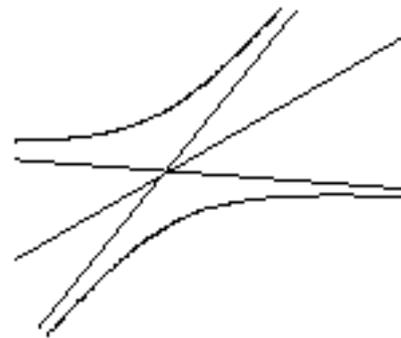# A Short Review of Some Classical Statistical Models
(Hopefully With Some Lessons for Differential Gene Expression Analyis)

Günther Sawitzki
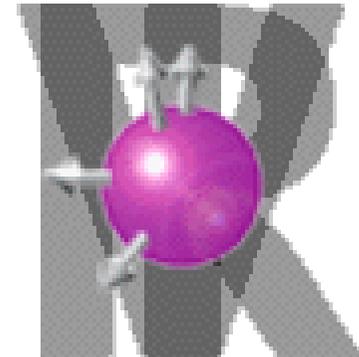StatLab Heidelberg
gs@statlab.uni-heidelberg.de

$$\hat{a}+\hat{b}x \pm \sqrt{2F_{6,4-6;2-6}\,s^{k}\left(\frac{1}{M}+\frac{(x-\overline{x})^{k}}{\sum(x_{i}-\overline{x})^{k}}\right)}$$

Universität Heidelberg
www.uni-heidelberg.de

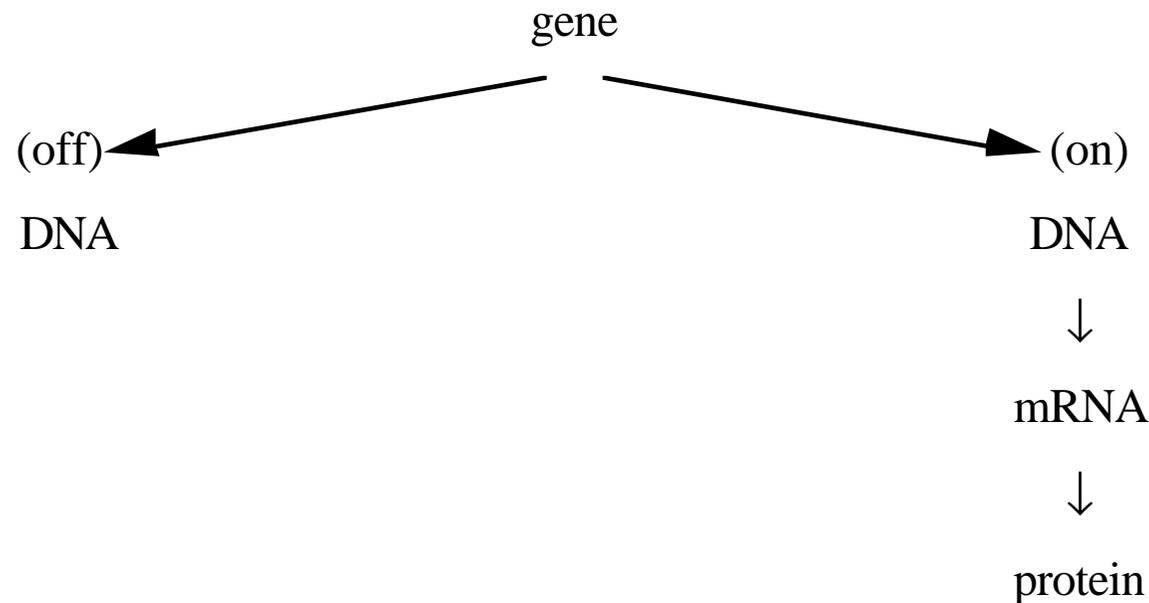**StatLab Heidelberg**
www.statlab.uni-heidelberg.de

Virtuelle Universität Oberrhein
www.viror.de

1. Experimental Background

Differential Gene Expression

A simplified view (a non-biologist's view - please correct)

```
                              gene
                            /      \
           (off) ◄─────────         ─────────► (on)

              DNA                              DNA

                                                ↓

                                              mRNA

                                                ↓

                                             protein
```

"gene" used loosely:

      maybe an experimentally defined DNA segment, i.e. defined by preparation

      maybe an allele, i.e. a specific gene variant
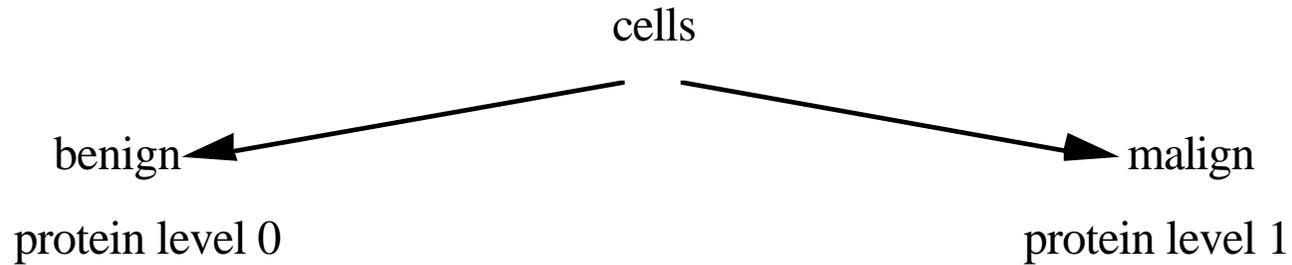
Switching on/off may be

      dynamic: inhibited $\leftrightarrow$ activated   by operon

      static: allele present $\leftrightarrow$ not present

Differential Gene Expression Diagnostics

A simplified view (a non-biologist's view - please correct)

General idea

cells

benign ← → malign

protein level 0   protein level 1

Open question: which proteins?
Working program: assume classification of cells as benign/malign given.

Classify proteins by protein level !

		benign
malign	low/low	low/high
	high/low	high/high

Next step: *classify genes !*

Pragmatic Considerations

DNA/RNA and proteins may be labeled
        e.g. dyes Cy5 (red), Cy3 (green)
        e.g. radioactive markers P

DNA/RNA bind to complementary chains
        allows hybridization technology for detection

DNA/RNA transcription accessible

DNA/RNA are "auto-reproductive"
        allows cloning

        DNA hybridization experiments are feasible on a large scale
        protein experiments require protein specific setup

        ⌒ *use DNA/RNA as substitute when many proteins are under question !*

Hybridization Experiments

Time line

clone cDNA   locate on spots →   ↓   ← label samples   ← extract RNA

probes →   samples

hybridize

↓

wash out

↓

scan label

intensity

↓

infer on bound

concentration

Hybridization Experiments (cont.)

Typical Figures

probes   select from $\approx$ 100 000 genes or fragments $\approx$ 8 000 per experimental unit

samples   1, 2 per case $\approx$ few to some 100 cases
                   most samples provide sufficient materieal for about 4 experimental units (hybridizations)
                   up to 8 experimental units are often achievable, exceptionally about 10

labels   2 or more dyes,
                   or radioactive $\approx$ 1 or 2 per experimental unit

Note: samples are not pure (tumor sample may contain benign cells).

Hybridization Experiments (cont.)
Typical Technologies

• glass slides & dye labeled samples

  scanner records colours in separate channels

  slides are not re-usable

  typical: two samples per slide, labeled Cy5 (red) and Cy6 (green)

• nylon filters & radiactive marker

  scanner gives integrated decay counts

  spotted filters may be re-used (advantage)

  hybridizations are sequential in time (disadvantage: possible decay of probes)

  typical: two filters for two samples, swapped in two runs:

  first: sample 1 one filter 1, sample 2 one filter 2

  then: sample 2 one filter 1, sample 1 one filter 2

Unified Terminology

  "Carrier"   slide or filter

  "Channel"   colour (for slide) or reuse (for filters)

Hybridization Experiments (cont.)

2 sample version simplified

Experiment

spots	i	$i: 1...\approx 8000$
genes allocated to spots	$gene(i)$	$gene(i) \in \{ 1...100\ 000\}$
samples	j	$j=1, 2$

Response

label	ℓ	corresponds to channel
label intensities	$Y_{i,\ell}$	$\ell = \ell(j) \triangleq j = 1,2$

Task

Classify genes		$gene$
or spots		$i: gene(i)=gene$
or observations		$(Y_{i,\ell})_{i:\ gene\ (i)=\ gene;\ \ell=1,\ 2}$

...more to come

2. Statistical Background

Statistical Basics

Use stochastic models to analyze data

observable	probability measure P
(experimental outcome)	distribution (distribution function $F = F_P$, $F(y) = P(Y \leq y)$)
Y $\leftrightarrow$	$\mathscr{L}(Y) = P$
	or $Y \sim P,\ Y \sim F$

Reference scenario

independent identical observations

stochastically **i**ndependent repeated experiments with **i**dentical result **d**istribution (**iid**)

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$\mathscr{L}(Y_i) = P$ identical, for all repetitions $i,\ i=1...n$

$\mathscr{L}(Y) = P^n$ product distribution for independent combination

corresponds to reproducible experiments

Parametrization

Sometimes

$$\mathscr{L}(Y) = P = P_{\vartheta}, \ \vartheta \in \Theta$$

$$\text{e.g. } P = N(\mu, \sigma^2)$$
$$\vartheta = (\mu, \sigma^2)$$

Note:

Focus is on distribution.

Parameters are only of technical importance.

Derived parameters

e.g. Location, Scale

Needs intelligence, not optimization.

for example location:

find $\gamma^* = \arg \min_{\gamma} \mathbf{E}|Y - \gamma|$ solution: $\gamma^* = med(Y)$

find $\gamma^* = \arg \min_{\gamma} \mathbf{E}|Y - \gamma|^2$ solution: $\gamma^* = \mathbf{E}(Y)$ if exists

Mean value $= E(Y)$ is computationally simple, but sample mean $\overline{Y} = \widehat{\mathbf{E}(Y)}$ has breakdown point 0.

Parametrization (cont.)

Moments (mean, variance…) follow linear algebra.

For Gaussian distributions: mean and variance are sufficient parametrization.

Gaussian distributions in statistics have a similar role as linear functions in analysis.

☞ Use Gaussian distribution as first approximation (and reduce computation to mean and variance).

Take-Away Message:	
	The world is not flat,
	and not all distributions are gaussian.

If better approximations are needed
          use higher expansions
          use transformation to improve approximation

Comparison of distributions

Simplified: two treatments, repeated measurements, shift only

Data $Y=(Y_{ij})$

i	sample	i=1, 2
j	repetition	$j=1\ldots J_i$

Model assumptions

independent observations

only shift in location

$Y_{ij} \sim F_i$ i=1, 2   independent

$$F_i(y) = F_0(y- \mu_i)$$

Test for shift:

test $\mathcal{H}_0 = \{\mu_1 = \mu_2\}$ against $\mathcal{H}_1 = \{\mu_1 \neq \mu_2\}$

Take-Away Message:	
	Shift is just one very special way how distributions can differ.

Comparison of distributions (cont.)

Two treatments, repeated measurements, shift only.

Abbreviations:

$$\overline{Y_{i,.}} = \frac{1}{J_i} \sum_{j=1...J_i} Y_{i,j} \qquad \text{for i=1, 2}$$

$$s^2_i = \frac{1}{J_i-1} \sum_{j=1...J_i} (Y_{i,j} - \overline{Y_{i,.}})^2$$

Solution for Gaussian distributions: t-test

Criterion:

$$t = (\overline{Y_{1,.}} - \overline{Y_{2,.}}) / s\sqrt{\frac{1}{J_1}+\frac{1}{J_2}} \quad \text{where } s^2 = ((J_1-1)s^2_1 + (J_2-1)s^2_2) / (J_1+J_2-2)$$

Test distribution:

If $\mu_1 = \mu_2$, the test statistics t has $t(J_1+J_2-2)$ distribution.

Take-Away Message:	
	! standardization applies to decision statistics, not data
	! standardization uses information from both sub samples.
	! Groupwise standardization using $(\overline{Y_{i,.}}/\sqrt{J_i}\, s^2_i)$ is feasible, but sub optimal.

Comparison of distributions (cont.) - non gaussian

Two treatments, repeated measurements, shift only.

Assume continuous, but not necessarily gaussian distribution:

Solution: Wilcoxon-test

$$R_{ij} \qquad \text{rank of } Y_{ij} \text{ in joined data set}$$

Criterion:

$$t_w = (\overline{R_{1,.}} - \overline{R_{2,.}}) / s \quad \text{where s depends on } J_1, J_2 \text{ only.}$$

Test distribution:

If $\mu_1 = \mu_2$, the test statistics t_w has (tabulated) Wilcoxon distribution.

Take-Away Messages:	
	! standardization implicit, reduced to sample size adaption.
	Asymptotic relative efficiency $\approx 95\%$.
	! transformation to normality can be avoided (sizable loss)

Comparison of distributions: Gaussian distribution, unequal variance

Simplified: two treatments, repeated measurements

Data $Y=(Y_{ij})$

i	sample	i=1, 2
j	repetition	$j=1\ldots J_i$

Model assumptions

      independent observations

      only shift in location

      $Y_{ij} \sim N(\mu_i, \sigma_i^2)$ i=1, 2   independent

Test for shift:

      test $\mathcal{H}_0 = \{\mu_1 = \mu_2\}$ against $\mathcal{H}_1 = \{\mu_1 \neq \mu_2\}$

Behrens-Fisher-Problem.

Take-Away Message:
Even simple tests for shift can lead to deep problems...

Comparison of distributions: Gaussian distribution, unequal variance

Simplified: two treatments, repeated measurements

Behrens-Fisher-Problem

Approximative solution: t-test
Criterion:

$$t = (\overline{Y_{1,.}} - \overline{Y_{2,.}}) / \sqrt{\frac{s^2_1}{J_1} + \frac{s^2_2}{J_2}}$$

Approximative test distribution:

If $\mu_1 = \mu_2$, the test statistics t has t $\left(\dfrac{1}{\dfrac{c^2}{J_1-1} + \dfrac{(1-c)^2}{J_2-1}} \right)$ distribution.

where

$$c = \frac{\dfrac{s_1^2}{J_1}}{\dfrac{s_1^2}{J_1} + \dfrac{s_2^2}{J_2}}.$$

Comparison of distributions: non-iid

Simplified: two treatments, repeated measurements, shift only, but casewise mean

Data $Y=(Y_{ij})$

i	sample	$i=1, 2$
j	repetition	$j=1\ldots J_i$

Model assumptions

independent observations

only shift in location

$Y_{ij} \sim N(\boldsymbol{\mu_{ij}}, \sigma^2)$ $i=1, 2$ independent

Test for shift:

test $\mathcal{H}_0 = \{\mu_{1j} = \mu_{2j}\}$ against $\mathcal{H}_1 = \{\mu_{1j} \neq \mu_{2j}\}$

$Y_{ij} = \mu_j + \alpha_i + err_{ij}$

$Y_{1j} - Y_{2j} = \alpha_1 + err_{1j} - \alpha_2 + err_{2j}$

matched pairs t-test based on $d_j = (Y_{1j} - Y_{2j})$; case mean μ_j cancels out.

Take-Away Message/ Home Exercise:
Using dye swap may cancel out spotting effects, thus reducing variance.

General Linear Models: Matrix Representation

$$E(Y) = X\beta \qquad \begin{array}{ll} Y & \text{values in } \mathbb{R}^N \\ X & N \times p \text{ - Matrix} \\ \beta & \in \mathbb{R}^p \end{array}$$

For example, simple linear regression $\quad y_i = \mu + \sum x_{ij}\,\alpha_j + err_i \qquad\qquad i=1..N,\ E(err_i)=0$

corresponds to $\quad Y = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,p} \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} + err.$

Simple one way analysis of variance model $\quad y_{ij} = \mu + \alpha_i + err_{ij} \qquad\qquad i=1..I,\ j=1..N_i,\ E(err_{ij})=0$

corresponds to $\quad Y = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1N_1} \\ \vdots \\ y_{J1} \\ \vdots \\ y_{1N_I} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & \vdots & & & \vdots \\ \vdots & 1 & 0 & & \vdots \\ \vdots & 0 & 1 & & \vdots \\ \vdots & & 1 & & 0 \\ \vdots & & 0 & & 1 \\ 1 & \vdots & & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix} + err$

General Linear Models: Space Representation

$$E(Y) \in \qquad \mathcal{M} = \left\{ X\beta : \beta \in \mathbb{R}^p \right\} \qquad \text{linear model}$$

If $\mathcal{M}_L$ is a linear model and $\mathcal{M}_H \subset \mathcal{M}_L$ is a linear subspace, then Y is said to satisfy the linear hypothesis $\mathcal{M}_H$ if $E(Y) \in \mathcal{M}_H$.

Example: One way classification

Let $\qquad\qquad \mathcal{M} = \mathbb{R}^N$

and $\qquad\qquad \mathcal{M}_1 = \{\mu \in \mathbb{R}^N : \mu_{ij} = \mu_i \in \mathbb{R} \text{ for } j=1..J_i, i=1..I\}$

$\mathcal{M}_1 \subset \mathcal{M}$ is a linear subspace in $\mathcal{M}$.

Y satisfies the linear model $\mathcal{M}_1$, if in each group i the observations Y_{ij} have same mean:

$\forall i \qquad\qquad \exists\, \mu_i \in \mathbb{R}: \forall j=1..J_i \qquad E(Y_{ij}) = \mu_i$.

Let $\qquad\qquad \mathcal{M}_0 = \{\mu \in \mathbb{R}^N : \mu_{ij} = \mu_0 \in \mathbb{R} \text{ for } j=1..J_i, i=1..I\}$

$\mathcal{M}_0 \subset \mathcal{M}_1$ is a linear subspace in $\mathcal{M}_1$.

Y satisfies the linear hypotheses $\mathcal{M}_0$, if all group means coincide (no difference between groups).

General Linear Models: Variance Decomposition

Theorem: Let Y be a random variable with values in $\mathbb{R}^n$, distributed as $N(0, \sigma^2 I)$. Let $\mathbb{R}^n = \mathcal{Y}_0 \oplus \ldots \oplus \mathcal{Y}_r$ be an orthogonal decomposition. Let $k_i = \dim \mathcal{Y}_i$ and $\pi_i = \pi \mathcal{Y}_i$ be the orthogonal projection on $\mathcal{Y}_i$, $i = 0 \ldots r$. Then

      (i) $\pi_0(Y), \ldots, \pi_r(Y)$ are independent random variables.

      (ii) $| \pi_i(Y) |^2 \sim \sigma^2 \chi^2(k_i)$ for $i = 0 \ldots r$.

Proof: $\rightarrow$ Probability theory, e.g. [Jørgensen 1993, 2.5 Theorem 3].

Note: quotient of any two, adjusted for dimensions, has F distribution, σ^2 cancels out.

Space	dim	contributes	remark
$\mathcal{Y}_0 = \mathcal{M}_0$	$k_0 = \dim(\mathcal{M}_0)$	$\widehat{\mathrm{Var}(\pi_0 Y)}$	
$\mathcal{Y}_1 = \mathcal{M} \ominus \mathcal{M}_0$	k_1 $= \dim(\mathcal{M}) - \dim(\mathcal{M}_0)$	$\widehat{\mathrm{Var}(\pi_1 Y)}$	$\left(^1/_{\dim(\mathcal{Y}_1)} \widehat{\mathrm{Var}(\pi_1 Y)}\right) \big/ \left(^1/_{\dim(\mathcal{Y}_2)} \widehat{\mathrm{Var}(\pi_2 Y)}\right)$ too large? $\curvearrowright \mathcal{M}_0$ rejected.
$\mathcal{Y}_2 = \mathcal{Y} \ominus \mathcal{M}$	$k_2 = N - \dim(\mathcal{M})$	$\widehat{\mathrm{Var}(\pi_2 Y)}$	
$\mathcal{Y}$	N	$\widehat{\mathrm{Var}(Y)}$	$\widehat{\mathrm{Var}(Y)} \overset{?}{=} \sum_h \widehat{\mathrm{Var}(\pi_h Y)}$ (*)

General Linear Models: Remarks

more work needed for other variance structures

variable decomposition only of intermediate use to define model spaces

decision procedure comes free (F-test)
no additional standardization needed

$\pi_{\mathcal{M}_i}(Y)$ is least squares estimator

for orthogonal spaces: decomposition unique
for non-orthogonal spaces: additional normalization needed

O/1 Loss and Quadratic Loss

Classical linear model

　　　　　quadratic loss function for estimation

on fit level

$$L(Y,\widehat{Y}) = \| Y - \widehat{Y} \|^2$$

For classification

　　　　　0/1 loss function

on fit level

$$L(Y,\widehat{Y}) = \delta(\kappa(Y),\widehat{\kappa(Y)}) \qquad 0 \text{ for } \kappa(Y)=\widehat{\kappa(Y)}$$
$$1 \text{ for } \kappa(Y)\neq\widehat{\kappa(Y)}$$

κ class

Quadratic loss makes mathematics easier.

Take-Away Message/ Home Exercise:	
	Optimal estimation for quadratic loss needs not be good for 0/1 classification.

1,2,3

Least squares consistenly optimal only for low dimensions.

E.g. for Gaussian linear mode in dimension ≥ 3:

arithmetic mean $\overline{Y}$ not admissible

Take-Away Message	
	There are dragons out there..

The Bottom Lines

Gaussian linear models **not** assumed to be correct/adequate

But

Try to define a Gaussian linear model

Learn from estimators/tests in these models

Advance to more adequate models

Sample Preparation

see video

<ftp://max.statlab.uni-heidelberg.de/projects/genex/ma.mov>

3. From Experiments to Statistical Models

The Very Simple Model

Spot i spotted with	gene (i)
RNA has concentration/Activity	$K(\text{gene }(i)) = K_\ell(\text{gene }(i))$ for gene(i)
labeled with chosen label ℓ.	
➡ Observed signal intensity	$I(i) = I_\ell(i)$

But unfortunately

    $I(i)$ depends on $K(\text{gene }(i))$,   but relation is unclear

        may depend on spotting

        may depend on marking

        may depend on general intensity

        may be affected by background

Assumption	$f: K(\text{gene }(i)) \mapsto I(i)$	is monotonous increasing.

Additional Factors:

    Spots arranged and hybridization done by carrier (carrier = filter, glass)

        f may depend on carrier

    Local variation on carrier may introduce dependency between spot intensitites

    Experiment may have different labels (different dyes used

    or different runs of experiment)

        f may depend on label

The Very Simple Caveats 1

Spot i spotted with $\qquad$ *gene (i)*

RNA has concentration/Activity $\qquad$ $K(gene\ (i)) = K_\ell(gene\ (i))$ for *gene(i)*

    labeled with chosen label ℓ.

➡ Observed signal intensity $\qquad$ $I(i) = I_\ell(i)$

Unfortunately

signal intensity $I(i)$ per spot is not directly observable

must be reconstructed/estimated from scanned image

↷ image segmentation/ spot finding

↷ local integration
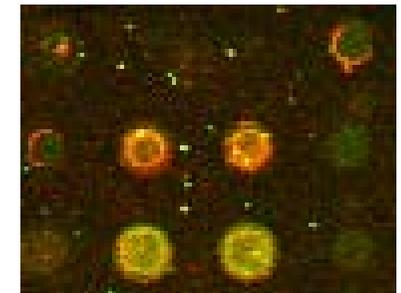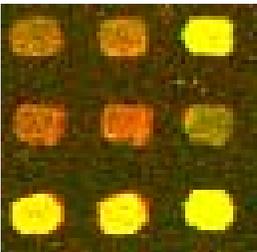
The Very Simple Caveats 2

Unfortunately

concentration to signal intensity transformation unknown

Assumption

$f: K(i) \mapsto I(i)$ is monotonous increasing

Typical assumption:

$f = f_\ell$ for label ℓ

where

$f_\ell = f_0 \, const_\ell$

but with unknown function f_0 and constant $const_\ell$.

The Very Simple Caveats 3

Unfortunately

        Background noise may affect intensity

        not additive

            ⤳ find adequate background correction

Additional sources of information:

        Background intensity as estimated from image processing

        Guide spots

            many repetitions, but intensities at untypical high level

        Missing spots

            repeated measurements, may serve to estimate error distribution at low level.

Suggestions for Notation

Conventional: factors etc. are encoded as indices

$$Y_{i,j,k} = \mu_i + \beta_j + \gamma_{j,k} + err_{\,i,j,k}$$ where Y may be concentration, intensity, …

1. Here: design may involve too many detail
1. Suggestion: encode experimental conditions as attributes - avoid too many indices
                 and use indexing by spot i (= spot position & carrier identification)
                 and channel c (color for glass, or experiment run# for filters)

$$Y_{i,c} = \mu_{gene(i)} + \beta_{carrier(i)} + \gamma_{channel(c),carrier(i)} + err_{\,...}$$
                 for mathematical purists: some notation is overloaded here
                 consider attributes as part of function specification

2. Here: Functional relation/scales are not clear
2. Suggestion: postpone until later - will need clarification
                 use additive notation for now (but don't take it too serious)
                 e.g. $Y_{i,c}$ may be log(Signal) ultimately, model may be multiplicative

Example for Notation: Two sample filter cross over design

filters	A,	B
sample groups	tumor,	normal

Filters are reused, and samples are exchanged

Hybridrization	1	2	3	4
Filter	A	B	A	B
# Filter usage	1	1	2	2
Sample	tumor	normal	normal	tumor

Channel here is #filter usage. Group=Group(channel, carrier) is sample group

After image processing: "Observed" variables are per spot

FG   foreground itensity

BG   background itensity.

with   I background corrected foreground itensity

$$I_{channel,spot} = \mu_{channel,gene,group} + \beta_{carrier,} + err$$

or $$= (\mu_{gene} + S_{gene,group}) + \beta_{carrier} + err$$

Note: Each gene is used in two spots per carrier; so gene = gene(carrier,spot)

Analysis Segmentation

Identify/fix moments in information flow which allow efficient modelling and estimation

variables must be well defined
and observable

complexity should be handable
ideally degrees of freedom low

error components should be decoupled
ideally, error terms stochastically independent

Analysis Segmentation (cont.)

Sample RNA
typically paired: control/tumor

Probe cDNA
typically **many** genes

…

Labeled RNA$_t$

typically one of two#

Spotted Carriers

Hybridized Carriers

Scanned Image

Spot Data

Difference Statistics

Differential Gene List

Analysis Segmentation (cont.)

Suggestion

        handle sample and probe preparation as internal quality control issue

        raw data for microarray analysis:
            scanned image data
                = x, y, channel, intensity

Pragmatic step: separate image analysis from statistical evaluation

        initial data for statistical analysis
            spot data
                = col, row, gene, channel background, channel foreground

Segmentation adds Items for ToDo List:
        specify gene detection based on spot data
        tune spot data preprocessing for best detection

Differential Gene Expression Primer

Model:

$$I(i) = f_i(K(i)) + err_i$$

where f_i is monotonous in $K(i)$.

f_i essentially depends only on attributes of i other than $gene(i)$.

"essentially": f_i may depend for example on the general concentration level

$$\overline{K} = \text{average}_{i': \, gene(i')=gene(i)} \, K(i)$$

Idea:

$$K(i) = K(gene(i), group(i))$$

$$K(tumor) > K(control)$$

Ideally, for example (fold change, multiplicative):

$$K(i) = K_0(i) * f_{group(i)} * err_i$$

$$I(i) = f_{label(i)} (K(i)) \, err'_i \qquad \text{where } f_{label(i)} = const_{label(i))} * f_0$$

$f_{group(i)}$ fold factor.

Unfortunately, f unknown.

Differential Gene Expression Primer (cont.)

Model:

$$I(i) = f_i(K(i)) + err_i$$
where f_i is monotonous in $K(i)$.

Idea:

$$K(i) = K(gene(i), group(i))$$
$$K(tumor) > K(control)$$

Unfortunately, f unknown.
Fortunately, f monotonous.

Suggestion (raw)

Use

$$Rk(i) = Rank_{i': group(i')=group(i)} I(i)$$

Recipe (two groups)

For spots i,j with gene(i) = gene(j), group(i) ≠ group(j)
call j differentially expressed if **Rk(j) >> Rk(i)**.

Differential Gene Expression Primer (cont.)

Summary table counts by within group ranks Rk(i)

Rel. Counts	tumor lower 25%	tumor 25%-50%	tumor 50%-75%	tumor upper 25%
control lower 25%	(ignore)		(switched up)	(definitively switched up)
control 25%-50%			(maybe switched up)	(switched up)
control 50%-75%	(switched down)	(maybe switched down)		
control upper 25%	(definitively switched down)	(switched down)		(ignore)

Detailed report

List genes by Rk(tumor) - Rk(control)

        ignore indifferent part

Differential for Expressions Primer: 2 Sample Filter Swap

How to combine observations the rough way.

1.   Use guide spots & missing to correct for background.
             This may be done per filtering to adjust for overall filter / gene effects.

2. Remove guide spot & missing data.

3. Rank data per filter (this depends on the chosen selection of genes???)

4. Combine repeated measurements for sequences, e.g. by taking mean ranks (or mid ranks).

5. Find differences within paired samples (same filter, tumor against normal) for the scores from 4.

6. Combine scores from swap over eg. by taking means between filters.

7. Use extreme high/low scores from 6 for identification of a differential step change For limits, use
                quadratic distribution product or gaussian approximation

8. double-check.

Differential Gene Expression: First Refinement

Primer model does not use background information
Quartiles are arbitrary bounds

Modifications:

Use background to define detection limit.

Call intensity *below detection limit* if $I(i) < BG_\alpha$ e.g $\alpha = 0.1\%$
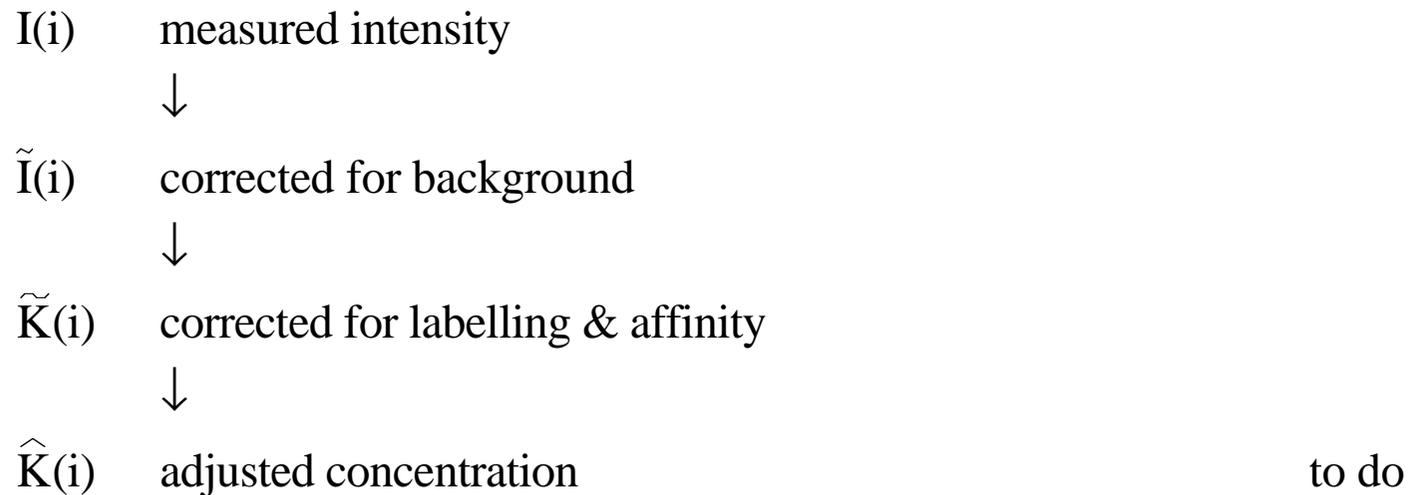   BG   background

Call intensity *above saturation limit* if $I(i) > FG_\alpha$ e.g $\alpha = 99.9\%$
   FG   foreground

Call difference *ignorable* if proportion p is within difference range   e.g. p = 10%
    or difference below detection limit

Differential Gene Expression: Critical Points

- Background may have local variation
- Determination of critical levels depend on experimentally controlled choice of genes (not a random sample, but controlled by design)
- Rank depends on ensemble
- Pure intensity effects are not detected, if rank is not changed.

Ideally.

$I(i)$ measured intensity

       $\downarrow$

$\tilde{I}(i)$ corrected for background

       $\downarrow$

$\tilde{\tilde{K}}(i)$ corrected for labelling & affinity

       $\downarrow$

$\hat{K}(i)$ adjusted concentration   to do

4. Towards a Data Analysis Protocol

Targets for Data Analysis

Quality Control

        per chip

        per batch

        par experiment series

                classical quality control

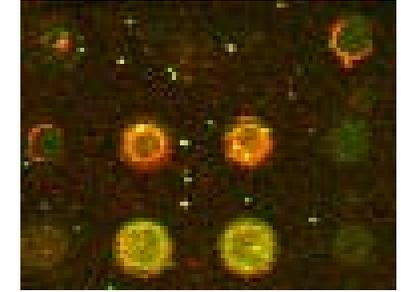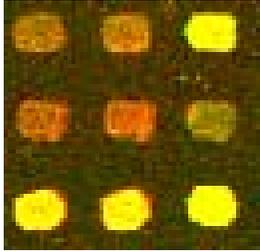Differential Gene Expression Detection

                classification?

                variable clustering?

Individual Diagnostics

                classification?

                test?

Step by Step - Step 1: Image Preprocessing

Segmentation

Local Integegration

... still to do: local model

After step 1: scan image reduced to 4 items per spot

	red	green
foreground	$Y_{fg,red}$	$Y_{fg,green}$
background	$Y_{bg,red}$	$Y_{bg,green}$

Step by Step - Step 1a: Check

Quality Control

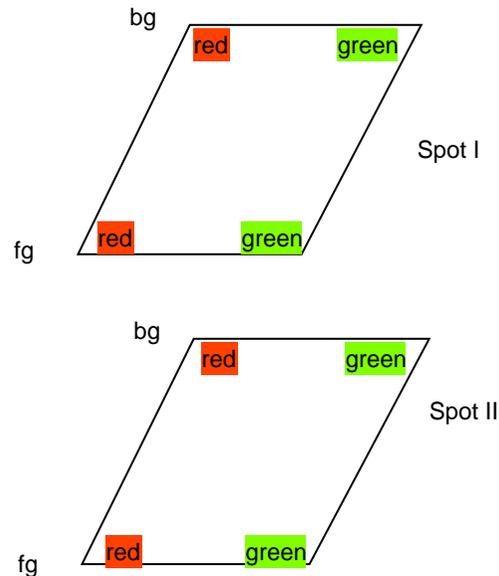   per chip

   per batch

   par experiment series

      classical quality contro

....Visualisation

Step by Step - Step 2: Identify Differentially Expressed Genes

e.g. two spots per gene.

per spot four items

Find decision rule to "identify" a differentially expressed gene.

If #spots large & Gaussian assumptions: t-Test

Find an adequate variation

Note: high number of genes may provide lateral information to model structural effects

Step by Step - Step 2a: Check Gene Replicates

e.g. two spots per gene.

after background correction

$$\text{check } | Y_{spot1} - Y_{spot2} |$$

correct for scale

correct for variance

modified version of

$$| Rk\ Y_{spot1} - Rk\ Y_{spot2} | / | Rk\ Y_{spot1} + Rk\ Y_{spot2} |$$

scaled for variance from uniform model.

Step by Step - Step 3: Classify Gene Families

-> Bertin

Step by Step - Step 3: Case Diagnostics