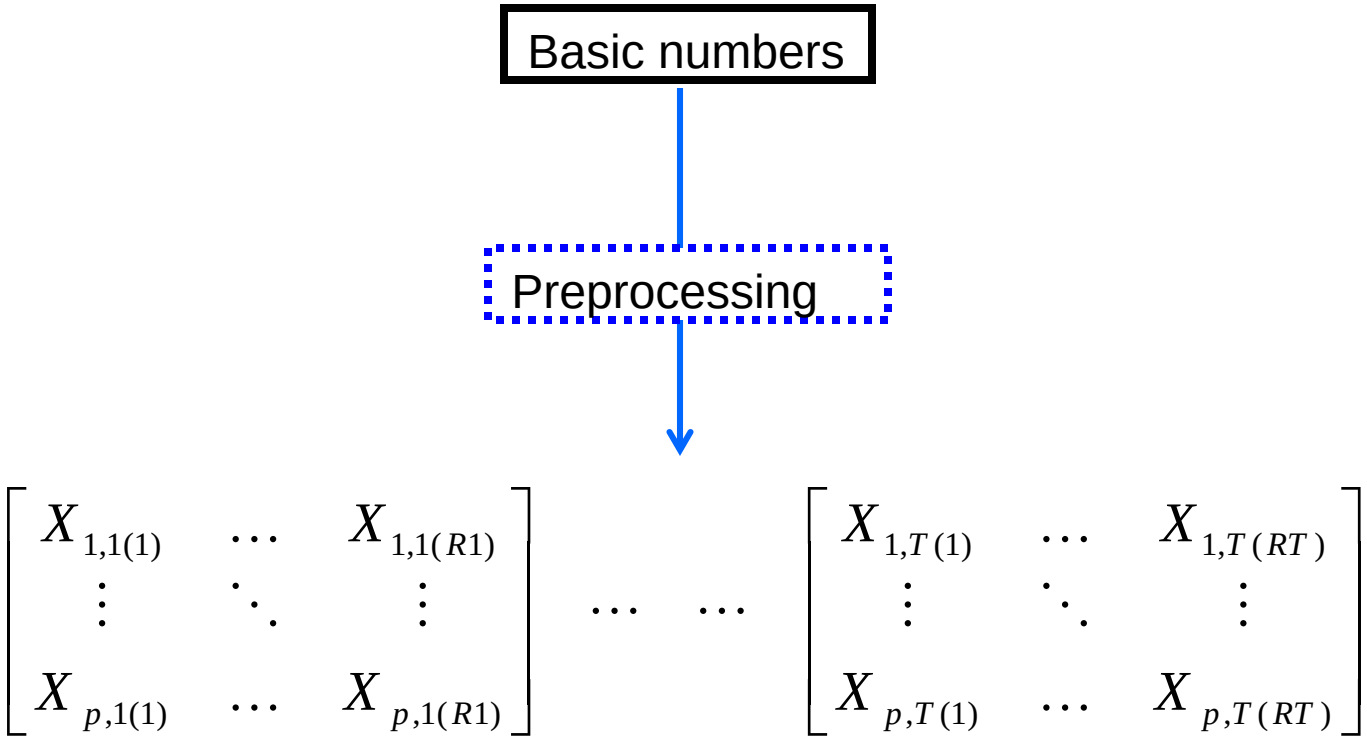
Preprocessing I:

Normalization and Missing Values



Comparable numbers, on a scale and format appropriate to the questions and the data analysis tools one intends to use. Possibly no missing values, possibly a reduced set of genes.

Normalization:

- Ensure comparability between two sets of expression measurements; experimental condition vs control.
- Important generalization: ensure comparability among several sets of measurements; multiple experimental conditions.
- Mitigate unwanted (non-experimental) variation in expression measurements.

For example: **List of possible sources of unwanted variation in spotted arrays**

Preparing the samples:
- mRNA preparation
- Reverse transcription to cDNA
- Dye labeling

Spotting the chips:
- PCR amplification
- Pin geometry and surface features
- Amount of cDNA transported by pins
- Amount of cDNA fixated on slide

Hybridization process:
- Hybridization parameters (temperature, time, amount of sample)
- Spatial dis-homogeneity of hybridization on the slide
- Non-specific hybridization

Image production and processing:
- Non-linear transmission, saturation effects, variations in spot shape
- Global background shining, and local overshining from neighboring spots

Et cetera …(source: Schuchhardt *et al*. 2000)

Computing scaling factor(s):

Relative "activity" on two colors, or on two chips (normalizing total equivalent to normalizing mean; normalizing median follows a similar logic, but robust to outliers)

$$\frac{Exper(i)}{Contr(i)} \times \rho$$

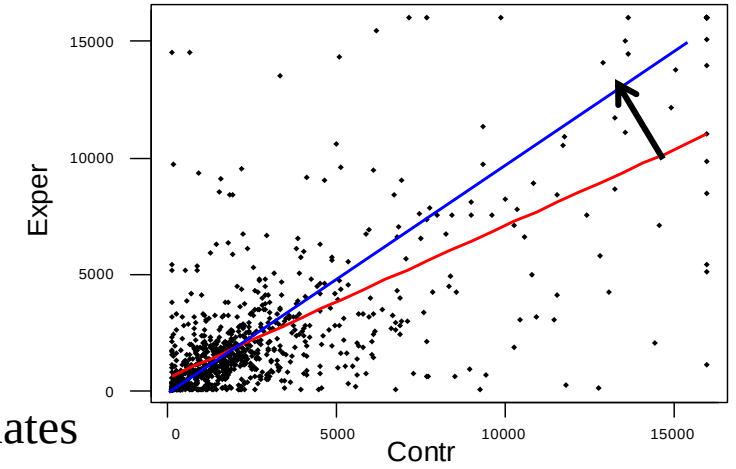
$$\rho = \frac{K_{Contr}}{K_{Exper}} \quad \text{e.g.} \quad K_w = \sum_i W(i)\ ,\ w = Exper, Contr$$

Computing linear trend(s):

Add the possibility of a translation

$$\frac{Exper(i)}{\alpha + \beta\, Contr(i)}$$

α, β intercept and slope, e.g. least square estimates



Global Normalization: scaling factors = "total signals"; LS estimates on all points.

Underlying rationale:

- only a small share of the "genes" is subject to significant experimental changes in expression
- changes tend to compensate as to not significantly affect the normalization quantities.

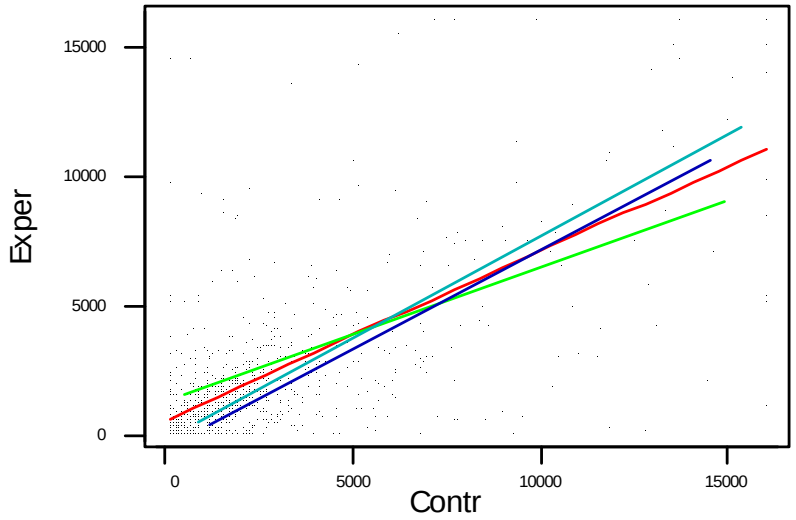
Issue: is this reasonable? Not in some experiments

Variations on the theme 1: (Finkelstein *et al*. 2002, spotted)

• compute normalization quantities separately for groups of genes, where the partition captures an obvious source of unwanted variation (e.g. pin/sector in spotted arrays)

• iteratively, excluding outlying genes at each iteration (outlying genes are those with sizeable differences between *Exper* and *Contr*).

$$\frac{Exper(i(g))}{\alpha_g + \beta_g\, Contr(i(g))}$$



Variations on the theme 2: (Hartemink *et al*. 2001, affy, multiple)

• **Control-based normalization**: compute normalization quantities on a subset of genes that ought not to show systematic experimental variation (housekeeping genes, spiked controls)

• through models describing variation sources, possibly using maximum likelihood estimation techniques, or Bayesian techniques that allow for priors (informative; for technical reasons).

i's = spiked controls

$$x_{ij} = m_i t_j e_{ij}$$

Assumptions: multiplicative error, normal on log scale

$$\tilde{x}_{ij} = s_j x_{ij} \qquad s_j = \frac{1}{t_j}$$

$$\log(x_{ij}) = \mu_i + \tau_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

$$\hat{s}_j = e^{-\hat{\tau}_j}$$

Maximum likelihood or maximum a posteriori estimation (prior on variances to avoid that spiked controls with small variance dominate too strongly the estimation)

Issue: are spiked controls a reliable tool for normalization? high variability, locations.

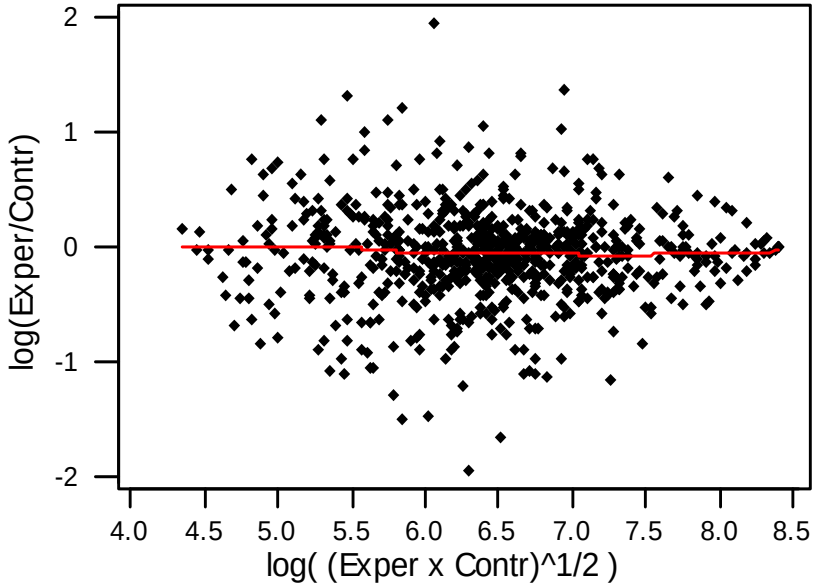
Variations on the theme 3: (Yang *et al*. 2001, spotted)

• compute normalization quantities using different scales and/or mappings for *Exper* and *Contr*, and consider mean relationships other than a linear trend (e.g. non-parametric lowess)

$$\log(Exper(i)/Contr(i))$$

$$-\varphi[\log(\sqrt{Exper(i)\times Contr(i)})]$$

(distribution of log-ratios centered at 0 at each average log-intensity level)



• again, can apply within gene groups

• adjust also for different variances about “mean relationships” within gene groups.

Missing Values:

Deleted rows (genes) with missing entries from the analysis?

… if the number of missing entries in a row is not too high we can retain the row, filling in the missing values according to some rationale.

Unsophisticated solutions:

• Averages over gene (spot) and/or condition (chip) replicates – if available

• Averages over conditions (average expression for the gene whose profile contains missing entries).

• For time courses, interpolation of nearby values in the profile.

More sophisticated solutions: (Troyanskaya *et al*. 2001)

• Identify a set of genes whose expression profile is similar to the one of the gene with missing entries and take averages over these genes, with weights inversely proportional to the similarity. Need to:

- choose a metric for profile similarity
- choose size of the neighbors set

• Form a set of basic expression patterns (e.g. principal components) and take linear combinations of them, with coefficients determined by proximity of the gene with missing entries to the patterns. Need to

- choose size of the patterns set
- iterate (from initial averaging imputations)

Missing value imputation: big research area in statistics.

If a data set contains a large share of missing entries, imputation can affect the analysis substantially (inducing spurious features).

In evaluating an imputation procedure, a core issue is what assumptions can be made on the nature of the process that produces missing entries. Let

X = data = $(X(obs),X(miss))$

R = indicators of whether elements of X are obs or miss

- Missing completely at random:
  $Pr(R|X(obs),X(miss)) = Pr(R)$
  does not depend on the values in X .
- Missing at random:
  $Pr(R|X(obs),X(miss)) = Pr(R|X(obs))$
  depends on the values in X only through the ones we get to observe.
- Missing NOT at random:
  $Pr(R|X(obs),X(miss))$
  depends also on the values we do not get to observe, the most complicated situation.

- Multiple Imputation (notes posted on web-site)
- For some analyses (e.g. mixture-based clustering) EM can perform imputation.