Designing a Complex-Omics Experiments

Xiangqin Cui

Section on Statistical Genetics
Department of Biostatistics
University of Alabama at Birmingham

1/7/2015

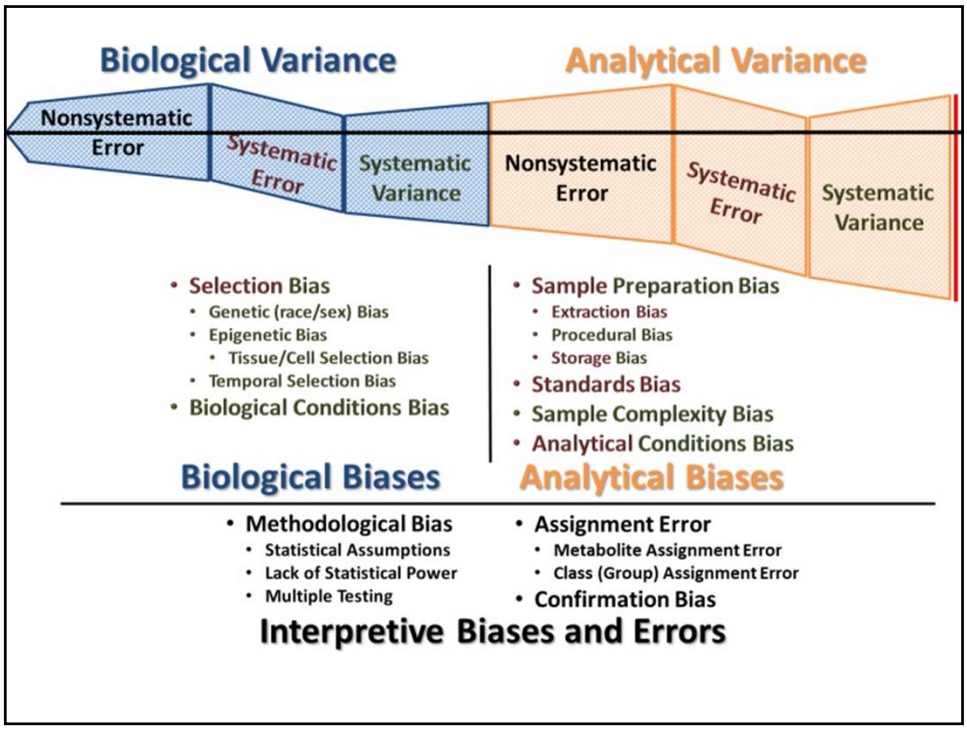
Some slides are from previous lectures of Grier Page

Experimental design

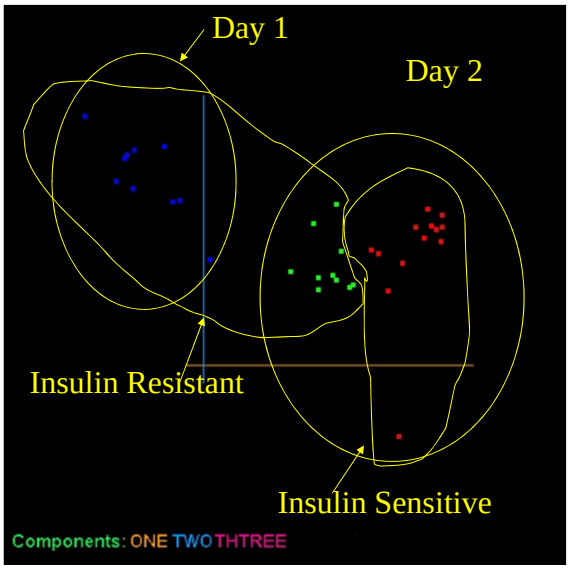
- ***Experimental design:*** is a term used about efficient methods for planning the collection of data, in order to obtain the maximum amount of information for the least amount of work. Anyone collecting and analyzing data, be it in the lab, the field or the production plant, can benefit from knowledge about experimental design.
  http://www.stat.sdu.dk/matstat/Design/index.html

The Myth That Metabolomics does not need a Hypothesis

- There always needs to be a biological question in the experiment.
- The question could be nebulous: What happens to the metabolome of this tissue when I apply Drug A.
- The purpose of the question drives the experimental design.
- Make sure the samples answer the question



UMSA Analysis



Experimental design general principals

- Randomization
- **Replication**
- **Blocking**
- Use of factorial experiments instead of the one-factor-at-a-time methods.
- Orthogonality

Randomization

- The experimental treatments are assigned to the experimental units (subjects) in a random fashion. It helps to eliminate effect of "lurking variables", *uncontrolled factors* which might vary over the length of the experiment.

Commonly used randomization method

- Number the objects to be randomized and then randomly draw the numbers using paper pieces in a hat or computer random number generator.

Example: Assign two treatments, Hormone and control, to 6 plants

1 2 3 4 5 6

Hormone treatment: (1,3,4) ; (1,2,6)

Control : (2,5,6) ; (3,4,5)

Design Issues in Omics Exp

- Known sources of non-biological error (not exhaustive) that must be addressed
  - Technician / post-doc
  - Reagent lot
  - Temperature
  - Protocol
  - Date
  - Location
  - Cage/ Field positions

Randomization in Omics Experiments

- Randomize samples in respect to treatments
- Randomize the order of handling samples.
- Randomize arrays/runs/gels/days in respect to samples

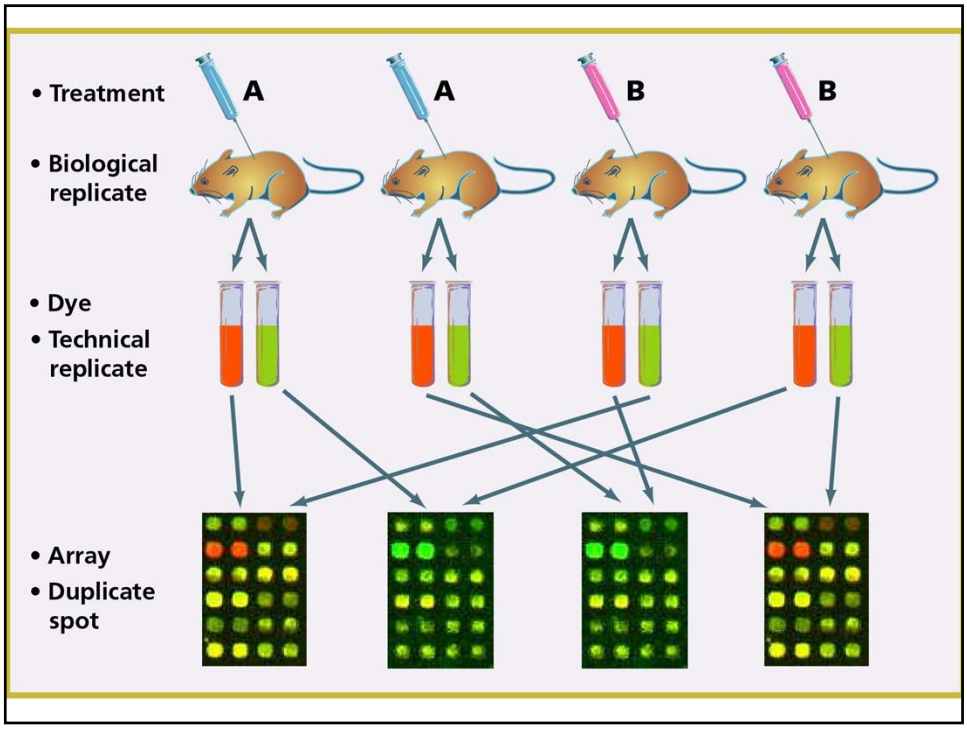
Replication

- **Replication** is repeating the creation of a phenomenon, so that the variability associated with the phenomenon can be estimated.

  Replications should not be confused with repeated measurements which refer to literally taking several measurements of a single occurrence of a phenomenon.

Replication in omics experiments

- What to replicate?
  - Biological replicates (replicates at the experimental unit level, e.g. mouse, plant, pot of plants…)
    - Experimental unit is the unit that the experiment treatment or condition is directly applied to, e.g. a plant if hormone is sprayed to individual plants; a pot of seedlings if different fertilizers are applied to different pots.
  - Technical replicates
    - Any replicates below the experimental unit, e.g. different leaves from the same plant sprayed with one hormone level; different seedlings from the same pot; Different aliquots of the same RNA extraction; multiple arrays hybridized to the same RNA; multiple spots on the same array.



Replication in Omics experiments

- Biological replicates are typically more important than technical replicates unless estimating the variation at different levels is the purpose of the experiment in evaluating the technology.
- Biological replicates are often more effective in increasing the power for detecting differentially expressed genes.

How Many to Replicate? ---Sample Size

- **Replication** is repeating the creation of a phenomenon, so that the variability associated with the phenomenon can be estimated.
- The accuracy of the estimation of the variability depends on the degree of freedom for estimating the variability.

  Degree of freedom (df) is a measure of the number of independent pieces of information on which the precision of a parameter estimate (e.g. variance) is based. The degrees of freedom for an estimate equals the number of observations (values) minus the number of additional parameters estimated for that calculation.

How many replicates? ---- Sample Size

Example: degree of freedom (df) for estimating the variance.

Using a 2x2 factorial design to examine the effects of two factors, A and B. Each factor has two levels.

ANOVA model:

$$y = \mu + A + B + A*B + \varepsilon$$

Factorial 2 x 2

	A1	A2
B1	r	r
B2	r	r

S.V.	(r=1)	(r=2)	(r=3)
μ	1	1	1
A	1	1	1
B	1	1	1
A*B	1	1	1
Var	0	4	8
Total	4	8	12

Sample Size and Power

•Sample size for a general two sample comparison

$$n = \frac{2\left(z_{(1-\alpha/2)} + z_{(1-\beta)}\right)^2}{(\delta/\sigma)^2}$$

- *n* increases as error, σ, increases.
- *n* increases as the difference between two means, δ,decreases.
- *n* increases as the significant level of the test, α, decreases.
- *n* increases as the power of the test, 1-β , increases.

Some R Power Packages in Bioconductor

- RNASeqPower
- Sizepower
- SSPA
- CSSP

Multilevel Replication and Resource allocation:

When there are both biological replications and technical replications.

Example: reference design with dye-swaps

Biological variation

Technical variation

$$EV = \frac{\sigma_M^2}{m} + \frac{\sigma_e^2}{mn}$$

Error variance of the fold change

m	**mouse / trt (biorep)**
n	**array pairs / mouse**
C_M	**cost / mouse**
C_A	**cost / array pair**

Note: to reduce EV increasing m (number of biological replicates) is more efficient.

Resource Allocation

Considering the error variance and the cost equations, we can obtain how many biological replicates and how many technical replicates to best allocate the money.

$$EV = \frac{\sigma_M^2}{m} + \frac{\sigma_e^2}{mn}$$ (reference design with dye-swaps)

$$Cost = mC_M + m \cdot nC_A$$

The optimum number of array pairs biological replicate:

$$n = \sqrt{\frac{\sigma_e^2}{\sigma_M^2} \cdot \frac{C_M}{C_A}}$$

Examples for resource allocation in early microarray experiments

- **Using variance components estimated from kidney in Project Normal data.**
- **No replicated spots on array**
- **Reference design**

Mouse price	Array price/pair	# of array pairs per mouse
$15	$600	1
$300	$600	1
$1500	$600	2

More efficient array level designs, such as direct comparisons and loop designs, can reduce the optimum number of arrays per mouse.

Pooling Biological Samples

Theoretically, pooling can reduce the biological variance but not the technical variances. The biological variance will be replaced by:

$$\sigma^2_{pool} = \frac{1}{k}\sigma^2_M$$

k : # of samples per pool
σ^2_M individual biorep variance
σ^2_{pool} pool variance

Note: It is often assumed that pooling will reduce the biological variance , therefore, be more efficient.

Potential problems of Pooling

- Reduced ability to estimate individual variability
- Prevent from identifying proper transformation and removing outliers.
- Not valid for classification studies (important for biomarker identification)
- Pooling samples is averaging at the raw level while the average of multiple samples is often after transformation (e.g. log2).
- The biological variability reduction is often smaller than 1/k.

$$\sigma^2_{pool} = \frac{1}{k^{\alpha}}\sigma^2_M$$

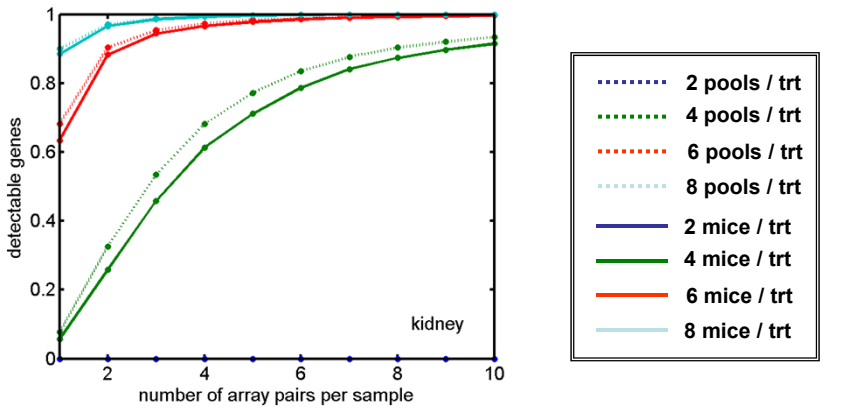
α *:* constant for the effect of pooling. $0 \leq \alpha \leq 1$
α = 1, pooling has maximum effect.
α = 0, pooling has no effect.
α < 0, pooling has negative effect.

Potential Advantage of Pooling

- When individual sample quantity is limited or technology is extremely expensive, pooling samples can increase the accuracy of the Fold Change estimation between two groups.
- Pooling has the potential to reduce the overall variance.

Example: Power Increase to Detect 2 fold change by Pooling in a mouse experiment (CAMDA 2002)

(Pool size *k* = 3, α = 1)



Significance level:
0.05 after Bonferroni correction

General Design Principles -- Continues

- Use of factorial experiments instead of the one-factor-at-a-time methods.
- Orthogonality: Factors are perpendicular to each other. Otherwise, the factors are called confounded or even nested.

To compare two treatments (T1, T2) and two strains (S1, S2)

	T1	T2
S1	T1S1	T2S1
S2	T1S2	T2S2

Blocking

- Some identified uninteresting but varying factors can be controlled through blocking.
  - COMPLETELY RANDOMIZED DESIGN
  - COMPLETE BLOCK DESIGN
  - INCOMPLETELY BLOCK DESIGNS

Completely Randomized Design

There is no blocking

➡ Example

- Compare two hormone treatments (trt and control) using 6 Arabidopsis plants.

1 2 3 4 5 6

Hormone trt: **(1,3,4); (1,2,6)**
Control : **(2,5,6); (3,4,5)**

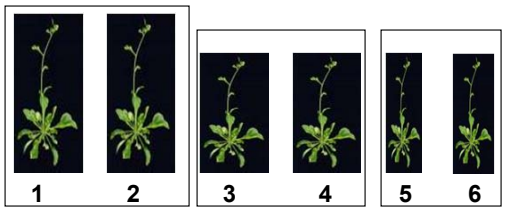
Note: Designs with one-color microarray is often completely randomized design.

Complete Block Design

➡ There is blocking and the block size is equal to the number of treatments.

Example:

- **Compare two hormone treatments (trt and control) using 6 Arabidopsis plants. For some reason plant 1 and 2 are taller, plant 5 and 6 are thinner.**

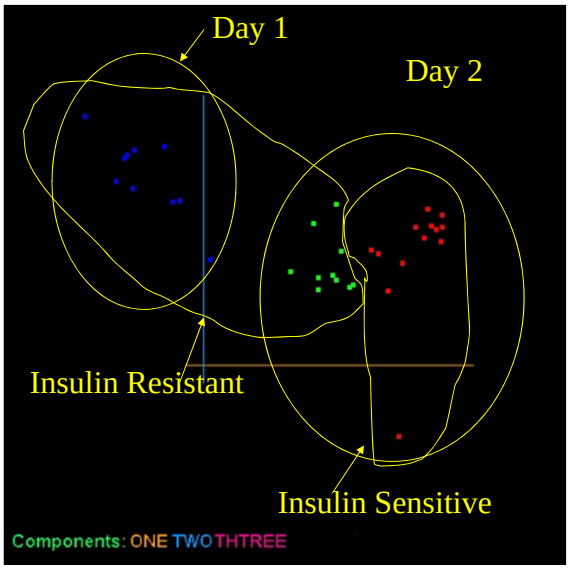


1 2 3 4 5 6

Hormone treatment: **(1,4,5) ; (1,3,6)**
Control : **(2,3,6) ; (2,4,5)**

⇨ Randomization within blocks

UMSA Analysis

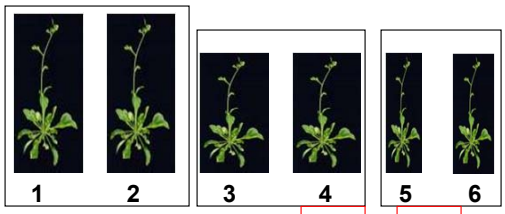


Incomplete Block Design

➡ There is blocking and the block size is smaller than the number of treatments.

Example:

- **Compare three hormone treatments (hormone level 1, hormone level 2, and control) using 6 Arabidopsis plants. For some reason plant 1 and 2 are taller, plant 5 and 6 are thinner.**



Hormone level1:	**(1,4)** ;	**(2,4)**
Hormone level2:	**(2,5)** ;	**(1,6)**
Control :	**(3,6)** ;	**(3.5)**

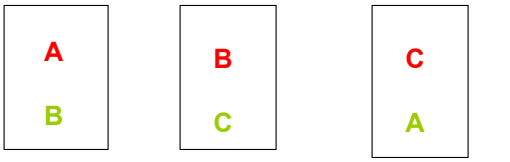
⇨ Randomization within blocks

Example: Incomplete Blocking in 2-color Microarray Experiments

• **In two-color platform, t**he arrays vary due to the loading of DNA quantity, spot morphology, hybridization condition, scanning setting… It is treated as a block of size 2, the samples are compared within each array.

• If there are two lots of chips in the experiment and there is large variation across chip lots. We can treat chip lots as a blocking factor.

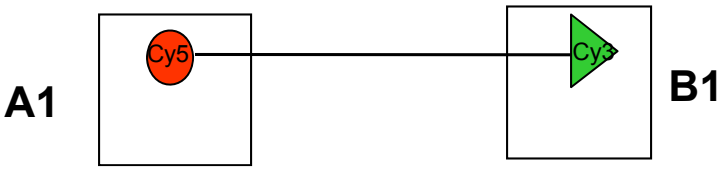
Example: compare three samples: A, B, C



Block (array) 1 **Block (array) 2** **Block (array) 3**

Designs for two-color microarrays

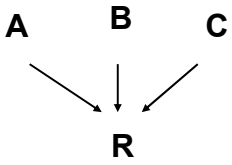
- For two color microarrays, we need to pair the samples and label them with Cy3 and Cy5 for hybridization to one array.– Blocking Designs
- How do we pair?



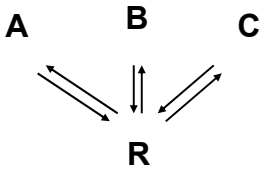
Reference design

All samples are compared to a single reference sample.
The reference sample is of no interest to the investigator.

Single reference

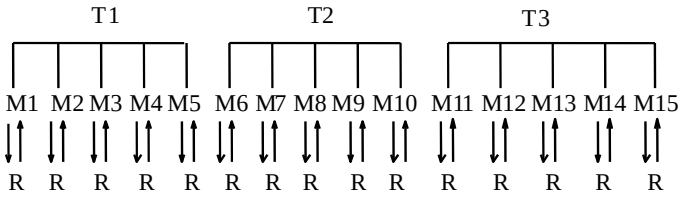


Double reference



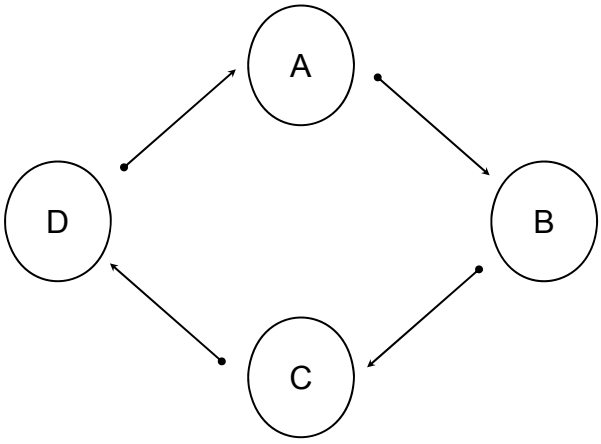
Example Microarray Experiment

comparing 3 treatments



Note: Arrows are used to represent a two-color microarray. The arrow head represents the red channel and the tail represents the green channel. T1 to T3, the three treatments; M1 to M15, the 15 plants; R, reference sample.

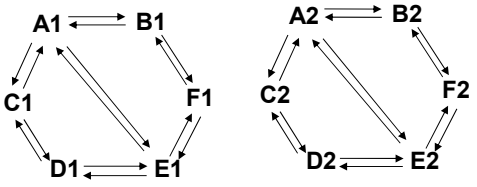
Loop Design

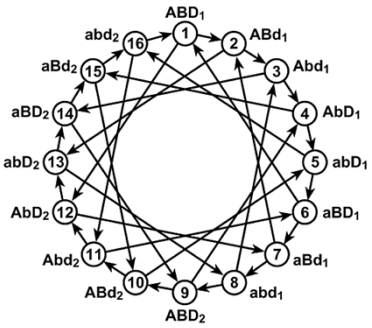


Complex loop designs

Two-factor factorial design at the high level

	Pera	l	DBA
High fat	A	B	C
Low fat	E	F	D





(Churchill and Oliver, 2001)

Statistical analyses

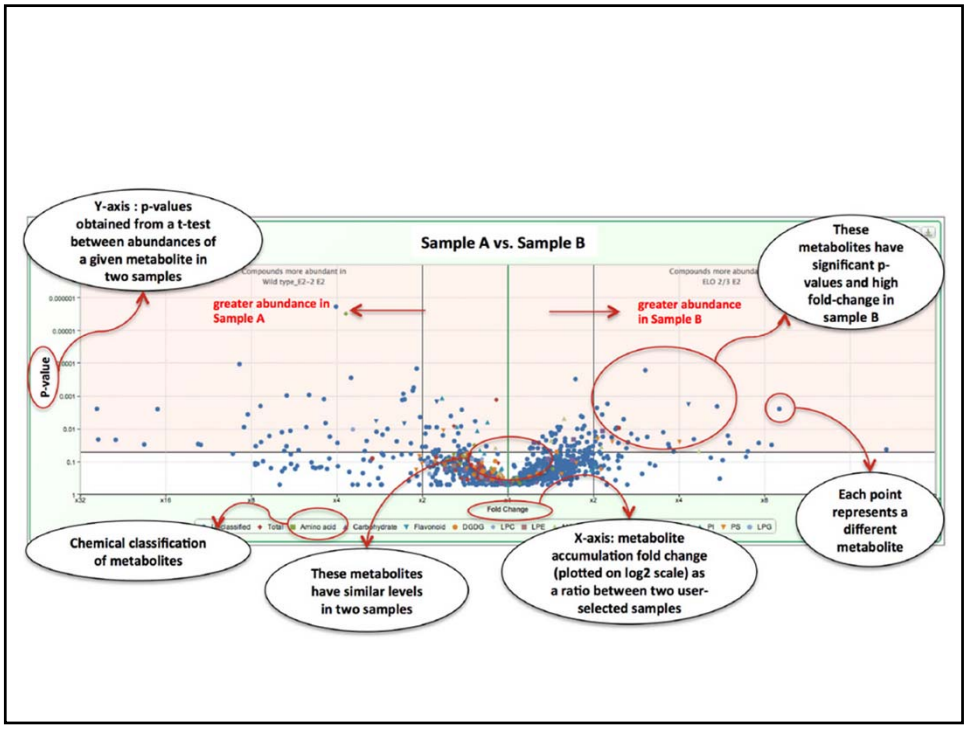
- Supervised analyses – linear models etc
  - Assume IID (independently identically distibuted)
  - Normality
  - Sometimes can rely on central limit
  - 'Weird' variances
  - Using fold change alone as a statistic alone is not valid.
  - 'Shrinkage' and or use of Bayes can be a good thing.
- False-discovery rate is a good alternative to conventional multiple-testing approaches.
- Pathway testing is desirable.

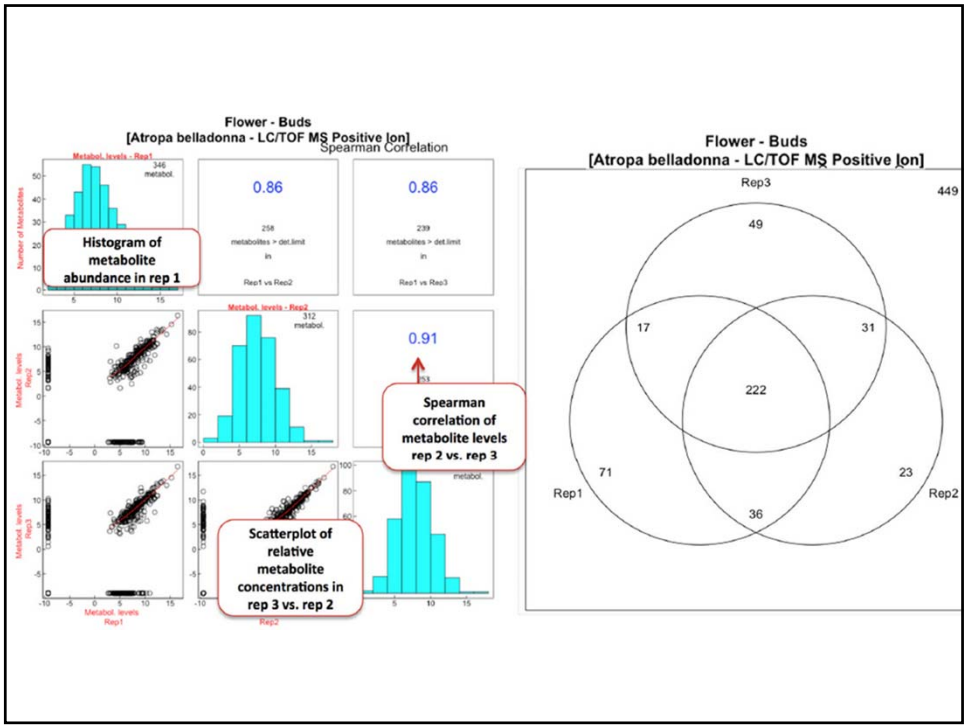
Classification

- Supervised classification
  - Supervised-classification procedures require independent cross-validation.
  - See MAQC-II recommendations Nat Biotechnol. 2010 August ; 28(8): 827–838. doi:10.1038/nbt.1665.
    - Wholly separate model building and validation stages. Can be 3 stage with multiple models tested
- Unsupervised classification
  - Unsupervised classification should be validated using resampling-based procedures.

Unsupervised classification - continued

- Unsupervised analysis methods
  - Cluster analysis
  - Principle components
  - Separability analysis
- All have assumptions and input parameters and changing them results in very different answers


Y-axis : p-values obtained from a t-test between abundances of a given metabolite in two samples
Sample A vs. Sample B
These metabolites have significant p-values and high fold-change in sample B
greater abundance in Sample A
greater abundance in Sample B
P-value
Fold Change
Each point represents a different metabolite
Chemical classification of metabolites
These metabolites have similar levels in two samples
X-axis: metabolite accumulation fold change (plotted on log2 scale) as a ratio between two user-selected samples


Flower - Buds
[Atropa belladonna - LC/TOF MS Positive Ion]
Spearman Correlation
0.86
0.86
0.91
Histogram of metabolite abundance in rep 1
Spearman correlation of metabolite levels rep 2 vs. rep 3
Scatterplot of relative metabolite concentrations in rep 3 vs. rep 2
Flower - Buds
[Atropa belladonna - LC/TOF MS Positive Ion]
Rep3
449
49
17
31
222
71
23
36
Rep1
Rep2

References

Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nature Genet. 32: 490-495, 2002.

Cui X and Churchill GA. How many mice and how many arrays? Replication in mouse cDNA microarray experiments, in "Methods of Microarray Data Analysis III", Edited by KF Johnson and SM Lin. Kluwer Academic Publishers, Norwell, MA. pp 139-154, 2003.

Gadbury GL, et al. Power and sample size estimation in high dimensional biology. Stat Meth Med Res 13: 325-338, 2004.

Kerr MK. Design considerations for efficient and effective microarray studies. Biometrics 59: 822-828, 2003.

Kerr MK and Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet. Res. 77: 123-128, 2001.

Kuehl RO. Design of experiments: statistical principles of research design and analysis, 2nd ed., 1994, (Brooks/cole) Duxbry Press, Pacific Grove, CA.

Page GP et al. The PowerAtlas: a power and sample size atlas for microarray experimental design and research. BMC Bioinformatics. 2006 Feb 22;7:84.

Rosa GJM, et al. Reassessing design and analysis of two-colour microarray experiments using mixed effects models. Comp. Funct. Genomics 6: 123-131. 2005.

Wit E, et al. Near-optimal designs for dual channel microarray studies. Appl. Statist. 54: 817-830, 2005.

Yand YH and Speed T. Design issues for cDNA microarray experiments. Nat. Rev. Genet. 3: 570-588, 2002.