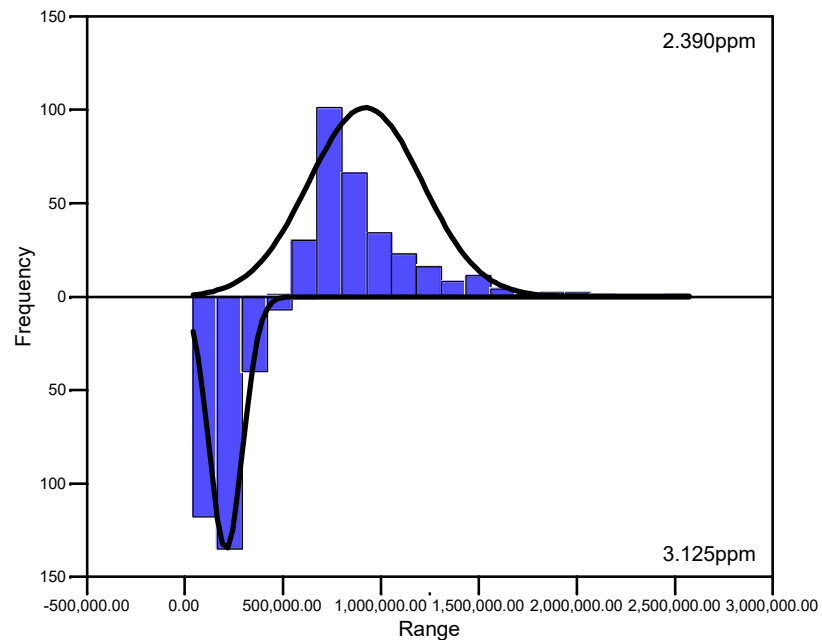
Obviously, converts log-normal distribution to a normal one



Log-transformation of data

Obviously, converts log-normal distribution to a normal one

But also makes it easier to compare between two different variables



Log-transformation of data

Obviously, converts log-normal distribution to a normal one

But also makes it easier to compare between two different variables

What effect does it have across a multivariate dataset?

Variance-stabilizing transformations

One of the assumptions behind many multivariate analyses such as PCA is that data have equal variance

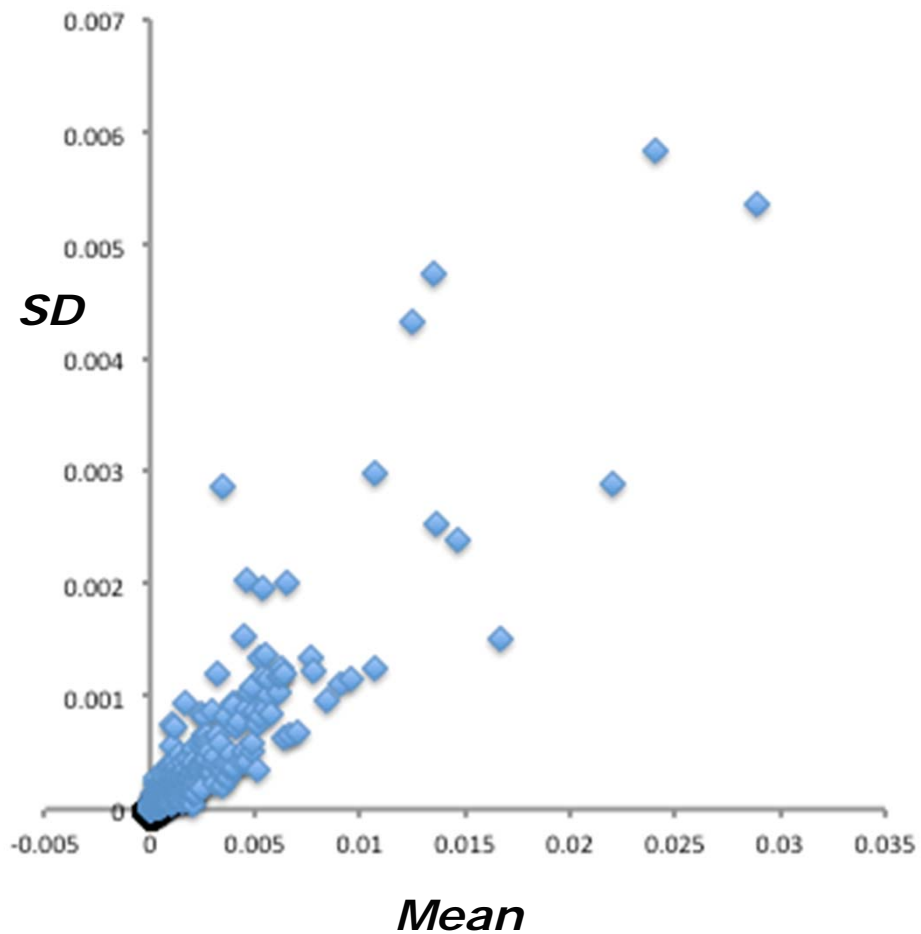
- Durbin and Rocke error model: generalized logarithmic (glog) transformation

$$y = \log \left(x + \sqrt{x^2 + c} \right)$$

- c is a constant based on the instrumental error, but it is usually estimated from experimental samples
 - Parsons et al. *BMC Bioinformatics* 8:234.
- However, started-log is simpler and I find it works essentially just as well:

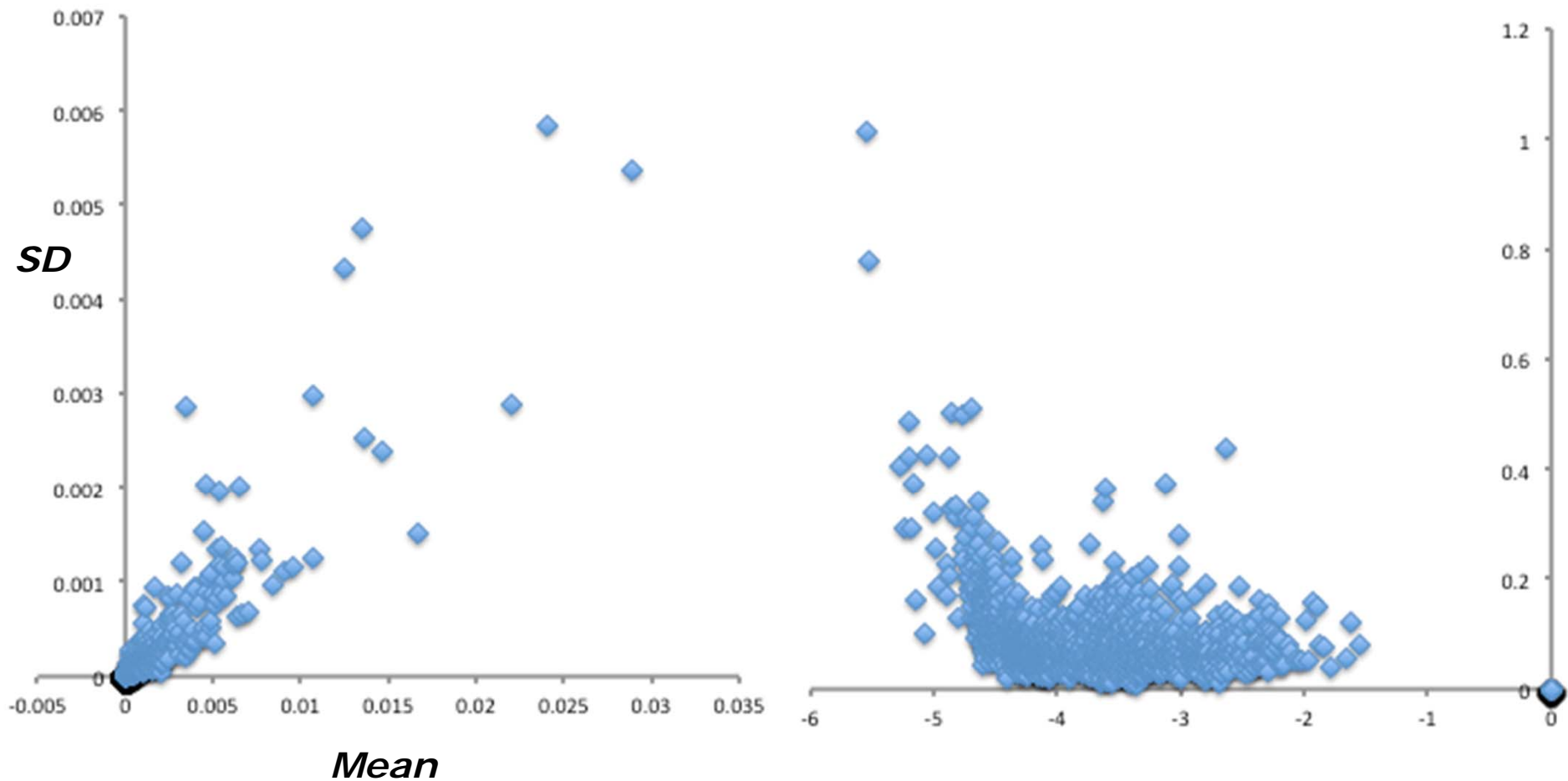
$$y = \log(x + c)$$

Variance-stabilizing transformations

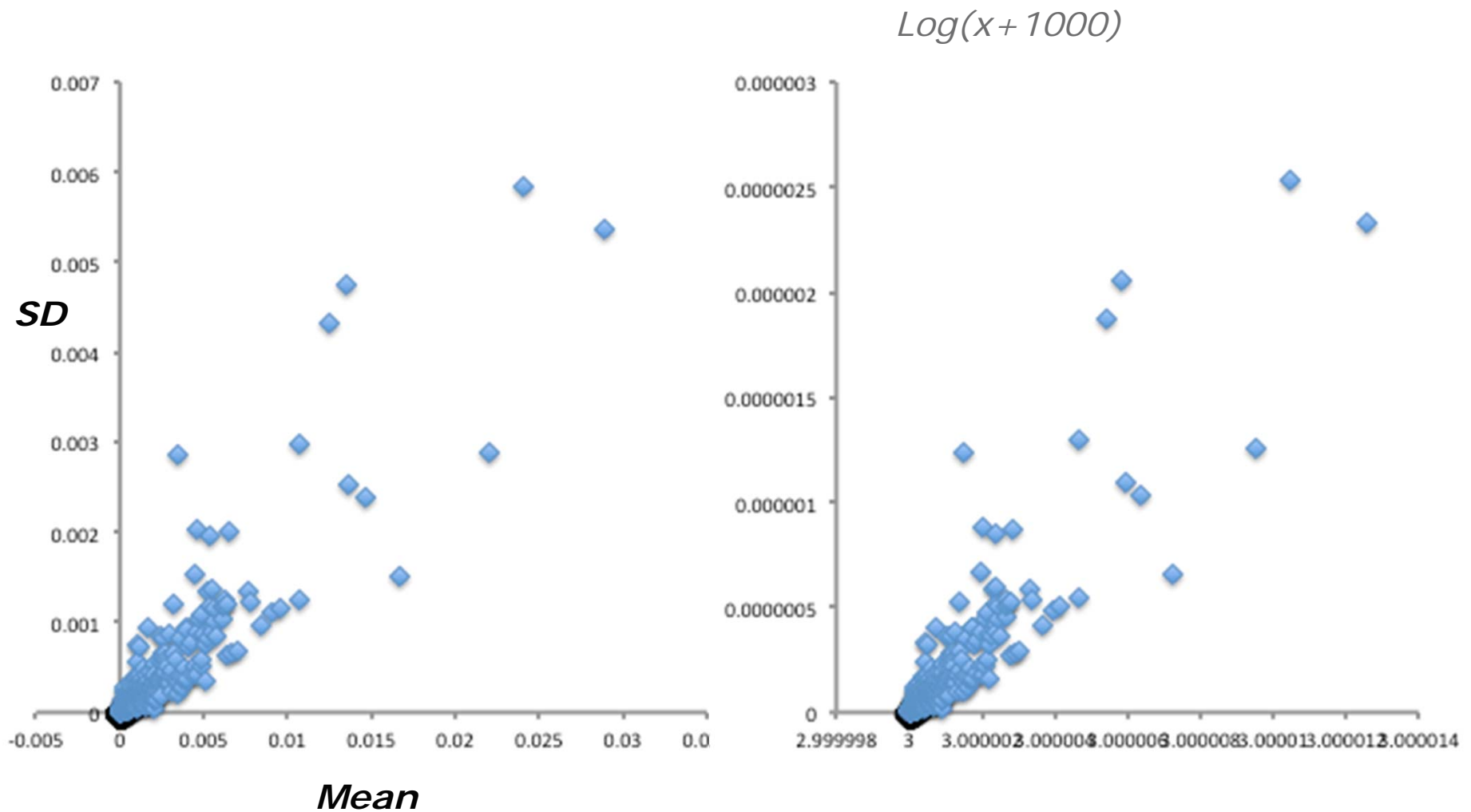


Variance-stabilizing transformations

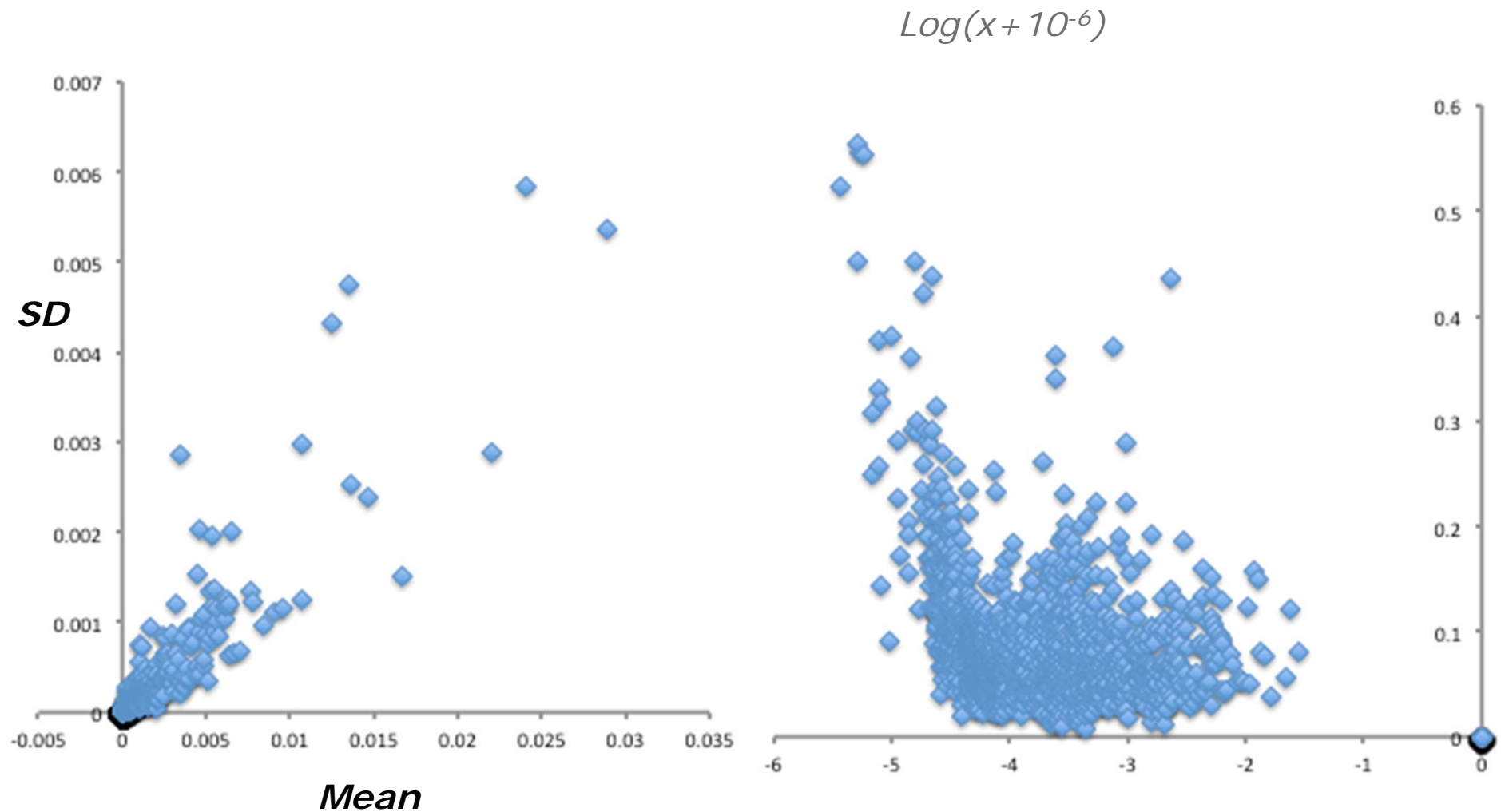
After log transformation



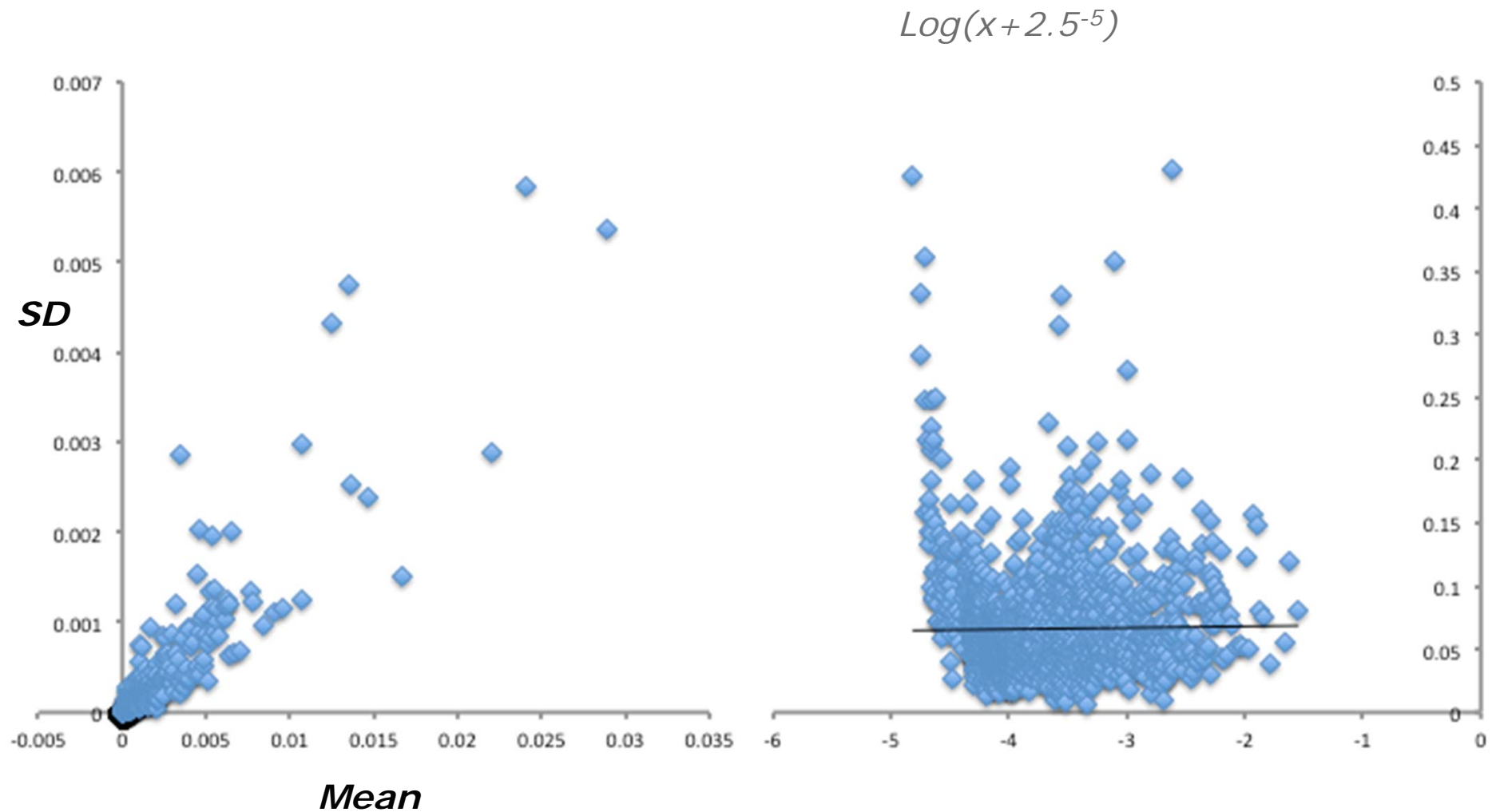
Variance-stabilizing transformations



Variance-stabilizing transformations



Variance-stabilizing transformations

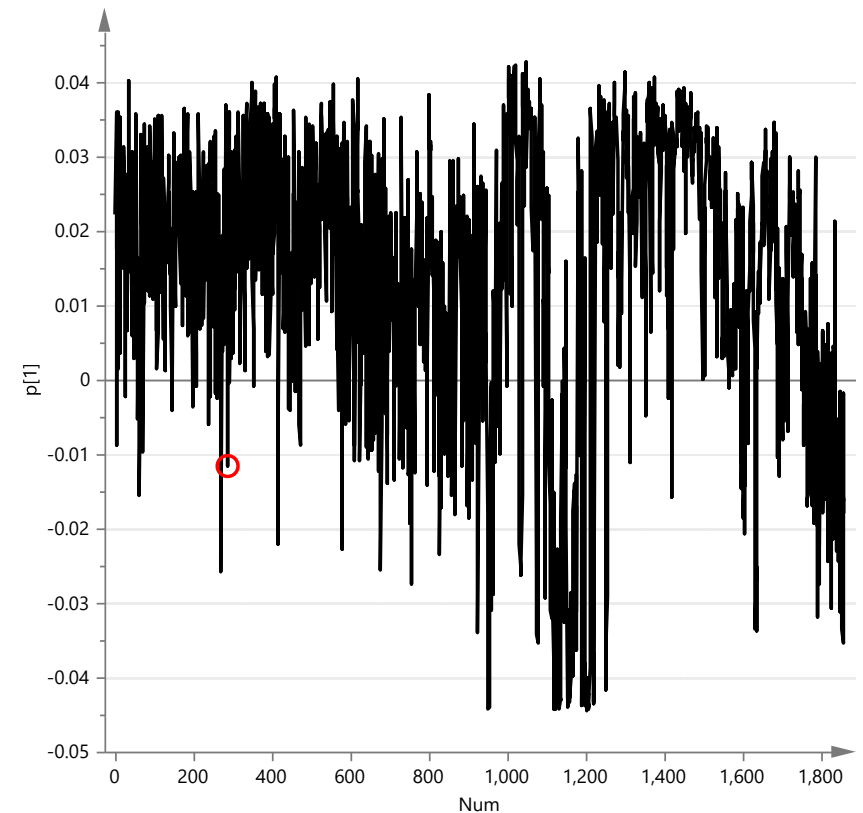
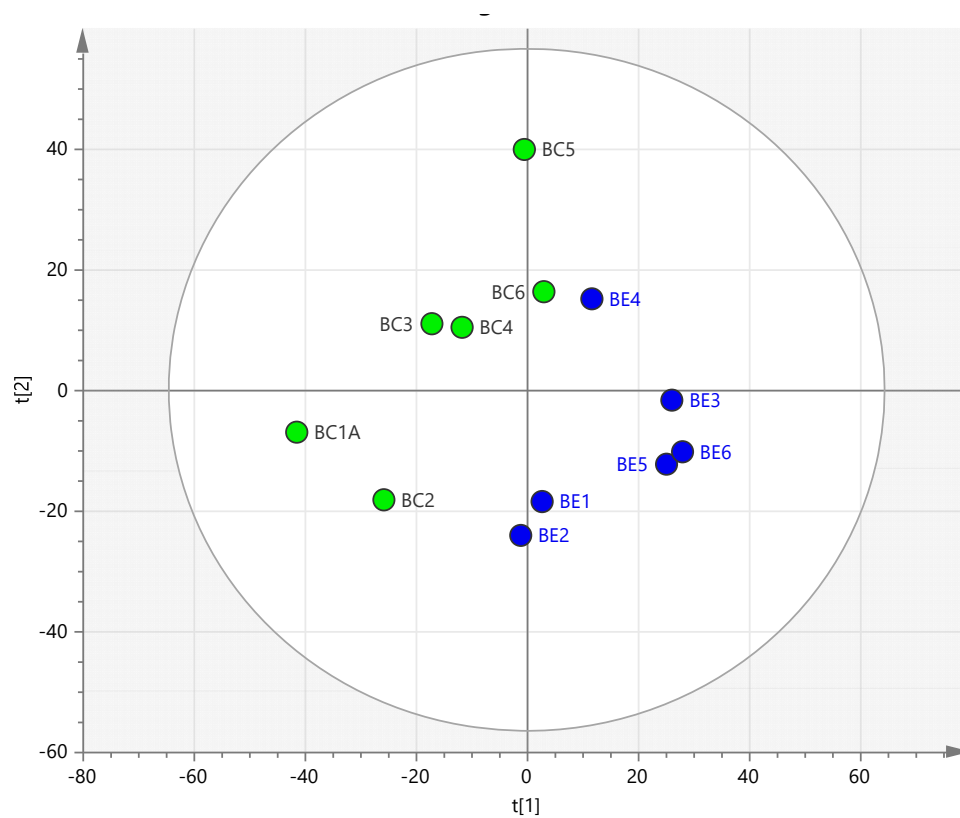


What is the practical value?

- Without some kind of transformation, PCA/PLS results dominated by high-concentration metabolites
- UV scaling can strongly overweight low-intensity peaks
 - Particular problem when you have noise regions in the data – somewhat different when you have peak detection
- Log transformation is better than Pareto scaling
 - In my subjective opinion!
 - Greater effect on *which* peaks are identified as influential

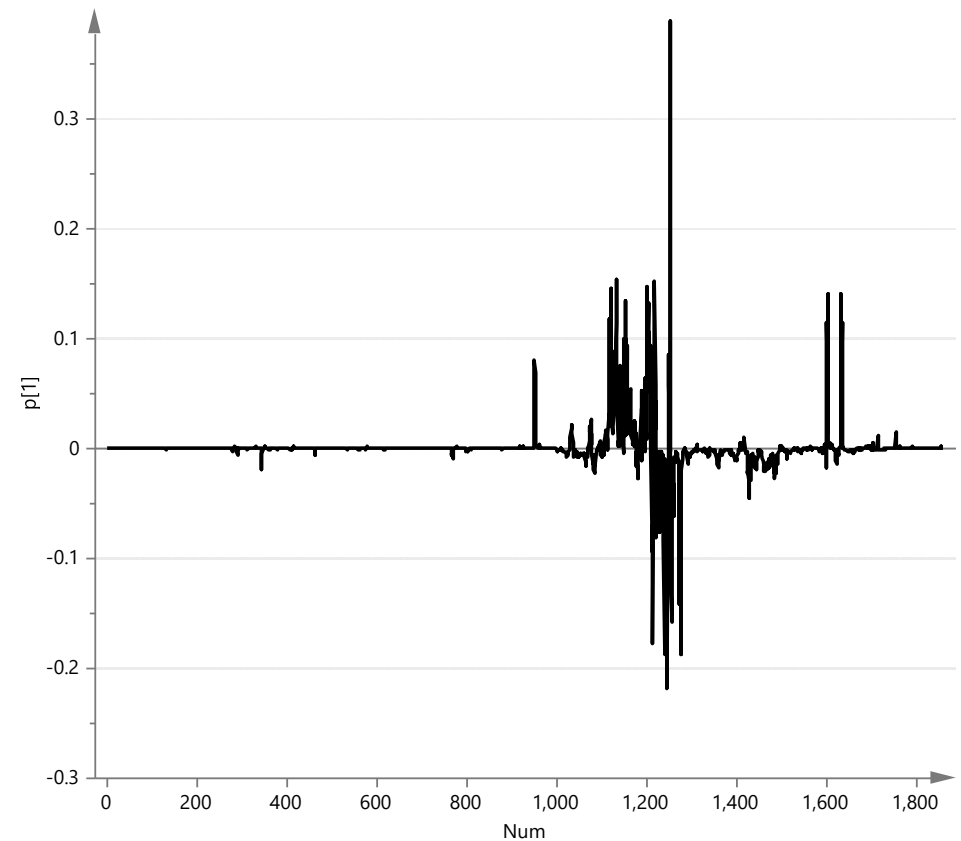
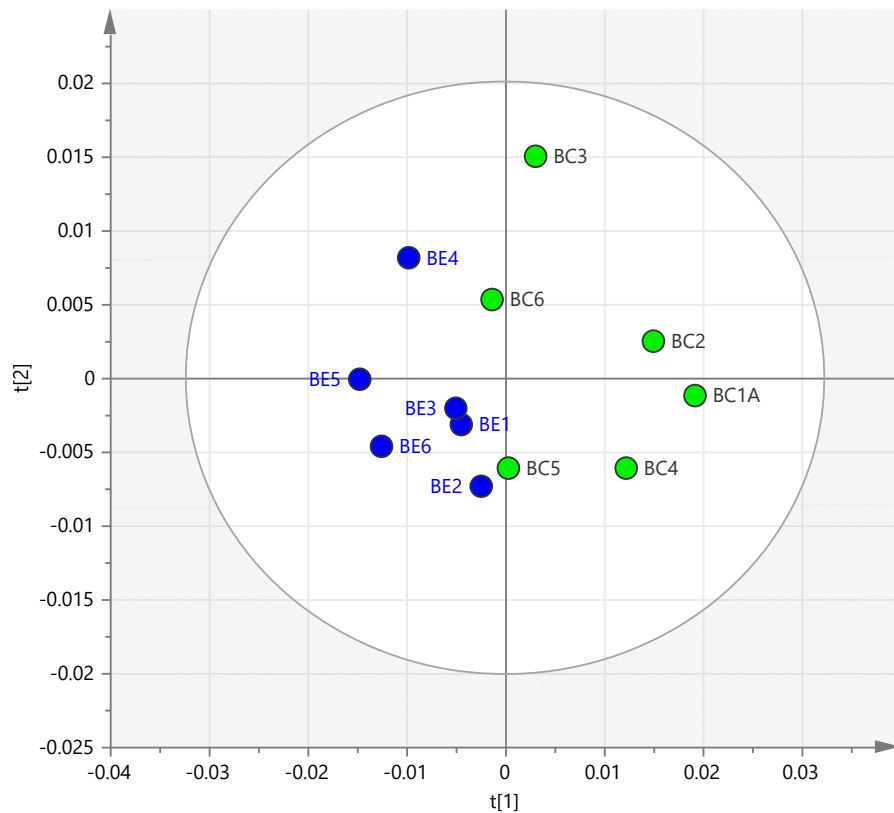
What is the practical value?

- Example set of real samples – NMR spectra, two groups
- Autoscaled



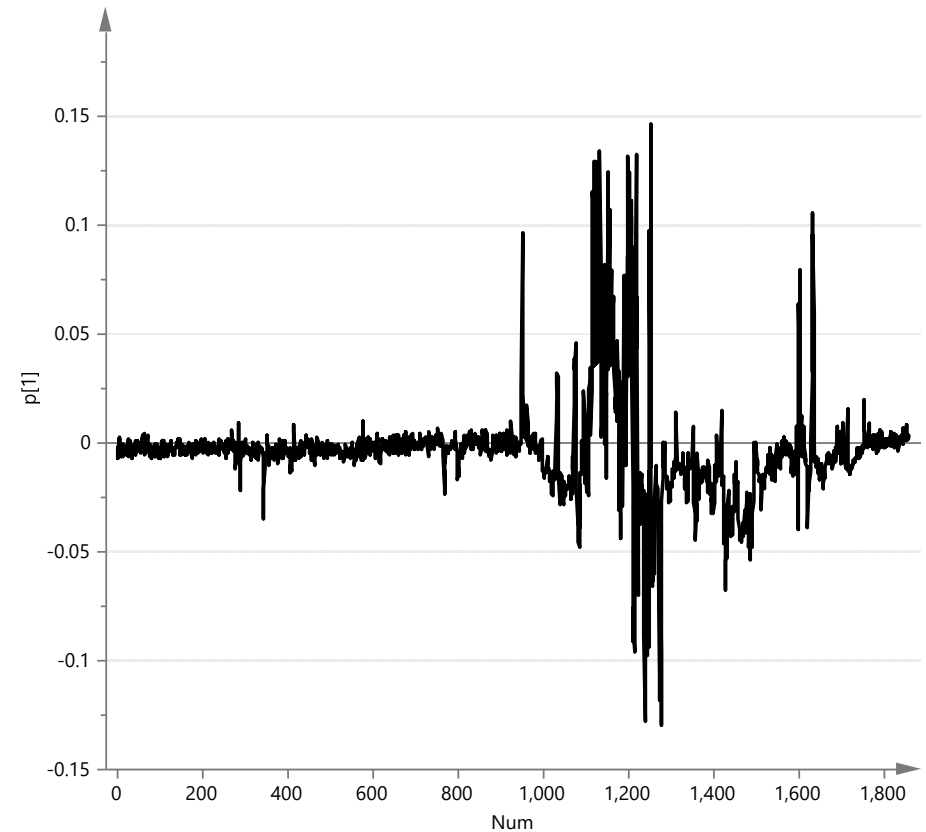
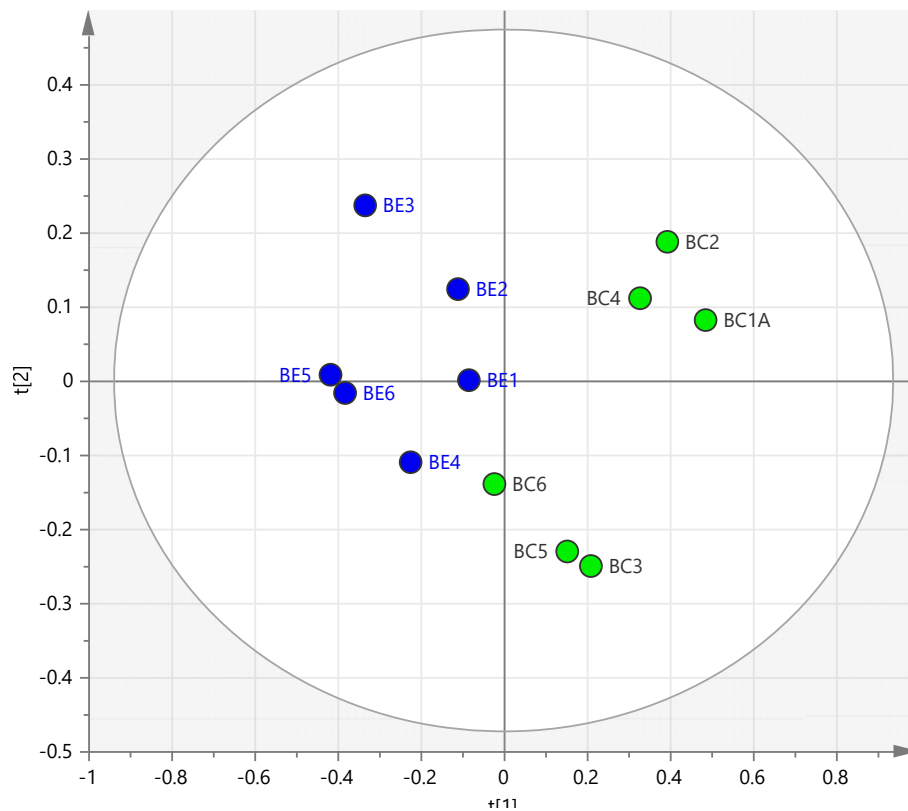
What is the practical value?

- Example set of real samples – NMR spectra, two groups
- Mean-centred



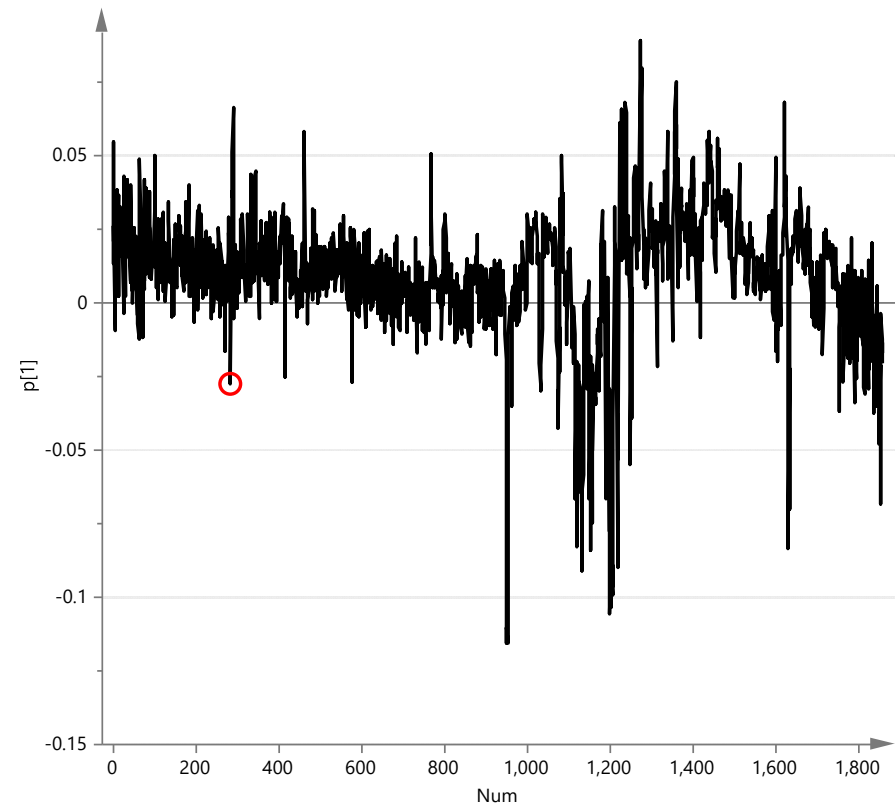
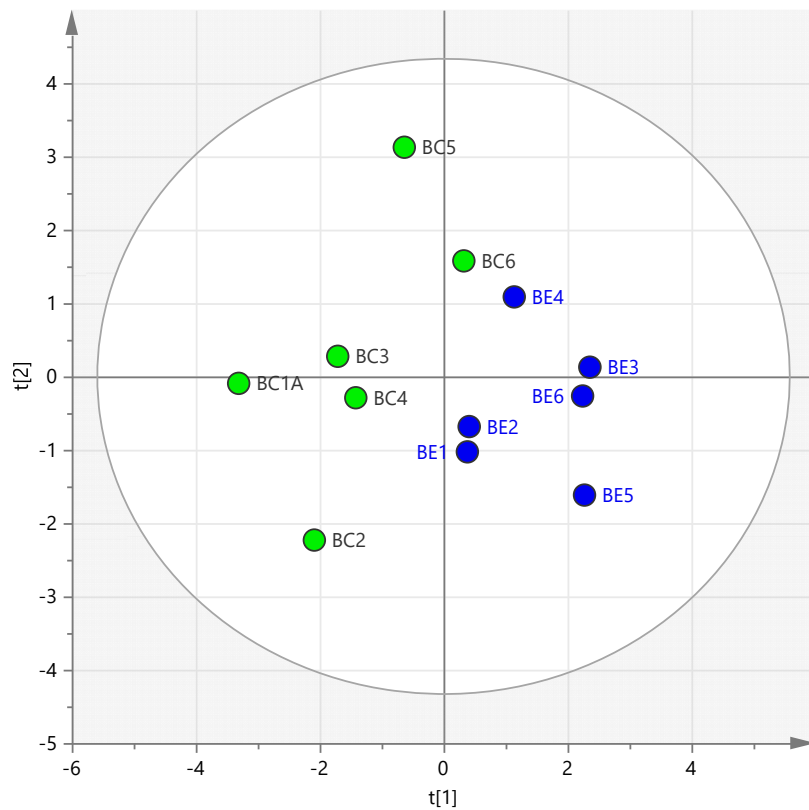
What is the practical value?

- Example set of real samples – NMR spectra, two groups
- Pareto-scaled



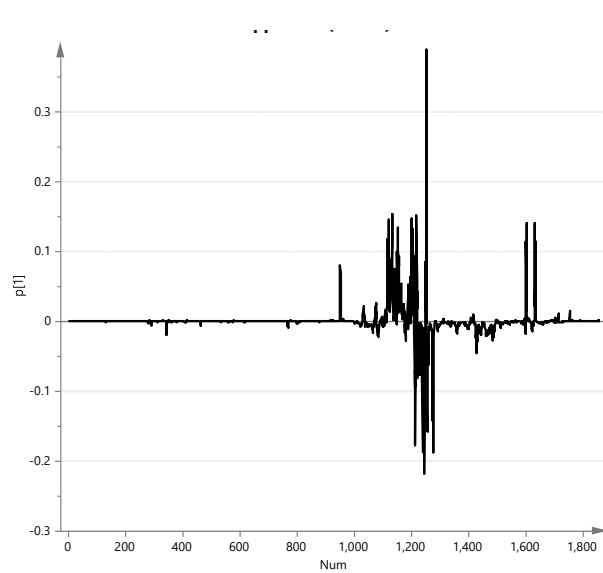
What is the practical value?

- Example set of real samples – NMR spectra, two groups
- $\text{Log}(x+C)$ transformed

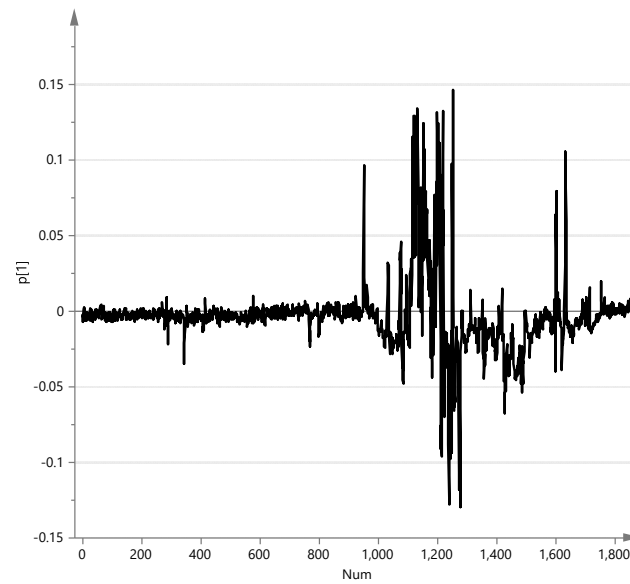


What is the practical value?

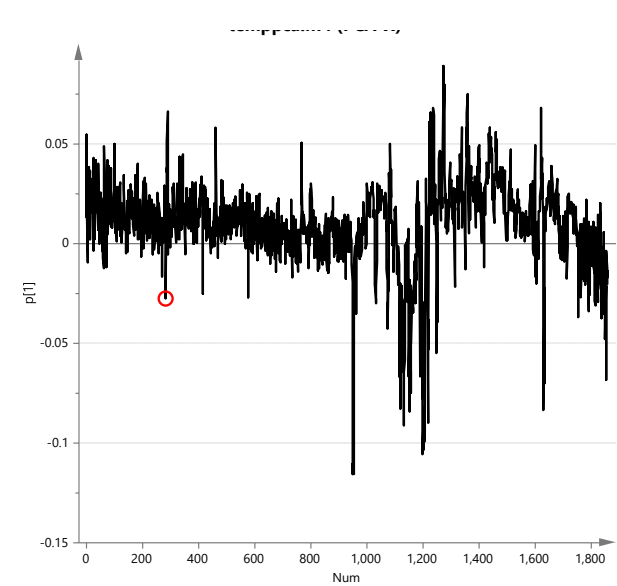
Mean-centered



Pareto-scaled



Log-transformed



One last suggestion for transformation -

Scale to unit variance in the control set only. Very simple thing to do!

Metabolomic profiling of heat stress: hardening and recovery of homeostasis
in *Drosophila*

Anders Malmendal,¹ Johannes Overgaard,^{1,2,3} Jacob G. Bundy,⁴ Jesper G. Sørensen,³
Niels Chr. Nielsen,¹ Volker Loeschcke,³ and Martin Holmstrup^{2,3}

Am J Physiol Regul Integr Comp Physiol 291: R205–R212, 2006.

One last suggestion for transformation -

Scale to unit variance in the control set only. Very simple thing to do!

Metabolomic profiling of heat stress: hardening and recovery of homeostasis in *Drosophila*

Anders Malmendal,¹ Johannes Overgaard,^{1,2,3} Jacob G. Bundy,⁴ Jesper G. Sørensen,³
Niels Chr. Nielsen,¹ Volker Loeschcke,³ and Martin Holmstrup^{2,3}

Am J Physiol Regul Integr Comp Physiol 291: R205–R212, 2006.

Cell, Vol. 102, 109–126, July 7, 2000, Copyright ©2000 by Cell Press

Functional Discovery via a Compendium of Expression Profiles

Timothy R. Hughes,*# Matthew J. Marton,*#
Allan R. Jones,* Christopher J. Roberts,*
Roland Stoughton,* Christopher D. Armour,*
Holly A. Bennett,* Ernest Coffey,* Hongyue Dai,*
Yudong D. He,* Matthew J. Kidd,* Amy M. King,*
Michael R. Meyer,* David Slade,* Pek Y. Lum,*
Sergey B. Stepaniants,* Daniel D. Shoemaker,*
Daniel Gachotte,† Kalpana Chakraburtt,‡
Julian Simon,§ Martin Bard,†
and Stephen H. Friend*||

One last suggestion for transformation -

Scale to unit variance in the control set only. Very simple thing to do!

Metabolomic profiling of heat stress: hardening and recovery of homeostasis in *Drosophila*

Anders Malmendal,¹ Johannes Overgaard,^{1,2,3} Jacob G. Bundy,⁴ Jesper G. Sørensen,³
Niels Chr. Nielsen,¹ Volker Loeschcke,³ and Martin Holmstrup^{2,3}

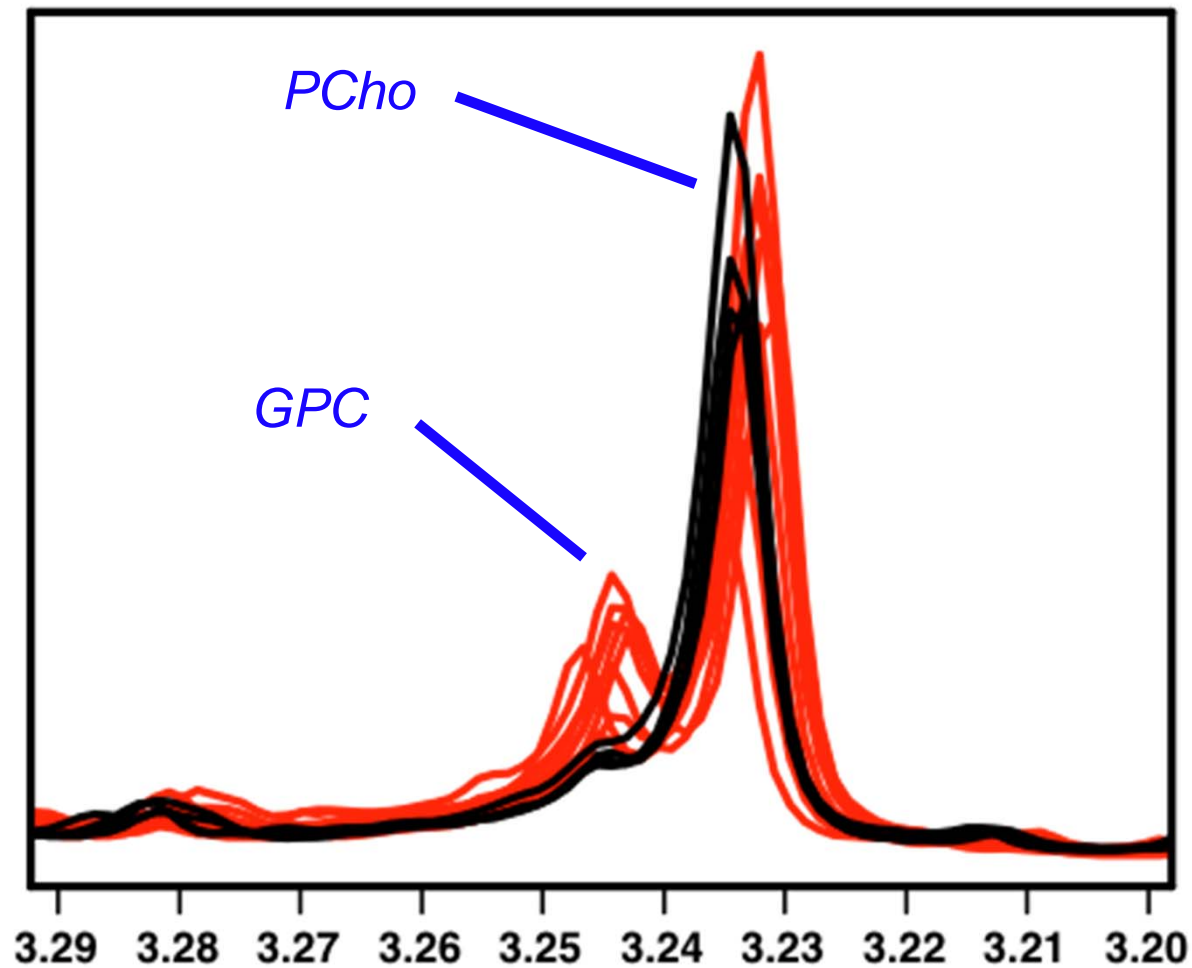
Am J Physiol Regul Integr Comp Physiol 291: R205–R212, 2006.

Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling

Hector C. Keun*, Timothy M.D. Ebbels, Henrik Antti, Mary E. Bollard,
Olaf Beckonert, Elaine Holmes, John C. Lindon, Jeremy K. Nicholson

Analytica Chimica Acta 490 (2003) 265–276

Don't forget the simple things



Univariate or multivariate analysis for finding markers?

Some cases where a multivariate approach is absolutely required

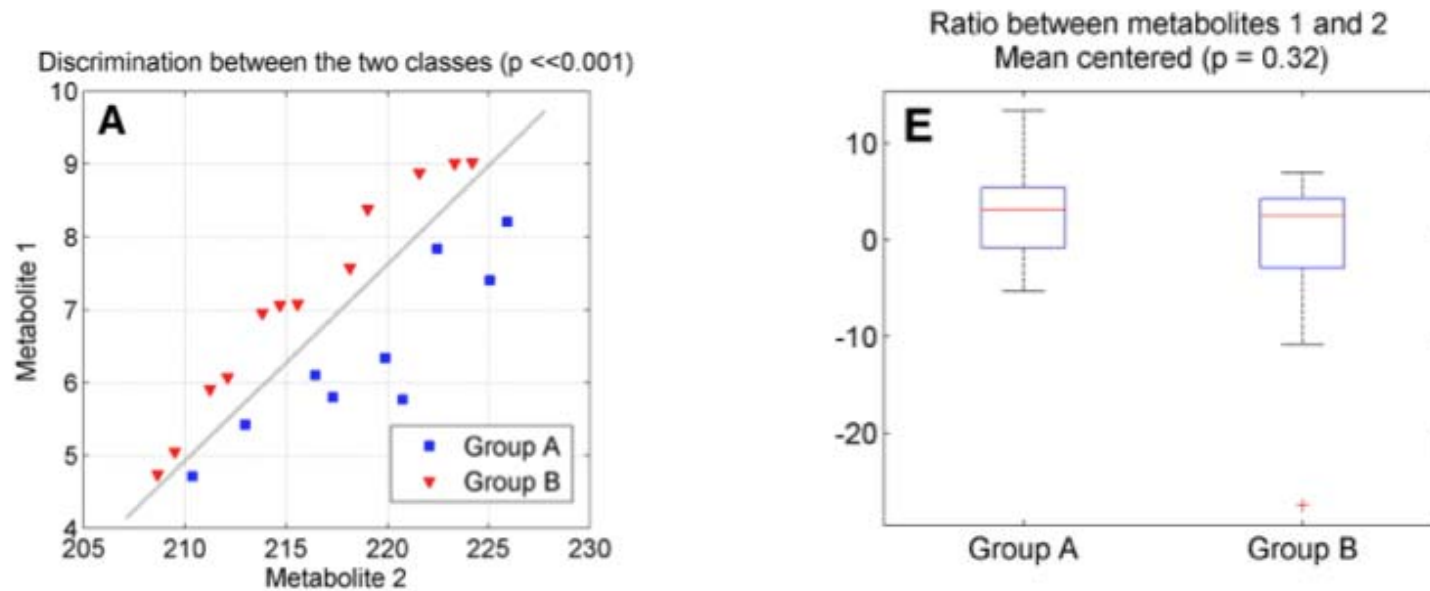
Reflections on univariate and multivariate analysis of metabolomics data

**Edoardo Saccenti · Huub C. J. Hoefsloot ·
Age K. Smilde · Johan A. Westerhuis ·
Margriet M. W. B. Hendriks**

**Metabolomics (2014) 10:361–374
DOI 10.1007/s11306-013-0598-6**

Univariate or multivariate analysis for finding markers?

Some cases where a multivariate approach is absolutely required

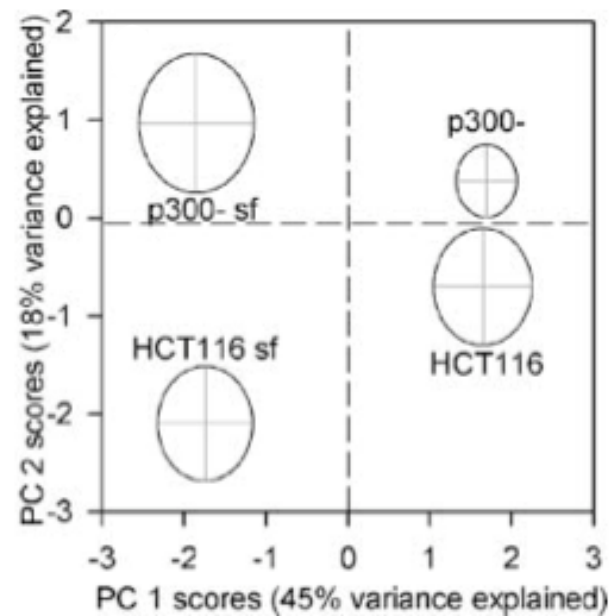


Is this ever found in a real-world situation, though?

Univariate or multivariate analysis for finding markers?

Metabolic Consequences of p300 Gene Deletion in Human Colon Cancer Cells

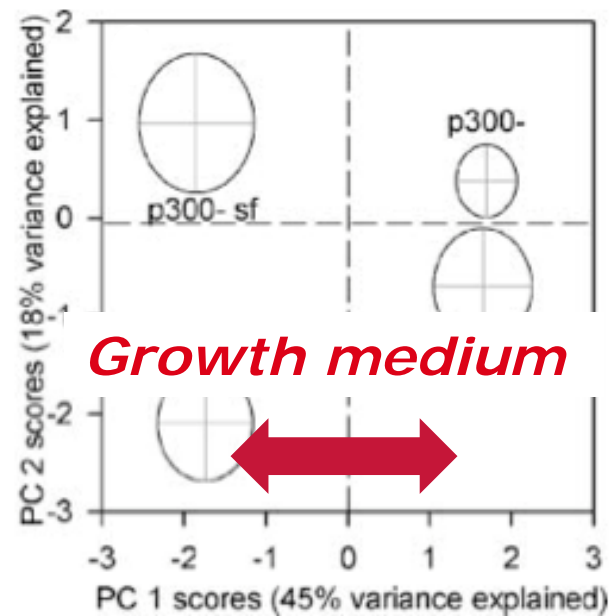
Jacob G. Bundy,¹ N. Gopalakrishna Iyer,² Michelle S. Gentile,² De-En Hu,¹
Mikko Kettunen,¹ Ana-Teresa Maia,² Natalie P. Thorne,² James D. Brenton,²
Carlos Caldas,² and Kevin M. Brindle¹



Univariate or multivariate analysis for finding markers?

Metabolic Consequences of p300 Gene Deletion in Human Colon Cancer Cells

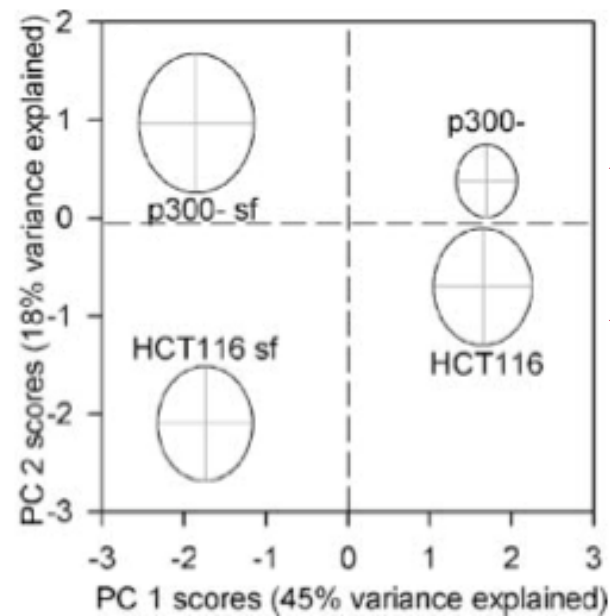
Jacob G. Bundy,¹ N. Gopalakrishna Iyer,² Michelle S. Gentile,² De-En Hu,¹
Mikko Kettunen,¹ Ana-Teresa Maia,² Natalie P. Thorne,² James D. Brenton,²
Carlos Caldas,² and Kevin M. Brindle¹



Univariate or multivariate analysis for finding markers?

Metabolic Consequences of p300 Gene Deletion in Human Colon Cancer Cells

Jacob G. Bundy,¹ N. Gopalakrishna Iyer,² Michelle S. Gentile,² De-En Hu,¹
Mikko Kettunen,¹ Ana-Teresa Maia,² Natalie P. Thorne,² James D. Brenton,²
Carlos Caldas,² and Kevin M. Brindle¹



Genotype

Why should we use multivariate methods?

Multivariate analyses are not the endpoint.

Louis Thurstone

- Psychologist and psychometrician (1887-1955)
 - Developed multivariate methods (factor analysis) in the context of understanding psychological properties
- “The exploratory nature of factor analysis is often not understood ... Factor analysis is useful, especially in those domains where basic and fruitful concepts are essentially lacking and where crucial experiments have been difficult to conceive. The new methods have a humble role. They enable us to make only the crudest first map of a new domain.”**



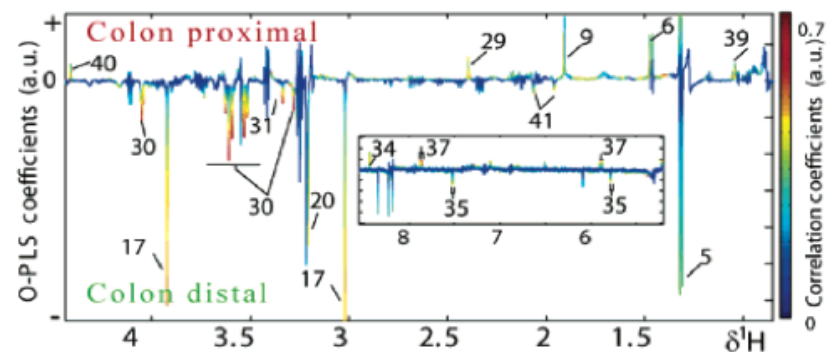
The Vectors of Mind (1947)

Differential metabograms

Only 5 references in the literature!

Effects of Probiotic *Lactobacillus Paracasei* Treatment on the Host Gut Tissue Metabolic Profiles Probed via Magic-Angle-Spinning NMR Spectroscopy

Francois-Pierre J. Martin,[†] Yulan Wang,[†] Norbert Sprenger,[‡] Elaine Holmes,[†] John C. Lindon,[†] Sunil Kochhar,[§] and Jeremy K. Nicholson^{*,†}



Differential metabograms – as O-PLS output

Trends

Trends in Analytical Chemistry, Vol. 28, No. 11, 2009

Notes on the practical utility of OPLS

Henri S. Tapp, E. Kate Kemsley

Differential metabograms – as O-PLS output

Trends

Trends in Analytical Chemistry, Vol. 28, No. 11, 2009

Notes on the practical utility of OPLS

Henri S. Tapp, E. Kate Kemsley

- *Demonstrate that they are directly equivalent to covariance/correlation plots*
- *Concluded that they are misleading in terms of representing data*
- *But this underestimates the value of this type of plot!*
 - *Essentially similar to a volcano plot*
 - *BUT adds in information on variable order*

Directly comparing multivariate *v.* univariate analyses

Example 1. Responses to a toxin: looking for expected metabolites

Example 2. Comparing different genotypes (cryptic species) from wild populations across several sites.

Directly comparing multivariate *v.* univariate analyses

Example 1. Responses to a toxin.

- Nematode *C. elegans* exposed to cadmium – interested in phytochelatin responses

Directly comparing multivariate *v.* univariate analyses

Example 1. Responses to a toxin.

- Nematode *C. elegans* exposed to cadmium – interested in phytochelatin responses
- Phytochelatins (PCs) are oligomers of glutathione – $(\text{GluCys})_n\text{Gly}$

Directly comparing multivariate *v.* univariate analyses

Example 1. Responses to a toxin.

- Nematode *C. elegans* exposed to cadmium – interested in phytochelatin responses
- Phytochelatins (PCs) are oligomers of glutathione – (GluCys)_nGly
- We wanted to know – are PCs responsive to cadmium in an animal species? (Previously only inferred from genetic evidence.)

The Metabolomic Responses of *Caenorhabditis elegans* to Cadmium Are Largely Independent of Metallothionein Status, but Dominated by Changes in Cystathionine and Phytochelatins

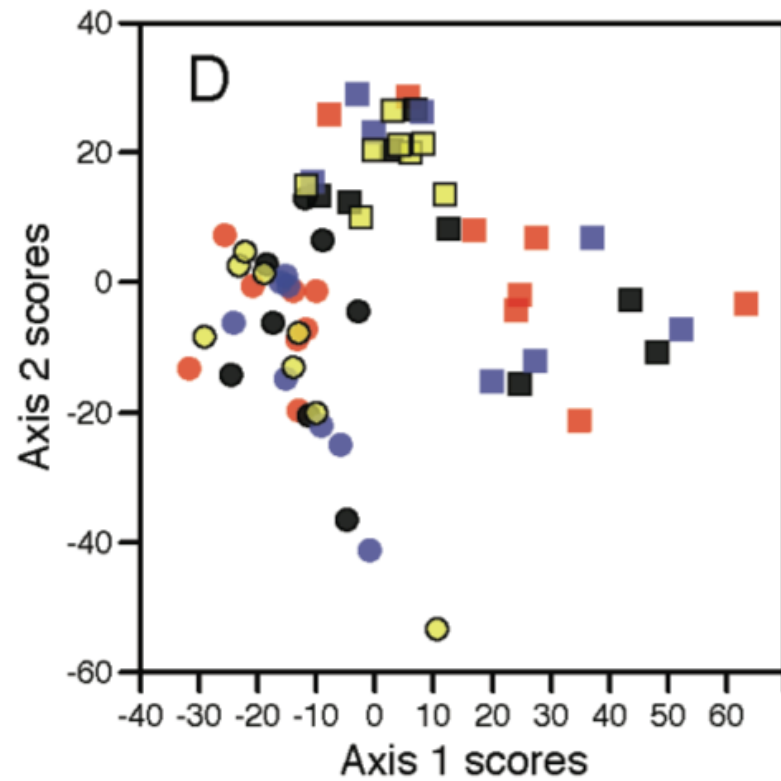
Samantha L. Hughes,^{†,‡,⊥,#} Jacob G. Bundy,^{§,#} Elizabeth J. Want,[§] Peter Kille,[‡] and Stephen R. Stürzenbaum^{*,†}

Multivariate/univariate comparison: phytochelatins

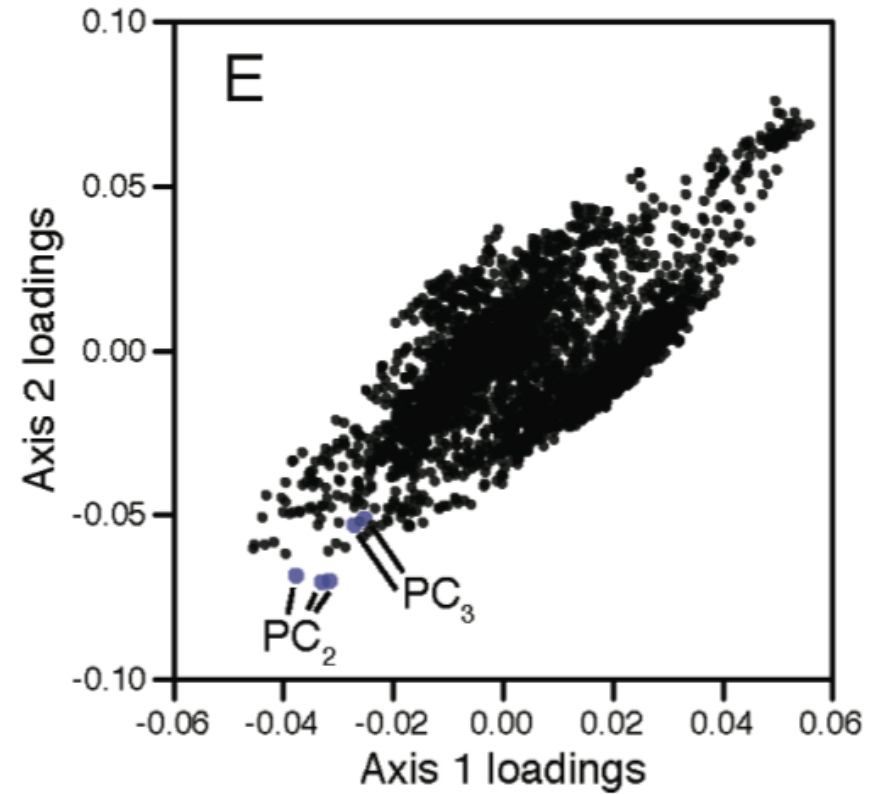
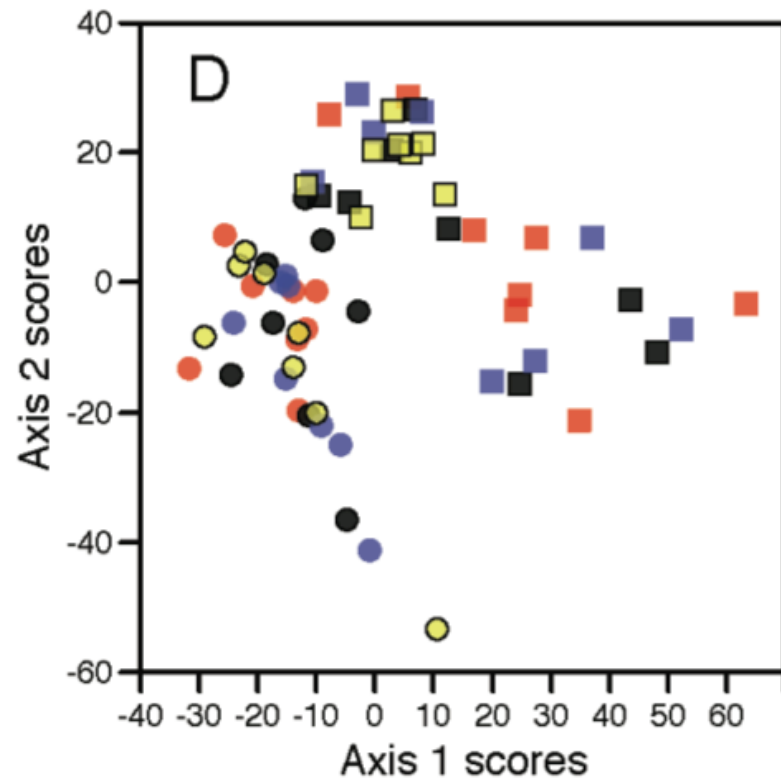
PC₂: GluCysGluCysGly

predicted MW 540.143

Multivariate/univariate comparison: phytochelatins



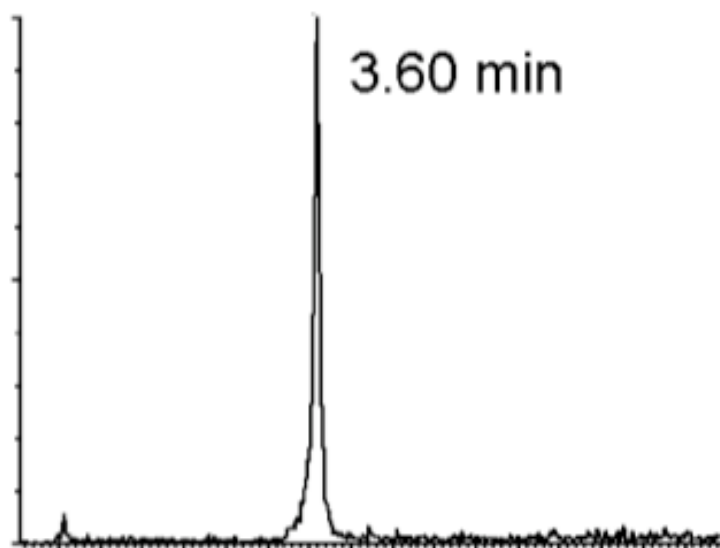
Multivariate/univariate comparison: phytochelatins



Multivariate/univariate comparison: phytochelatins

PC₂: GluCysGluCysGly

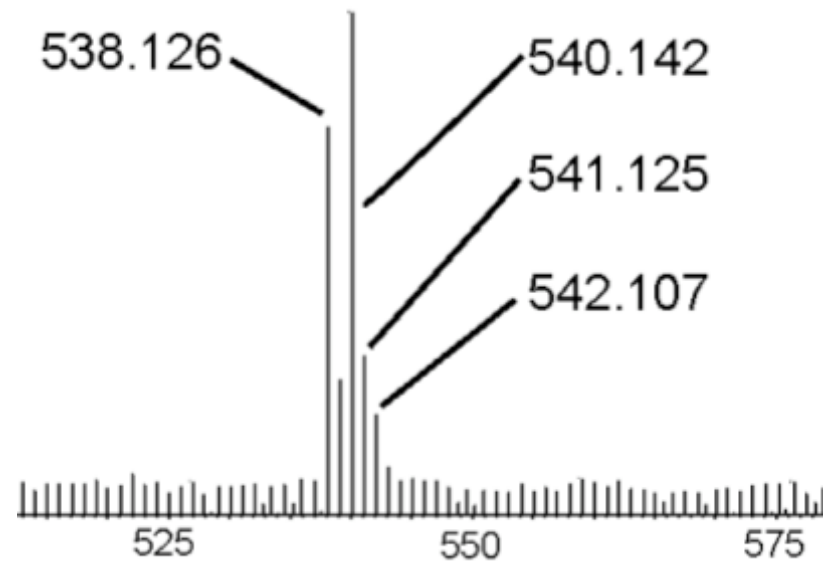
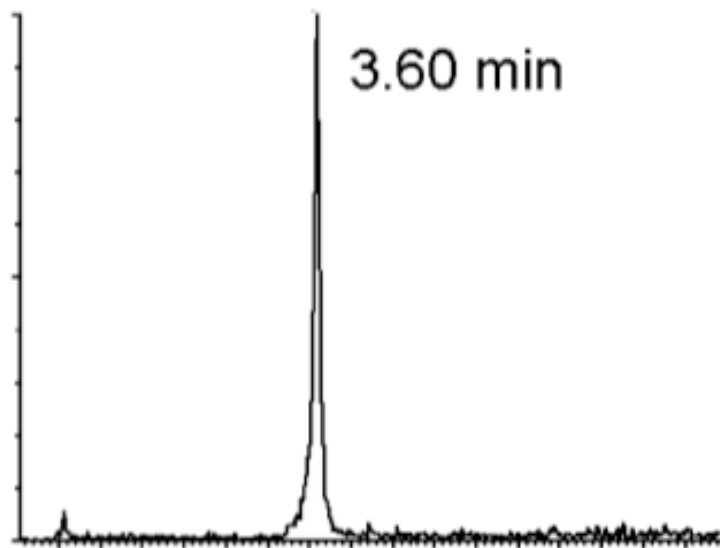
predicted MW 540.143



Multivariate/univariate comparison: phytochelatins

PC₂: GluCysGluCysGly

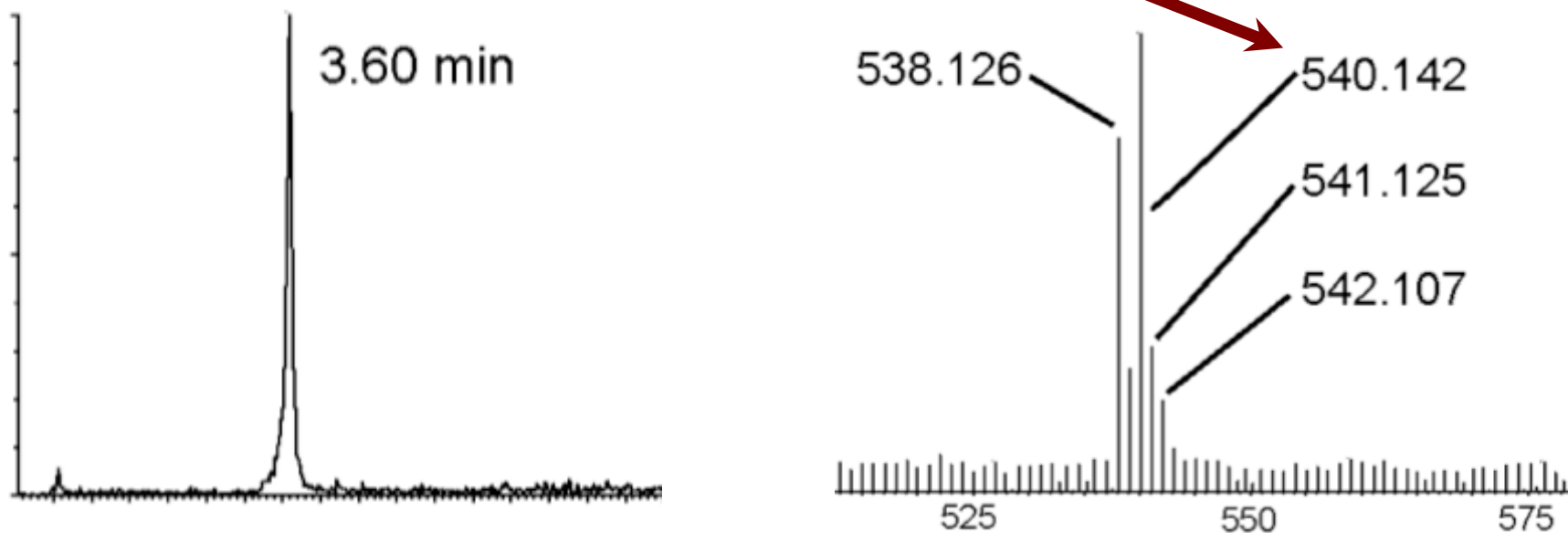
predicted MW 540.143



Multivariate/univariate comparison: phytochelatins

PC₂: GluCysGluCysGly

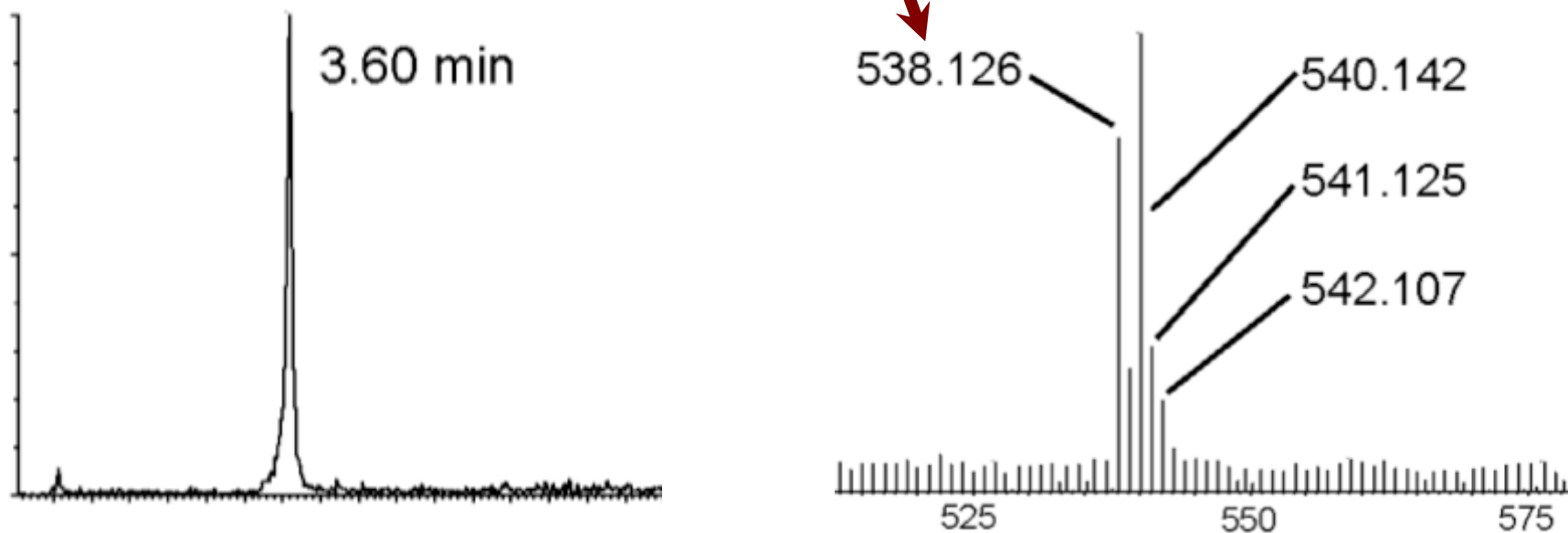
predicted MW 540.143



Multivariate/univariate comparison: phytochelatins

PC_2 : GluCysGluCysGly

predicted MW 540.143 – 2H



Multivariate/univariate comparison: phytochelatins

Univariate results?

4821 peaks detected after
processing (XCMS)

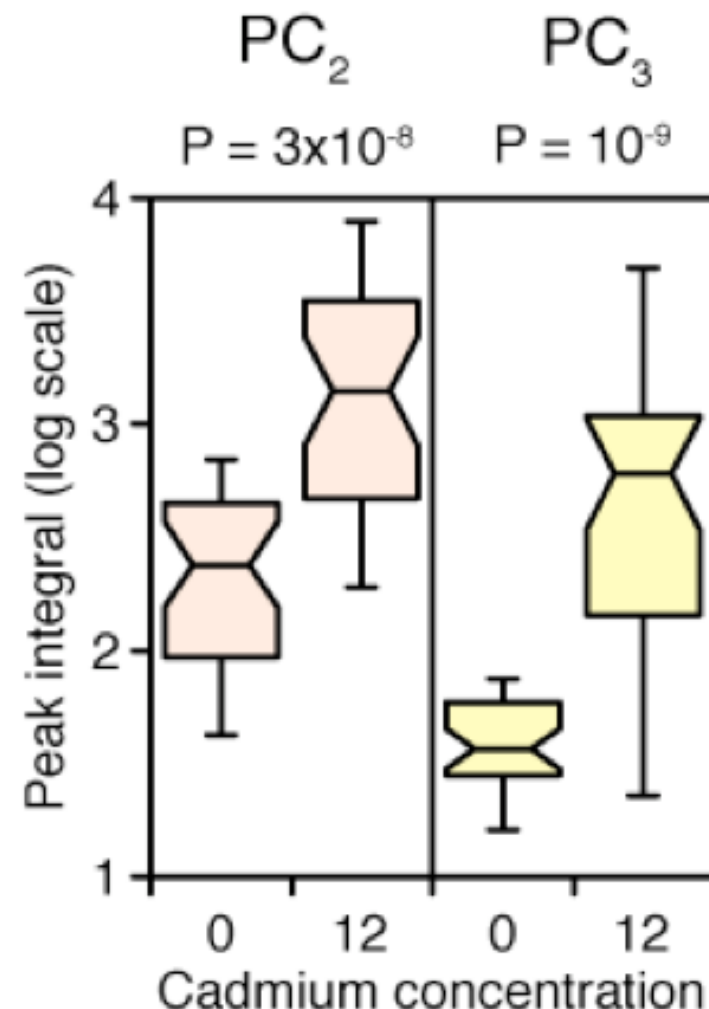
- PC₃ putative: 535/4821
- PC₃ putative: 966/4821
- PC₂ putative: 2143/4821
- PC₂ putative: 3630/4821

Multivariate/univariate comparison: phytochelatins

Univariate results?

4821 peaks detected after processing (XCMS)

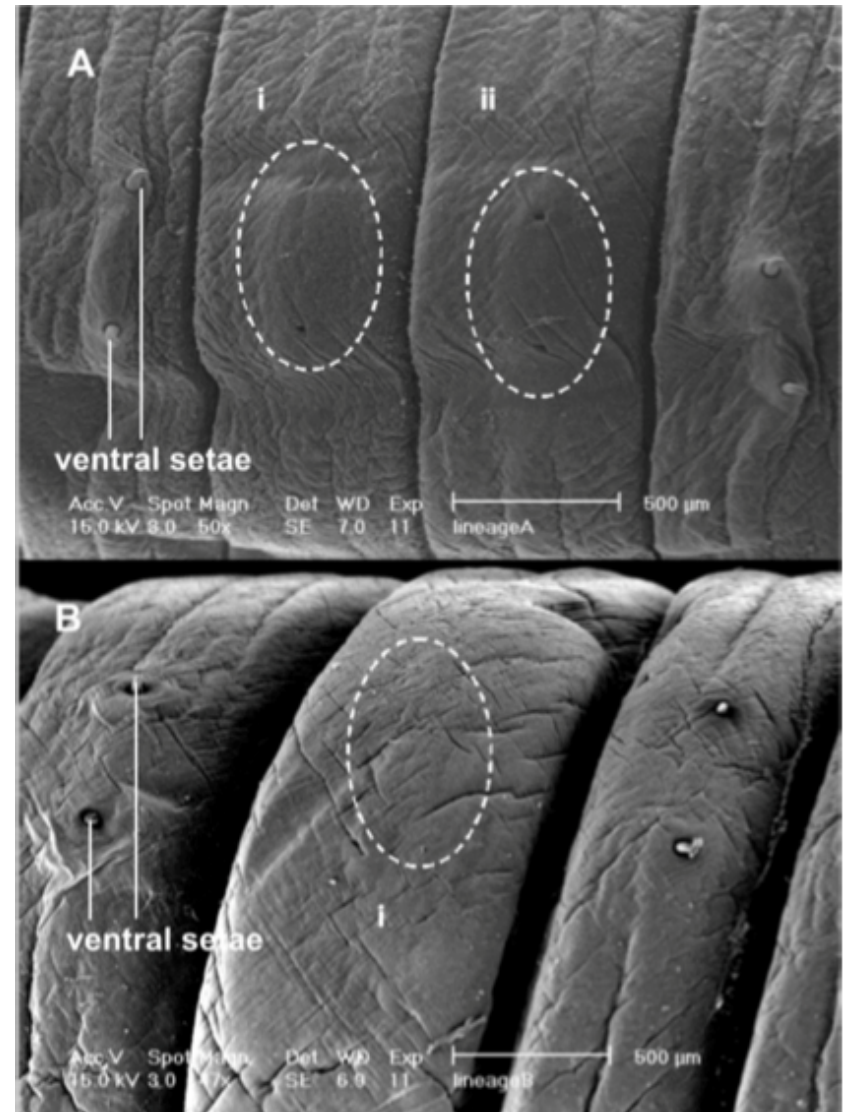
- PC₃ putative: 535/4821
- PC₃ putative: 966/4821
- PC₂ putative: 2143/4821
- PC₂ putative: 3630/4821



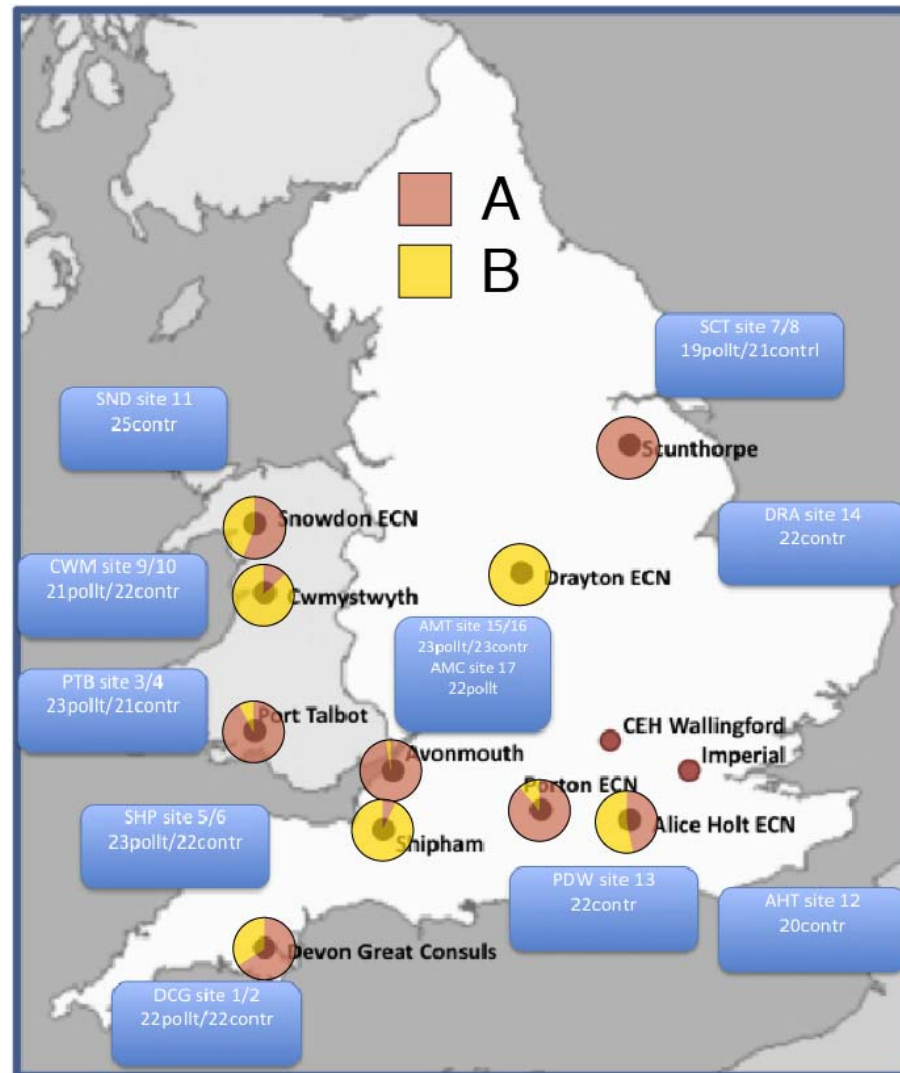
Multivariate/univariate comparison: cryptic species

Example 2. Distinguishing cryptic species.

- Earthworm species are hard to tell apart!
- *Lumbricus rubellus* is present as two putative cryptic species in the UK
- We sampled wild populations from multiple sites within the UK, and analysed tissue extracts by NMR – complex set of site effects!



Multivariate/univariate comparison: cryptic species



Earthworm hunting



Earthworm hunting



Devon Great Consols – arsenic contaminated site



Cwmystwyth cottage



Multivariate/univariate comparison: cryptic species

Univariate analysis: needs to allow for site effects

- Filter data for sites with ≥ 4 individuals from both genotypes
 - Dataset reduced from 17 sites and ~200 worms to 5 sites and <100 worms
 - Simple linear models identified 4 possible markers:

δ 2.22, δ 2.89, δ 3.10, δ 7.70

- Significant even when allowing for other factors (soil pH, soil organic carbon, soil moisture)
- How did a multivariate approach compare?

Multivariate/univariate comparison: cryptic species

Univariate analysis: needs to allow for site effects

- Filter data for sites with ≥ 4 individuals from both genotypes
 - Dataset reduced from 17 sites and ~200 worms to 5 sites and <100 worms
 - Simple linear models identified 4 possible markers:

δ 2.22, δ 2.89, δ 3.10, δ 7.70

- Significant even when allowing for other factors (soil pH, soil organic carbon, soil moisture)
- How did a multivariate approach compare?

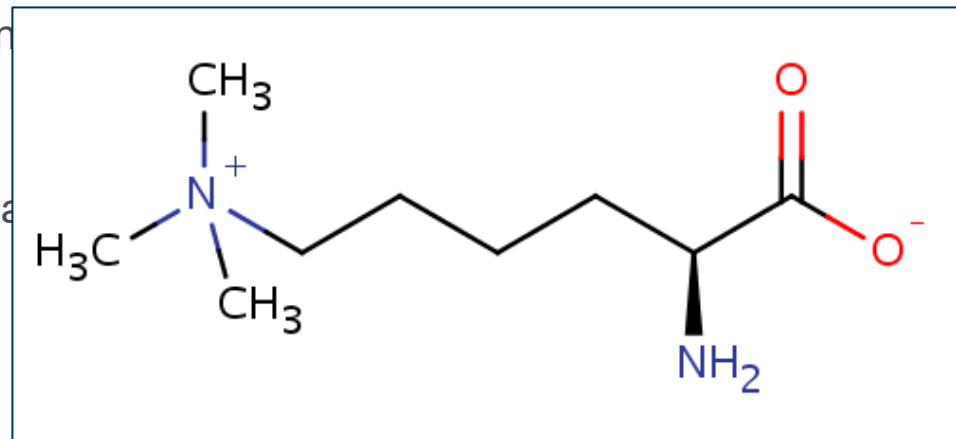
Multivariate/univariate comparison: cryptic species

Univariate analysis: needs to allow for site effects

- Filter data for sites with ≥ 4 individuals from both genotypes
 - Dataset reduced from 17 sites and ~200 worms to 5 sites and <100 worms
 - Simple linear models identified 4 possible markers:

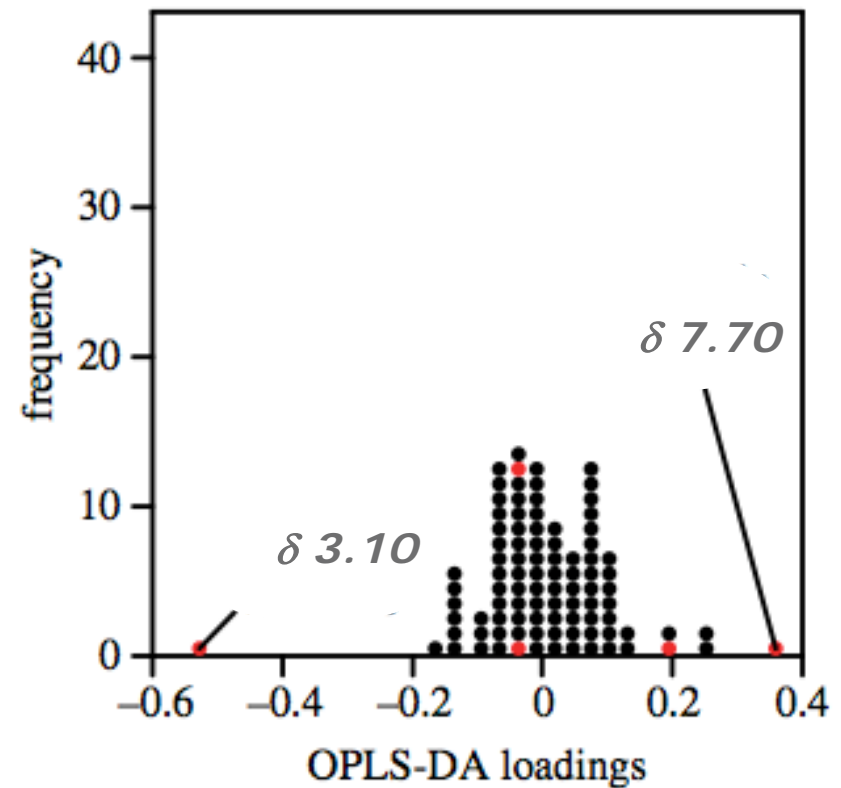
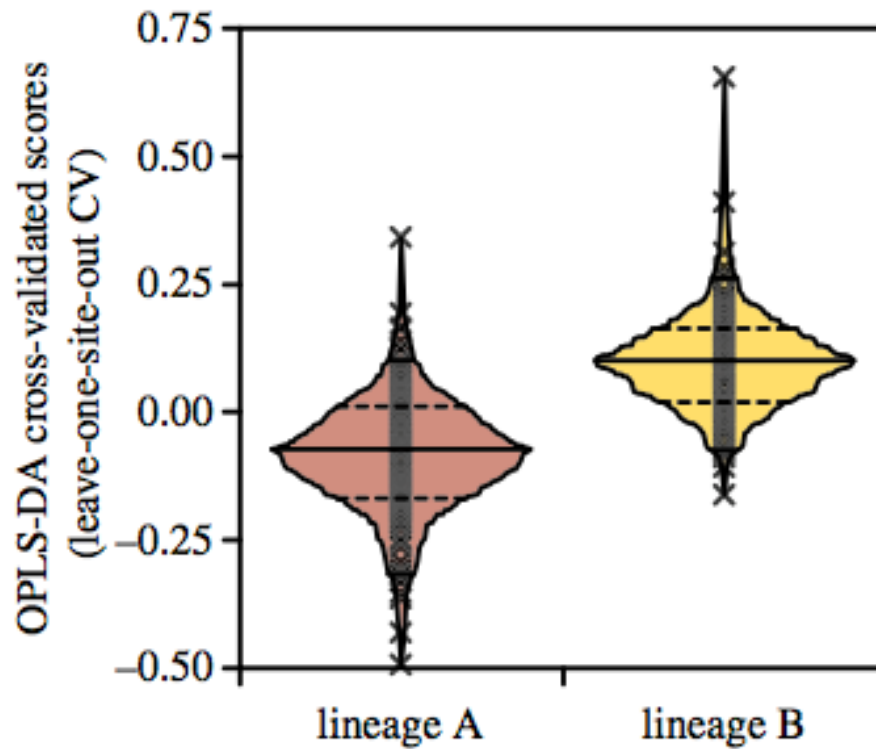
δ 2.22, δ 2.89, δ 3.10, δ 7.70

- Significant even when allowing for site effects (e.g. carbon, soil moisture)
- How did a multivariate approach



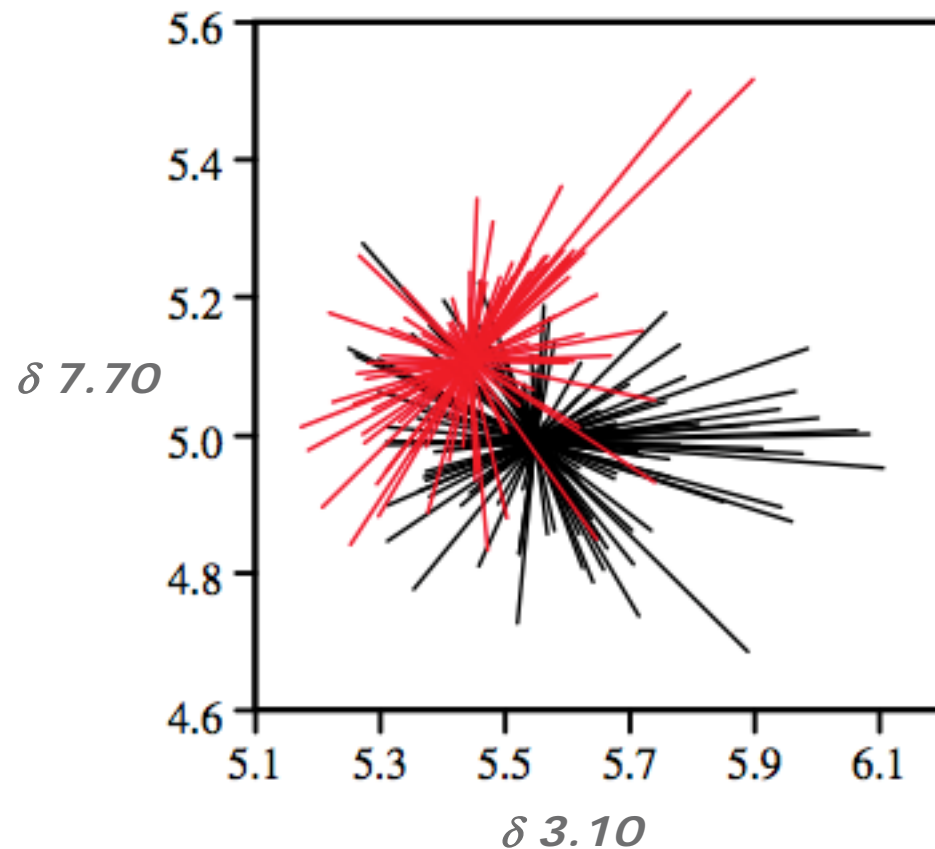
Multivariate/univariate comparison: cryptic species

OPLS-DA (one correlated and two orthogonal axes)



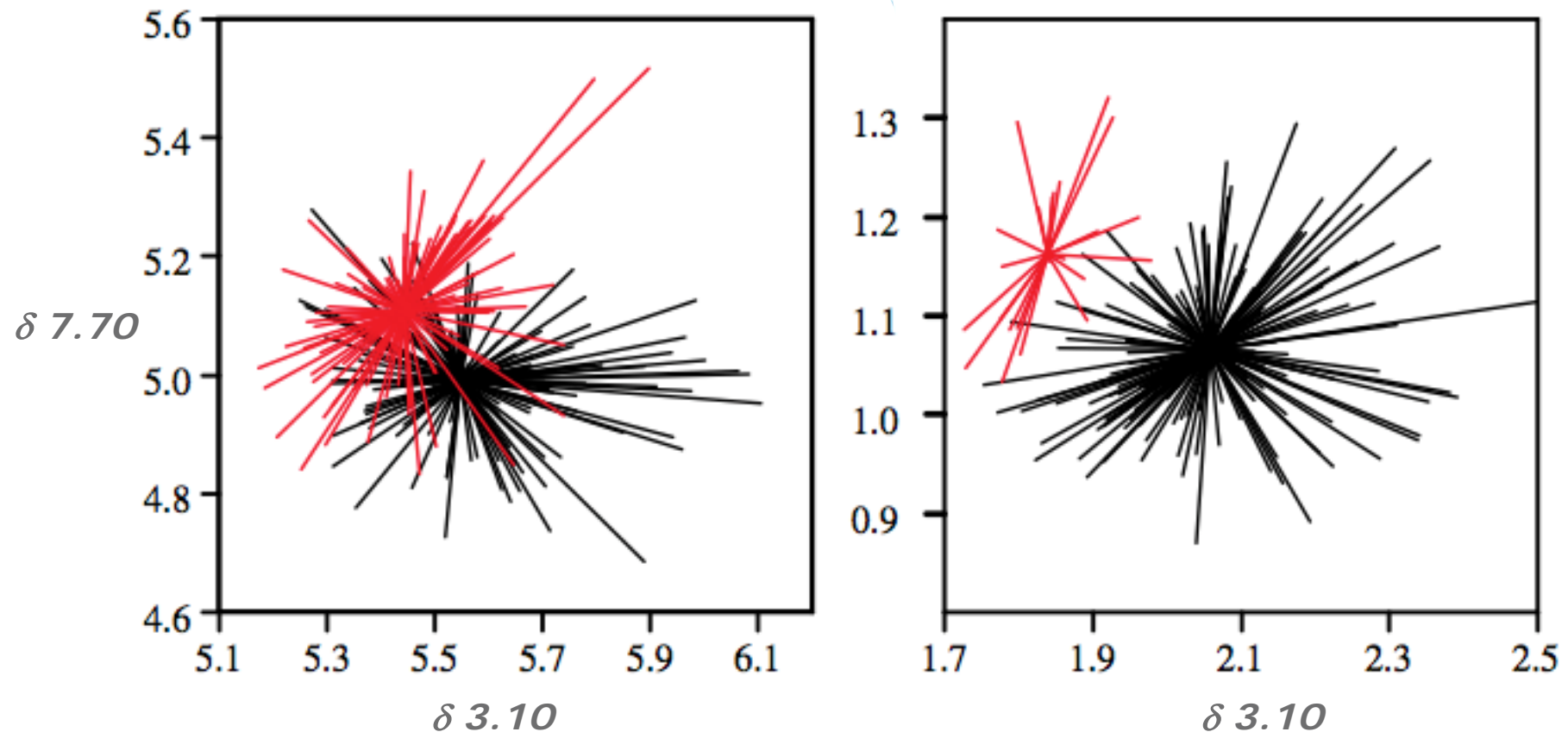
Multivariate/univariate comparison: cryptic species

Comparing two variables, $\delta 3.10$ and $\delta 7.70$



Multivariate/univariate comparison: cryptic species

Comparing two variables, $\delta 3.10$ and $\delta 7.70$



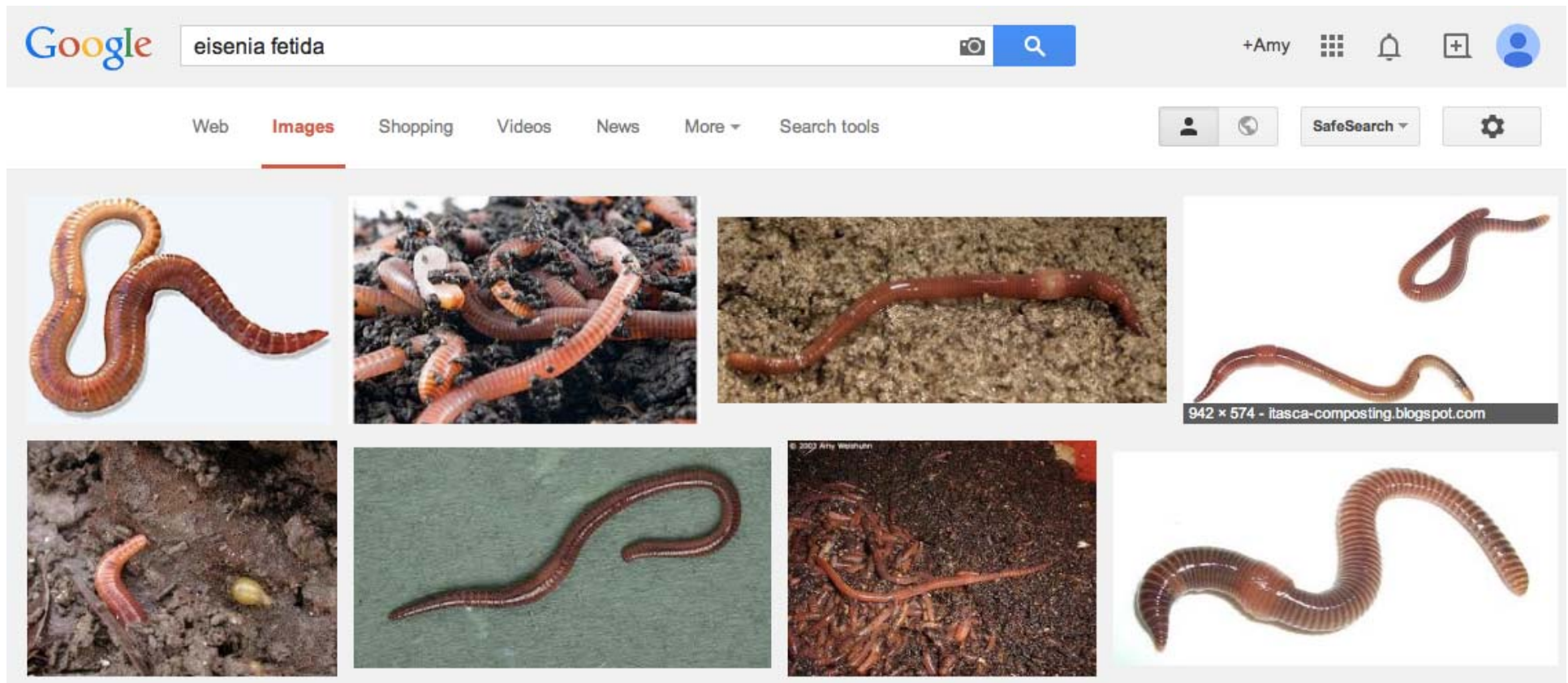
Are we overfitting our data?

Data analysis is easy when the results are clear-cut!

Are we overfitting our data?

Data analysis is easy when the results are clear-cut!

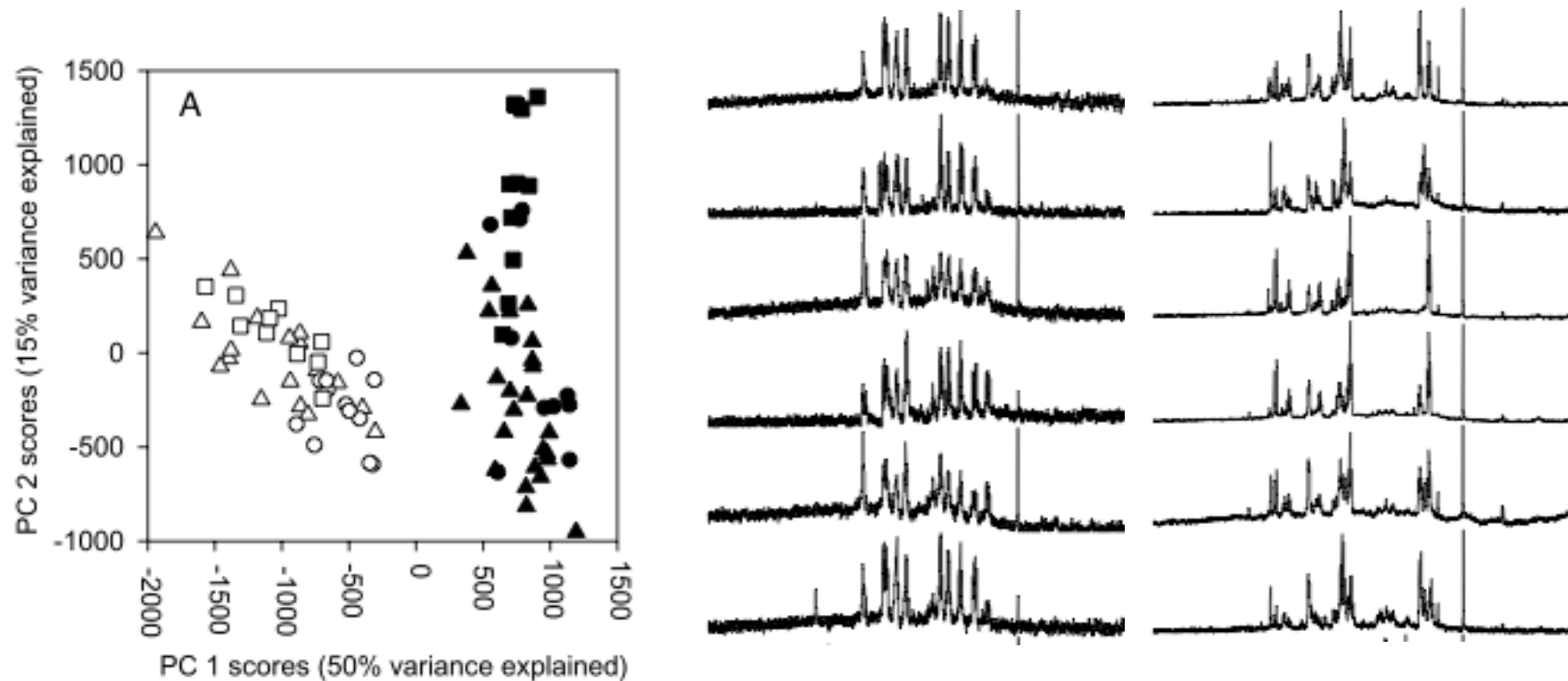
Eisenia fetida v *Eisenia andrei* – sister species



Are we overfitting our data?

Data analysis is easy when the results are clear-cut!

Eisenia fetida v *Eisenia andrei* – sister species



Are we overfitting our data?

Data analysis is easy when the results are clear-cut!

- Things are more problematic when results are less transparent
- Ultimately, the author's responsibility to ensure that results are robust
- **Cannot rely on software to do everything for us!**

Need to use common-sense about analyses

Same set of earthworms/sites: **Can we distinguish worms from complex and polluted sites?**

- PLS-DA?

N= 302

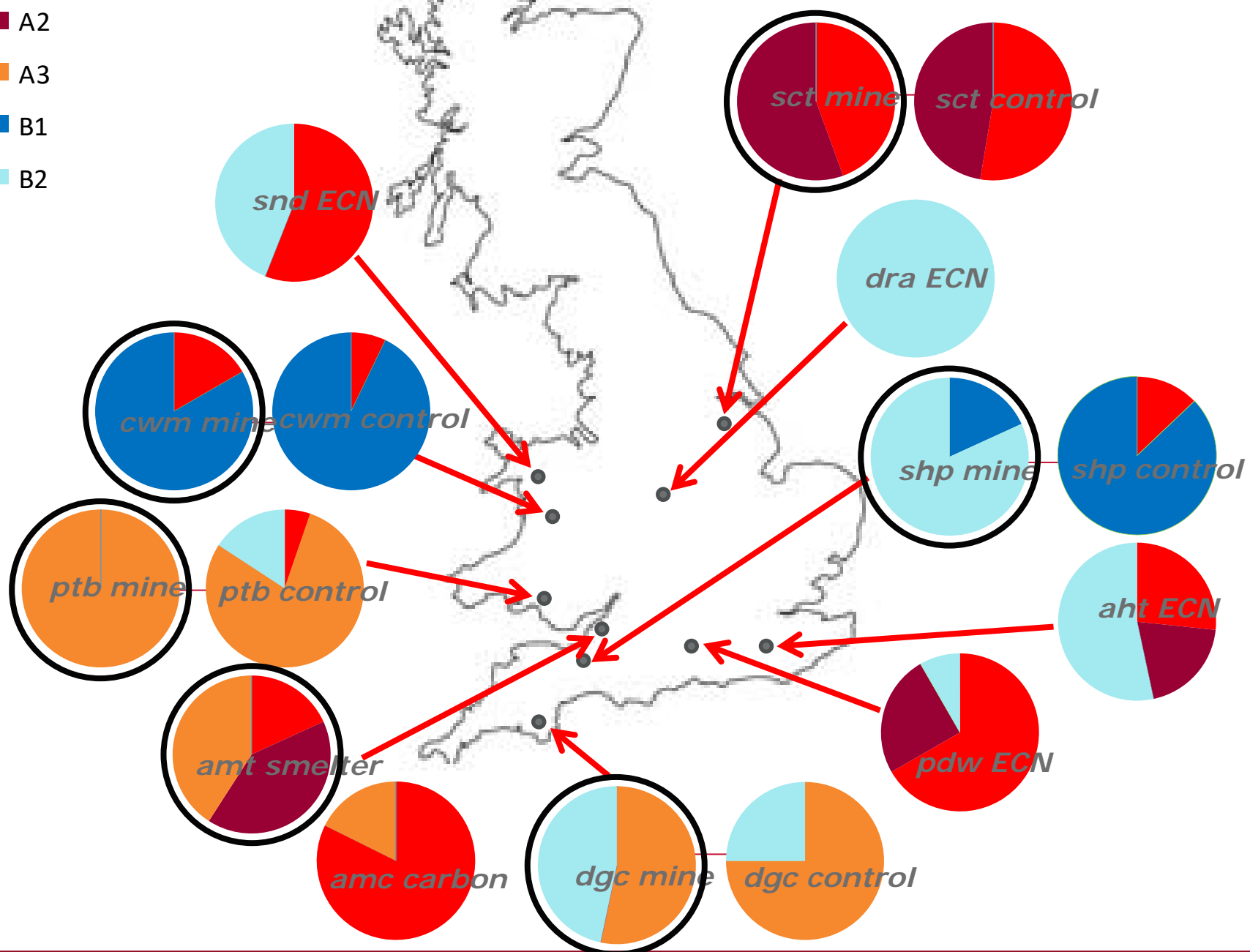
■ A1

■ A2

■ A3

■ B1

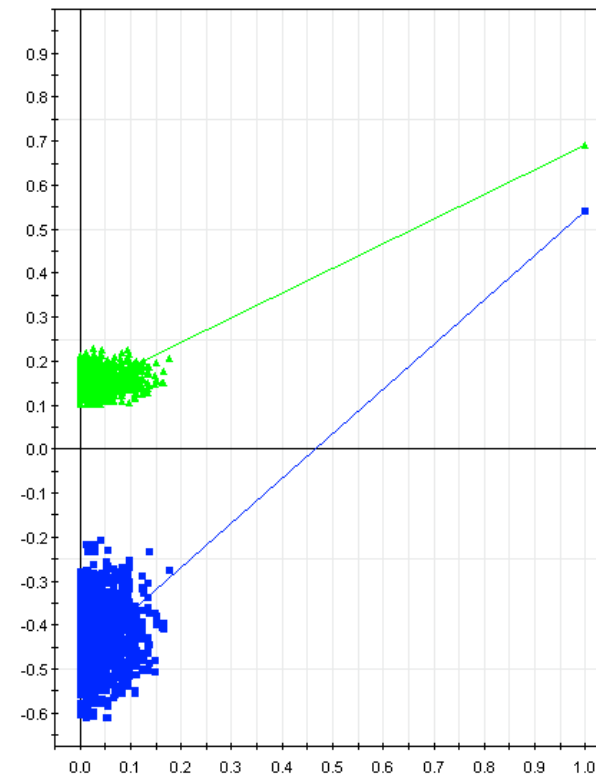
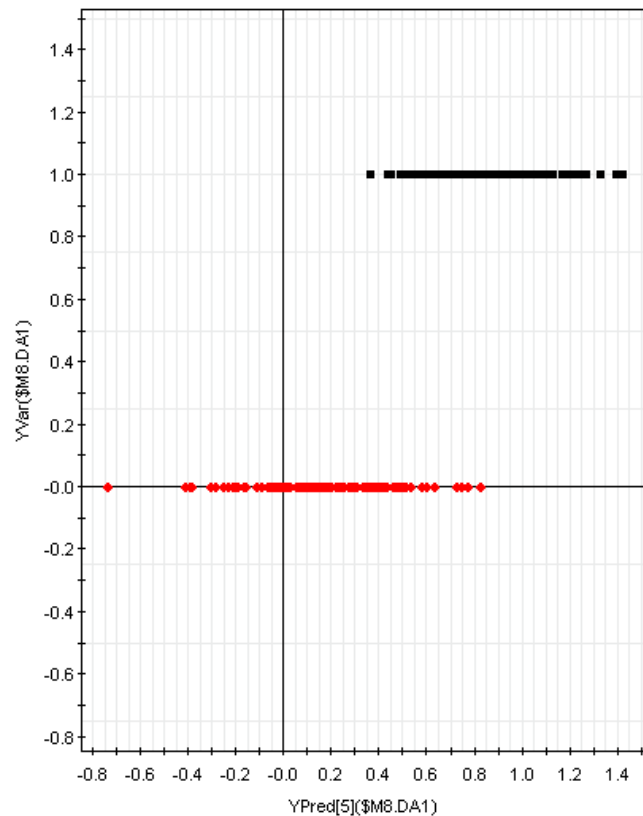
■ B2



Need to use common-sense about analyses

Same set of earthworms/sites: **Can we distinguish worms from complex and polluted sites?**

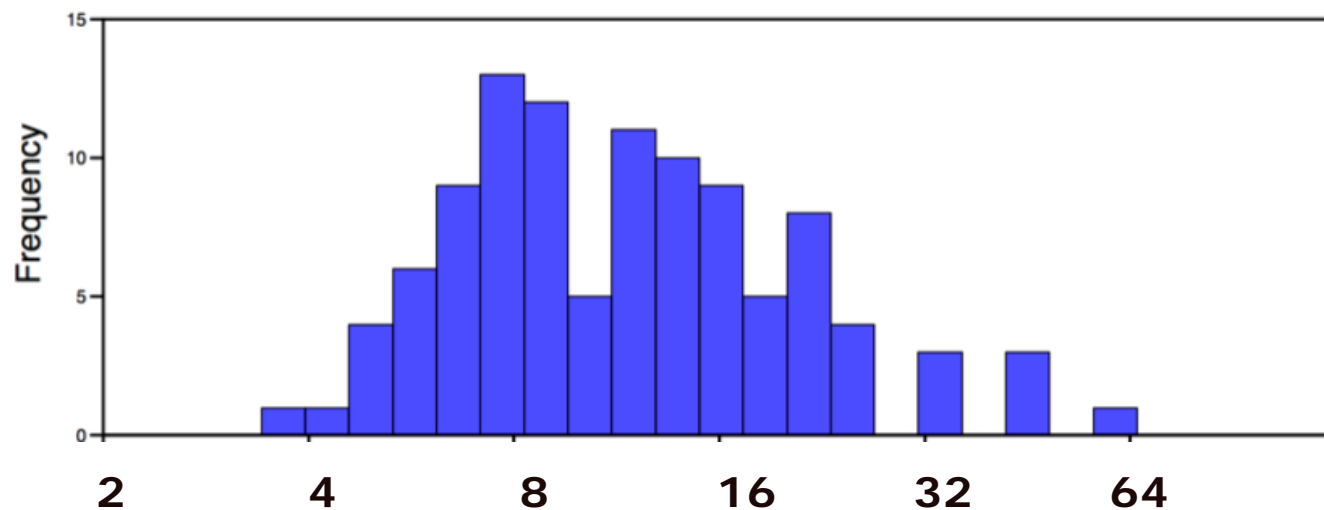
- PLS-DA?



Need to use common-sense about analyses

Same set of earthworms/sites: **Can we distinguish worms from complex and polluted sites?**

- PLS-DA?
- Data are confounded with the site the worms come from
 - In this particular case, between-site \gg within-site variation



Need to use common-sense about analyses

Same set of earthworms/sites: **Can we distinguish worms from complex and polluted sites?**

- PLS-DA?
- Data are confounded with the site the worms come from
 - In this particular case, between-site >> within-site variation
 - Ultimately, all the analysis is telling us is, can we distinguish one site from another
 - Switching the cross-validation to leave-one-group out solves the problem (can be done easily within Simca-P, for instance)
 - **In this case, no significant model can be fitted**

Avoiding overfitting – simplifying analyses

Allowing for one factor frequently clarifies the effects of another

Avoiding overfitting – simplifying analyses

Allowing for one factor frequently clarifies the effects of another

Optimized Phenotypic Biomarker Discovery and Confounder Elimination via Covariate-Adjusted Projection to Latent Structures from Metabolic Spectroscopy Data

Joram M. Posma,^{*,†,§} Isabel Garcia-Perez,^{†,¶} Timothy M. D. Ebbels,[†] John C. Lindon,[†]
Jeremiah Stamler,[#] Paul Elliott,^{§,⊥} Elaine Holmes,^{†,⊥,‡} and Jeremy K. Nicholson^{*,†,⊥,‡}

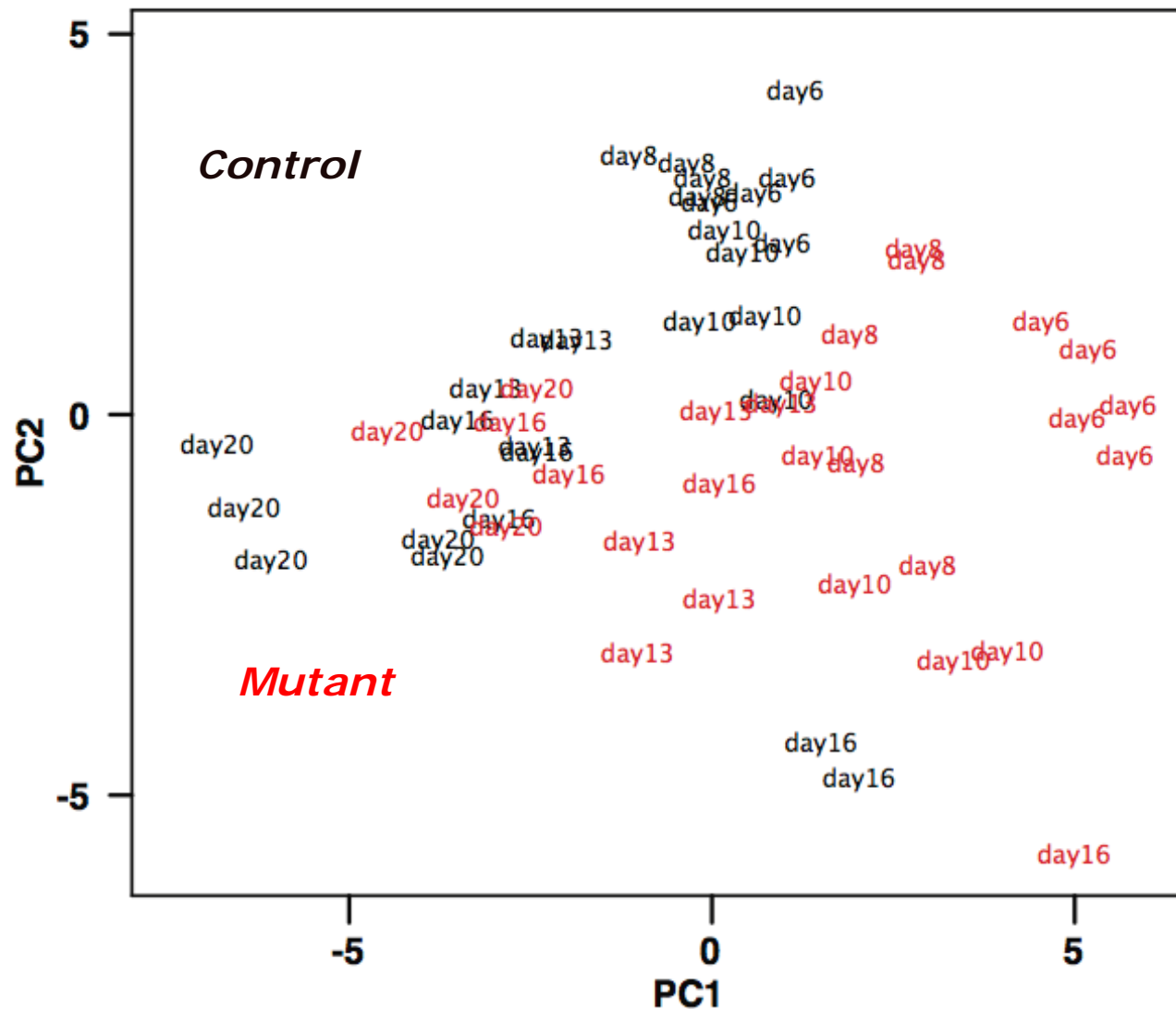
Avoiding overfitting – simplifying analyses

Allowing for one factor frequently clarifies the effects of another

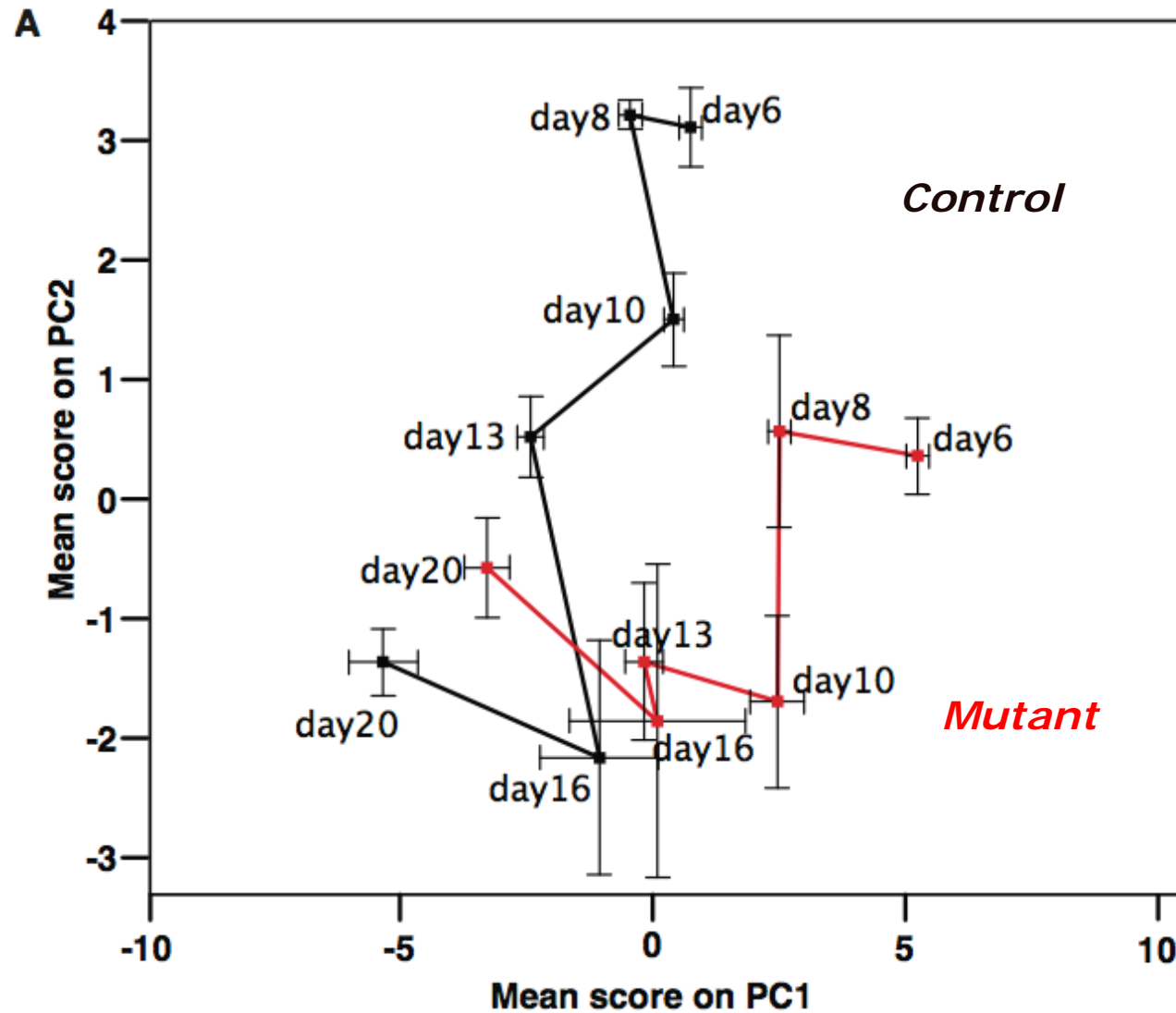
Example: designed experiment comparing wild-type and long-lived nematodes over time

- Two genotypes
- Five timepoints

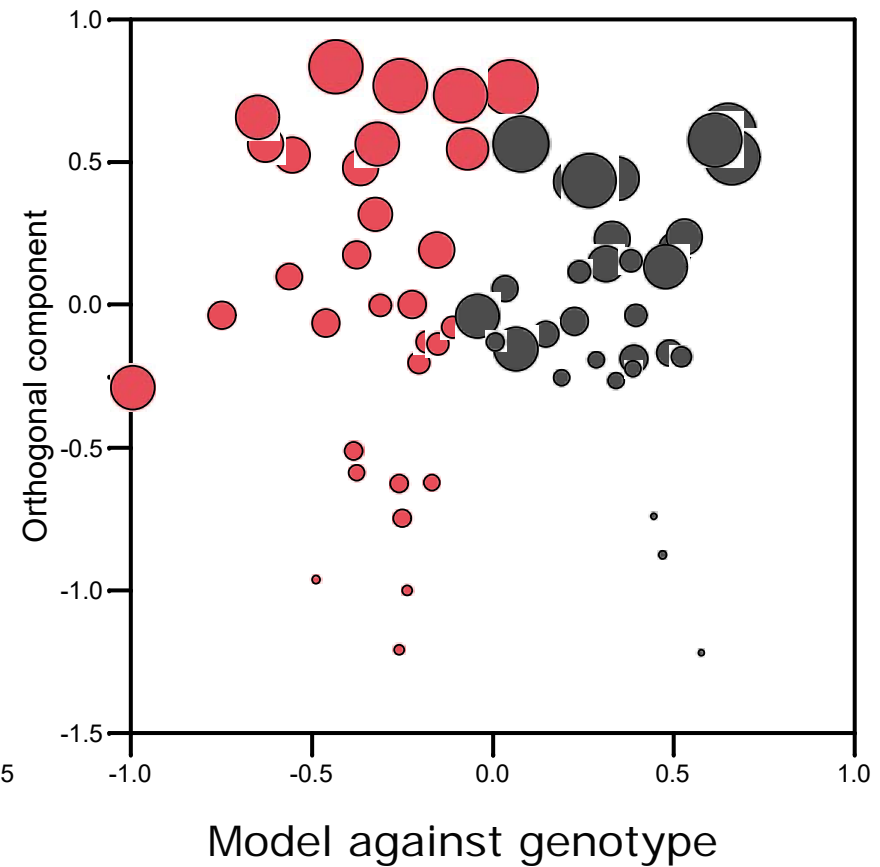
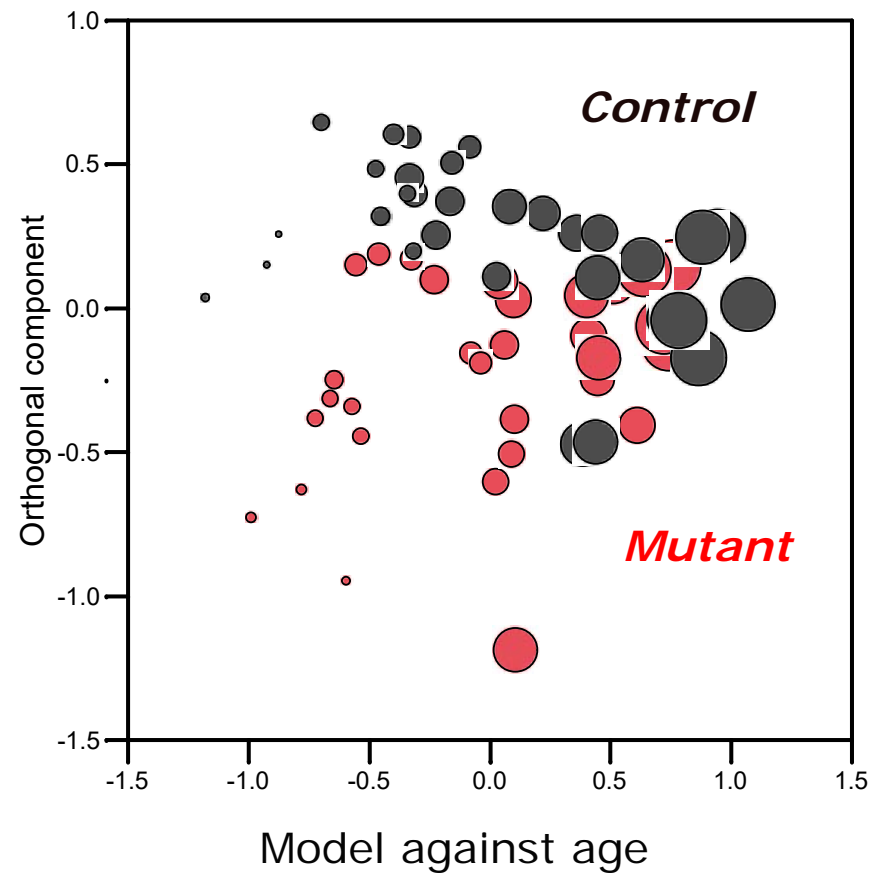
Avoiding overfitting – simplifying analyses



Avoiding overfitting – simplifying analyses



OPLS: 'unsupervised' orthogonal component



Conclusions

- **“Everything should be made as simple as possible, but not simpler”**
 - applies to selection of data analyses as well
- Do not just apply a single transformation to your data when analysing it
 - But I find that log transformation tends to give me the “best” results overall
- Multivariate and univariate analyses are both useful
 - Back up MVA with traditional statistical analyses of key biomarkers/variables
- Make sure your own results are robust
 - We usually know when we are squeezing data beyond a sensible point ...

