



University of
Zurich^{UZH}

Institute of Molecular Life Sciences

limma, Affymetrix, RMA, Independent filtering

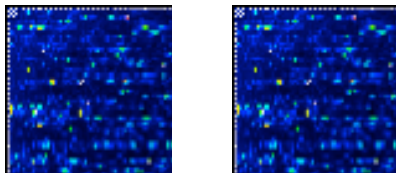
- Review of last week: moderation
- design matrices + contrast matrix
- limma mathematical theory
- Affymetrix arrays + RMA

Mark D. Robinson, Statistical Genomics, IMLS

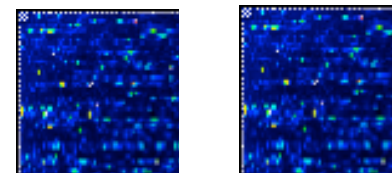


A very common experiment

Mutant x 2



WT x 2



Gene X



Which genes are differentially expressed?

$n_1 = n_2 = 2$ microarrays

~30,000 features (e.g., genes) measured



Ordinary t-tests (1-colour)

$$t_g = \frac{\bar{y}_{\text{mu}} - \bar{y}_{\text{wt}}}{s_g c}$$

gives very high false discovery rates

$$c = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Residual df = 2



t-tests with common variance

$$t_{g,\text{pooled}} = \frac{\bar{y}_{\text{mu}} - \bar{y}_{\text{wt}}}{s_0 c}$$

with residual standard deviation s_0 pooled
across genes

More stable, but ignores gene-specific variability

$$c = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$



Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_0^2 d_0 + s_g^2 d_g}{d_0 + d_g}$$

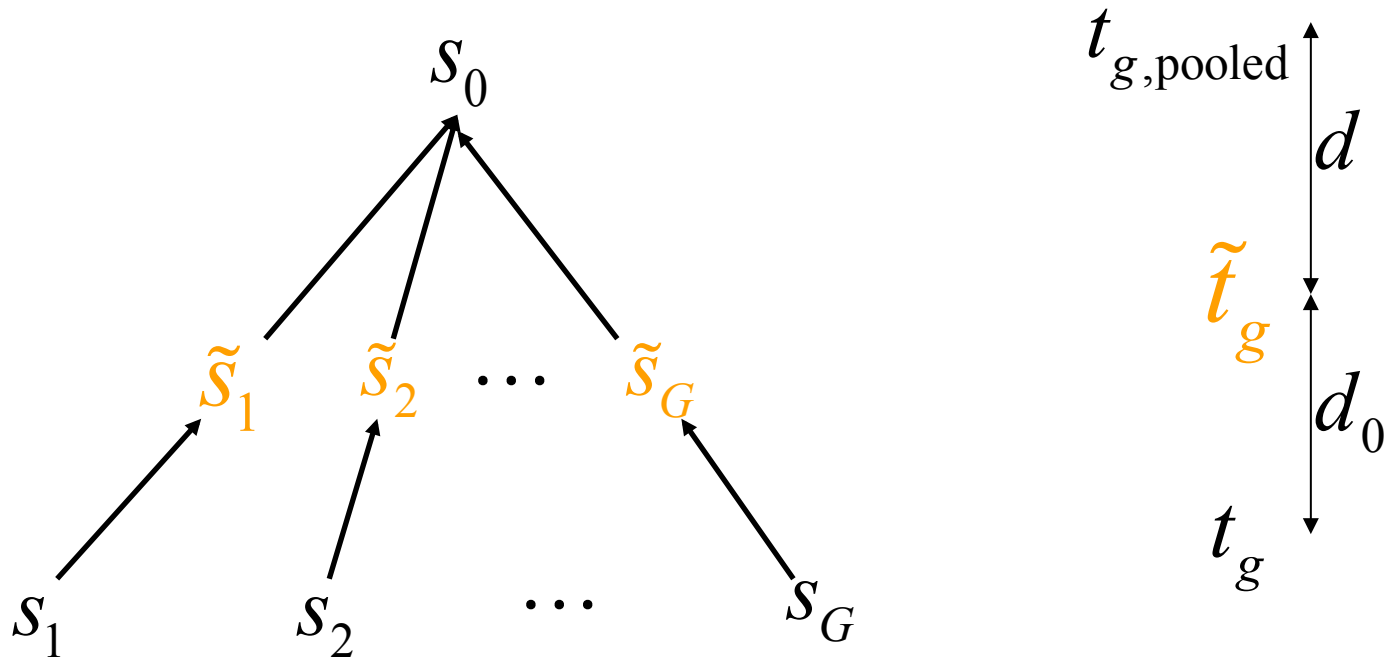
Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

Baldi & Long 2001, Wright & Simon 2003, Smyth 2004



Shrinkage of standard deviations



The **data decides** whether \tilde{t}_g should be closer to $t_{g,\text{pooled}}$ or to t_g



What layers to add today

- Where does the moderated variance come from?
- Why the degrees of freedom add: $d_0 + d$
- A bit about how to get the hyperparameters (d_0 and s_0)
- Design matrices + contrast matrices in practice



**University of
Zurich^{UZH}**

Institute of Molecular Life Sciences

Exercise:

where does the t-distribution come from?

15 minutes: in groups or individually, use the resources provided and/or search the web to explain the fundamental theory of a t-test/t-distribution.



Unexpected mathematics: Why do degrees of freedom add?

The construction of the classical t-statistic:

$$Z = (\bar{X}_n - \mu) \frac{\sqrt{n}}{\sigma}$$
$$V = (n - 1) \frac{S_n^2}{\sigma^2}$$
$$T \equiv \frac{Z}{\sqrt{V/\nu}} = (\bar{X}_n - \mu) \frac{\sqrt{n}}{S_n},$$

Stated another way → Exercise (optional): what are a, b above?

If T is distributed as $(a/b)^{1/2} Z/U$ where $Z \sim N(0, 1)$ and $U \sim \chi_\nu$, then T has density function

$$p(t) = \frac{a^{\nu/2} b^{1/2}}{B(1/2, \nu/2) (a + bt^2)^{1/2 + \nu/2}}$$



Review: Bayes theorem, conditional probability

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}.$$

$$\Pr(A) = \sum_n \Pr(A | B_n) \Pr(B_n)$$

Disease Test	+ (1%)	- (99%)
"+"	85%	1%
"_"	15%	99%

- "+" – event that you test positive
- event that you have the disease
- test negative
- do not have disease

Sketch:

$$P(+ | "+") = \frac{P("+ | +)P(+)}{P("+")} = \frac{P("+ | +)P(+)}{P("+ | +)P(+)+P("+ | -)P(-)}$$



Exercise: Derive the posterior

Data

$$s_g^2 \sim \sigma_g^2 \frac{\chi_{d_g}^2}{d_g}$$

Prior

$$\frac{1}{\sigma_g^2} \sim s_0^2 \frac{\chi_{d_0}^2}{d_0}$$

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta}$$

Posterior

$$E\left(\frac{1}{\sigma_g^2} \mid s_g^2\right) = \frac{d_0 + d_g}{s_0^2 d_0 + s_g^2 d_g}$$

Optional exercise

Sketch: i) Let $x=s^2$, $\theta=\sigma^{-2}$; ii) Using the functional form of chi-squared distribution, calculate only the numerator (since denominator does not contain θ); iii) collect terms and see if you can identify the distribution and the parameters of it; iv) What is the mean of this distribution?



Linear Models

- In general, need to specify:
 - Dependent variable
 - Explanatory variables (experimental design, covariates, etc.)
- More generally:

$$y = X\alpha + \epsilon$$

Diagram illustrating the components of the linear model equation $y = X\alpha + \epsilon$:

- y : vector of observed data (indicated by a blue arrow)
- X : design matrix (indicated by a red arrow)
- α : Vector of parameters to estimate (indicated by an orange arrow)



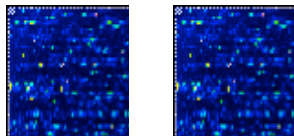
Linear Models for microarrays

- Combined estimation of precision (moderated variance)
- Extensible to arbitrarily complicated experiments (multiple groups, factorial designs, time courses, paired designs, etc.)
 - NB: only special cases of mixed models are covered
- **Design matrix**: specifies experimental condition of each sample
- **Contrast matrix**: specifies which comparisons are of interest

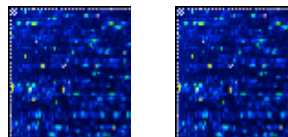


Analysis of Variance → Linear model

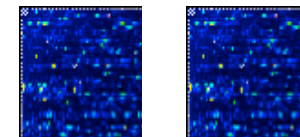
WT x 2



Cond A x 2



Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

α_1 = wt log-expression

α_2 = Cond A - wt

α_3 = Cond B - wt

$$E[y_1] = E[y_2] = \alpha_1$$

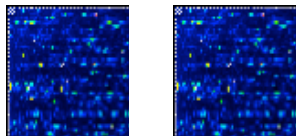
$$E[y_3] = E[y_4] = \alpha_1 + \alpha_2$$

$$E[y_5] = E[y_6] = \alpha_1 + \alpha_3$$

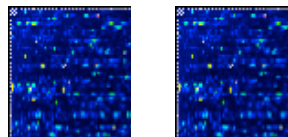


Analysis of Variance → Linear model, alternative parameterization

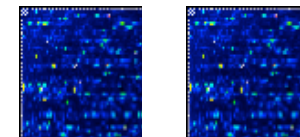
WT x 2



Cond A x 2



Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$\alpha_1 = \text{wt log-expression}$
 $\alpha_2 = \text{Cond A log-expression}$
 $\alpha_3 = \text{Cond B log-expression}$

$$E[y_1] = E[y_2] = \alpha_1$$

$$E[y_3] = E[y_4] = \alpha_2$$

$$E[y_5] = E[y_6] = \alpha_3$$



Linear Model Estimates – `lmFit()`

Obtain a linear model for each gene g

$$E(\underline{y}_g) = X\alpha_g$$
$$\text{var}(\underline{y}_g) = W_g^{-1}\sigma_g^2$$

Estimate:

coefficients

$$\hat{\alpha}_{gj}$$

standard deviations

$$s_g$$

standard errors

$$\text{se}(\hat{\beta}_{gj})^2 = c_{gj}s_g^2$$



An example use of design and contrast matrices

design matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$E[y_1]=E[y_2]=\alpha_1$
 $E[y_3]=E[y_4]=\alpha_2$
 $E[y_5]=E[y_6]=\alpha_3$

contrast matrix

$$\beta = C\alpha = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_2 \end{bmatrix}$$



Contrasts -- `contrasts.fit()`

A *contrast* is any linear combination of the coefficients α_j which we want to test equal to zero.

Define contrasts

$$\beta_g = C^T \alpha_g$$

where C is the *contrast matrix*.

Want to test

$$H_0 : \beta_{gj} = 0$$

vs

$$H_a : \beta_{gj} \neq 0$$



Limma / Analysis of Variance

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

$$F = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}} = \frac{SS_{\text{Treatments}}/(I-1)}{SS_{\text{Error}}/(n_T - I)}$$

The moderated t -statistics also lead naturally to moderated F -statistics which can be used to test hypotheses about any set of contrasts simultaneously. Appropriate quadratic forms of moderated t -statistics follow F -distributions just as do quadratic forms of ordinary t -statistics. Suppose that we wish to test all contrasts for a given gene equal to zero, i.e., $H_0 : \beta_g = 0$. The correlation matrix of $\hat{\beta}_g$ is $R_g = U_g^{-1}C^TV_gCU_g^{-1}$ where U_g is the diagonal matrix with unscaled standard deviations $(v_{gj})^{1/2}$ on the diagonal. Let r be the column rank of C . Let Q_g be such that $Q_g^TR_gQ_g = I_r$ and let $\mathbf{q}_g = Q_g^T\mathbf{t}_g$. Then

$$F_g = \mathbf{q}_g^T\mathbf{q}_g/r = \mathbf{t}_g^TQ_gQ_g^T\mathbf{t}_g/r \sim F_{r,d_0+d_g}$$



Unexpected mathematics: Why do degrees of freedom add?

The construction of the classical t-statistic:

$$Z = (\bar{X}_n - \mu) \frac{\sqrt{n}}{\sigma}$$
$$V = (n - 1) \frac{S_n^2}{\sigma^2}$$
$$T \equiv \frac{Z}{\sqrt{V/\nu}} = (\bar{X}_n - \mu) \frac{\sqrt{n}}{S_n},$$

Stated another way → Exercise (optional): what are a, b above?

If T is distributed as $(a/b)^{1/2} Z/U$ where $Z \sim N(0, 1)$ and $U \sim \chi_\nu$, then T has density function

$$p(t) = \frac{a^{\nu/2} b^{1/2}}{B(1/2, \nu/2) (a + bt^2)^{1/2 + \nu/2}}$$



Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 \mid \beta = 0) = \int p(\hat{\beta} \mid \sigma^{-2}, \beta = 0) p(s^2 \mid \sigma^{-2}) p(\sigma^{-2}) d(\sigma^{-2})$$

The integrand is

$$\begin{aligned} & \frac{1}{(2\pi v \sigma^2)^{1/2}} \exp\left(-\frac{\hat{\beta}^2}{2v\sigma^2}\right) \\ & \times \left(\frac{d}{2\sigma^2}\right)^{d/2} \frac{s^{2(d/2-1)}}{\Gamma(d/2)} \exp\left(-\frac{ds^2}{2\sigma^2}\right) \\ & \times \left(\frac{d_0 s_0^2}{2}\right)^{d_0/2} \frac{\sigma^{-2(d_0/2-1)}}{\Gamma(d_0/2)} \exp\left(-\sigma^{-2} \frac{d_0 s_0^2}{2}\right) \\ & = \frac{(d_0 s_0^2/2)^{d_0/2} (d/2)^{d/2} s^{2(d/2-1)}}{(2\pi v)^{1/2} \Gamma(d_0/2) \Gamma(d/2)} \\ & \quad \sigma^{-2(1/2+d_0/2+d/2-1)} \exp\left\{-\sigma^{-2} \left(\frac{\hat{\beta}^2}{2v} + \frac{ds^2}{2} + \frac{d_0 s_0^2}{2}\right)\right\} \end{aligned}$$



Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 | \beta = 0) = \int p(\hat{\beta} | \sigma^{-2}, \beta = 0) p(s^2 | \sigma^{-2}) p(\sigma^{-2}) d(\sigma^{-2})$$

$$= \frac{(d_0 s_0^2 / 2)^{d_0/2} (d/2)^{d/2} s^{2(d/2-1)}}{(2\pi v)^{1/2} \Gamma(d_0/2) \Gamma(d/2)}$$

$$\sigma^{-2(1/2+d_0/2+d/2-1)} \exp \left\{ -\sigma^{-2} \left(\frac{\hat{\beta}^2}{2v} + \frac{ds^2}{2} + \frac{d_0 s_0^2}{2} \right) \right\}$$



σ^{-2} is chi-squared (or gamma)

$$f(x; k) = \begin{cases} \frac{x^{(k/2)-1} e^{-x/2}}{2^{k/2} \Gamma(\frac{k}{2})}, & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

http://en.wikipedia.org/wiki/Chi-squared_distribution



Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 \mid \beta = 0) = \int p(\hat{\beta} \mid \sigma^{-2}, \beta = 0) p(s^2 \mid \sigma^{-2}) p(\sigma^{-2}) d(\sigma^{-2})$$

$$\begin{aligned} p(\hat{\beta}, s^2 \mid \beta = 0) \\ = \frac{(1/2v)^{1/2} (d_0 s_0^2/2)^{d_0/2} (d/2)^{d/2} s^{2(d/2-1)}}{D(1/2, d_0/2, d/2)} \left(\frac{\hat{\beta}^2/v + d_0 s_0^2 + d s^2}{2} \right)^{-(1+d_0+d)/2} \end{aligned}$$



Unexpected mathematics: Why do degrees of freedom add?

$$p(\hat{\beta}, s^2 | \beta = 0) \\ = \frac{(1/2v)^{1/2} (d_0 s_0^2/2)^{d_0/2} (d/2)^{d/2} s^{2(d/2-1)}}{D(1/2, d_0/2, d/2)} \left(\frac{\hat{\beta}^2/v + d_0 s_0^2 + d s^2}{2} \right)^{-(1+d_0+d)/2}$$

The null joint distribution of \tilde{t} and s^2 is

$$p(\tilde{t}, s^2 | \beta = 0) = \tilde{s} v^{1/2} p(\hat{\beta}, s^2 | \beta = 0)$$

http://en.wikipedia.org/wiki/Random_variable#Distribution_functions_of_random_variables

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Unexpected mathematics: Why do degrees of freedom add?

If T is distributed as $(a/b)^{1/2}Z/U$ where $Z \sim N(0, 1)$ and $U \sim \chi_\nu$, then T has density function

$$p(t) = \frac{a^{\nu/2} b^{1/2}}{B(1/2, \nu/2)(a + bt^2)^{1/2 + \nu/2}}$$

$$p(\tilde{t}, s^2 \mid \beta = 0) = \frac{(d_0 s_0^2)^{d_0/2} d^{d/2} s^{2(d/2-1)}}{B(d/2, d_0/2)(d_0 s_0^2 + ds^2)^{d_0/2 + d/2}} \times \frac{(d_0 + d)^{-1/2}}{B(1/2, d_0/2 + d/2)} \left(1 + \frac{\tilde{t}^2}{d_0 + d}\right)^{-(1+d_0+d)/2}$$

This shows that \tilde{t} and s^2 are independent with

$$s^2 \sim s_0^2 F_{d, d_0}$$

and

$$\tilde{t} \mid \beta = 0 \sim t_{d_0+d}.$$



Aside: Marginal Distributions to calculate

Fun fact: Under usual likelihood model, s_g is independent of the estimated coefficients.

Under the hierarchical model, s_g is independent of the moderated t-statistics instead

$$s_g^2 \sim s_0^2 F_{d, d_0}$$

Thus, the set of s_g can be used to estimate d_0 and s_0



**University of
Zurich^{UZH}**

Institute of Molecular Life Sciences

Affymetrix + RMA + IRLS



Affymetrix probe design

Early platforms (11 or 20 probes in a set), 25bp probes, 3' biased

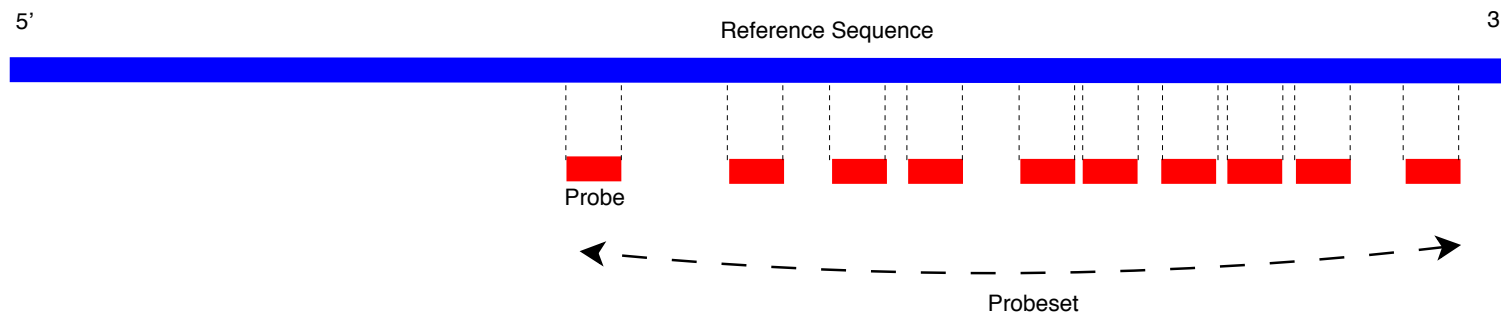


Figure 1.1: Multiple probes interrogating the sequence for a particular gene make up probesets.



Figure 1.2: Pefect Match and Mismatch Probes.



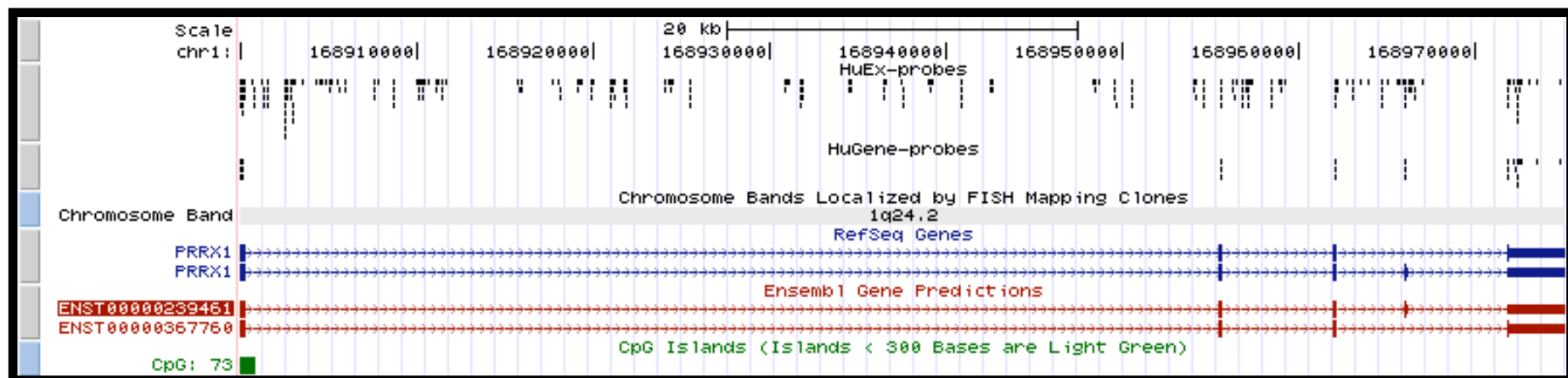
Latest Affymetrix design: “whole transcript” arrays

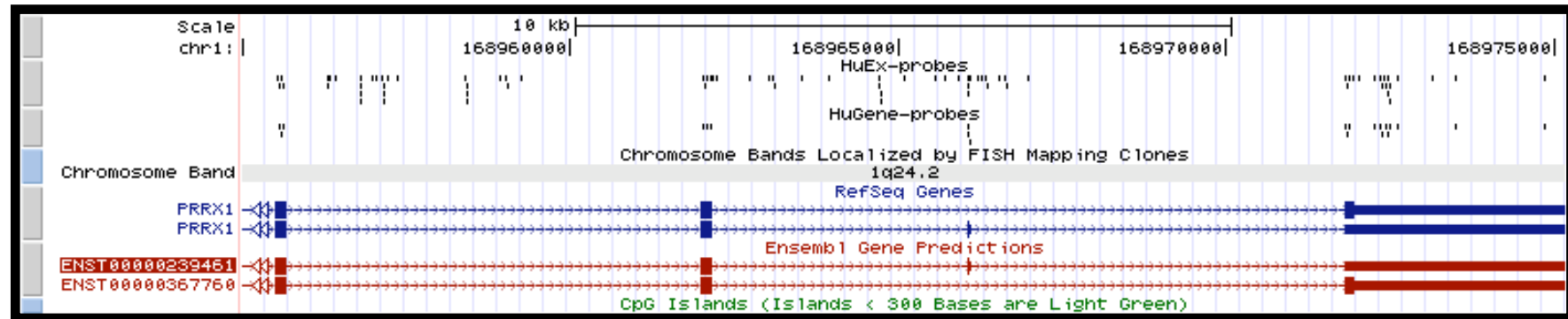
Still 25 base pair probes, multiple probes per transcript (“probesets”)

No more mismatch probes.

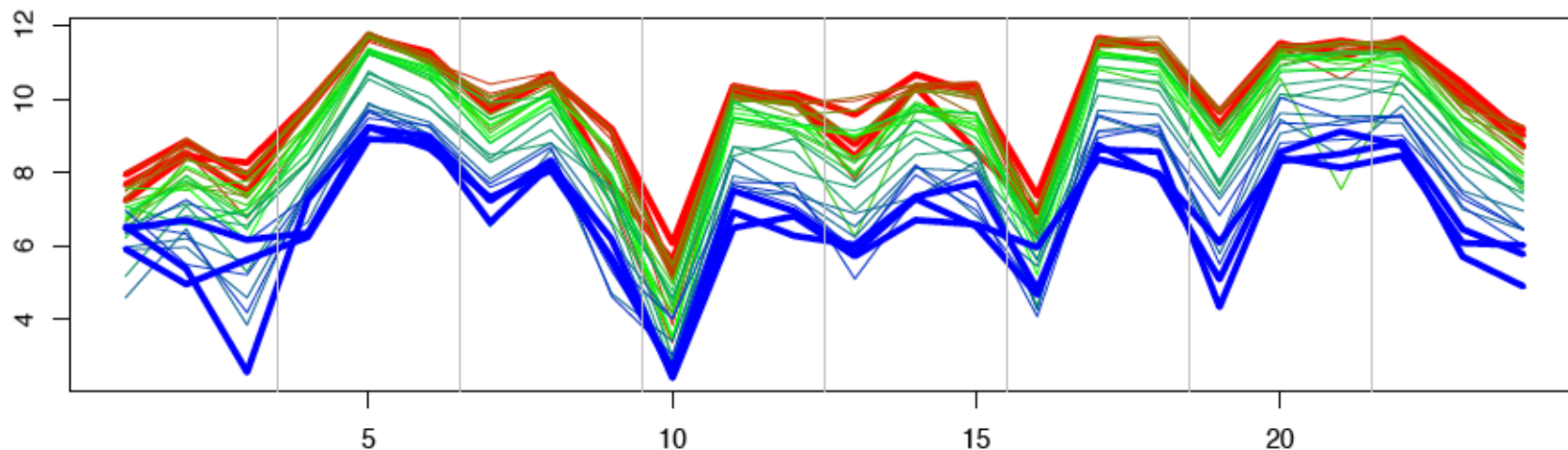
Reference Sequence

- HuExon: *Human Exon 1.0 ST* (~40 probes per gene, 4 probes per “exon”, annotated and predicted transcripts)
- HuGene: *Human Gene 1.0 ST* (~25 probes per gene, annotated genes only)
- [NEW in 2013](#): HTA (Human Transcriptome Array): updated content + junction probes





HuGene data [red=heart,blue=brain,mixtures] 10 ENSG00000116132



- Data for one gene that is differentially expressed between heart (red is 100% heart) and brain (blue is 100% brain).
- 11 mixtures x 3 replicates = 33 samples (33 lines)
- Note the parallelism: probes have different **affinities**



“Summarization”: Going from probesets to summarized expression level

MAS 4.0

$$AvDiff = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j)$$

MAS 5.0

$$CT_j = \begin{cases} MM_j, & \text{if } MM_j < PM_j \\ \text{less than } PM_j, & \text{if } MM_j \geq PM_j \end{cases}$$

$$signal = TukeyBiweight\{\log(PM_j - CT_j)\}$$

dChip (MBEI)

$$PM_{ij} - MM_{ij} = \theta_i \cdot \phi_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

θ_i expression index
 ϕ_j probe-specific affinity
 ε_{ij} noise component

RMA, GCRMA



Robust multichip analysis (RMA)

Exploration, normalization, and summaries of high density oligonucleotide array probe level data

RAFAEL A. IRIZARRY*

Department of Biostatistics, Johns Hopkins University, Baltimore MD 21205, USA
rafa@jhu.edu

BRIDGET HOBBS

Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia

FRANCOIS COLLIN

Gene Logic Inc., Berkeley, CA, USA

YASMIN D. BEAZER-BARCLAY, KRISTEN J. ANTONELLIS, UWE SCHERF

Gene Logic Inc., Gaithersburg, MD, USA

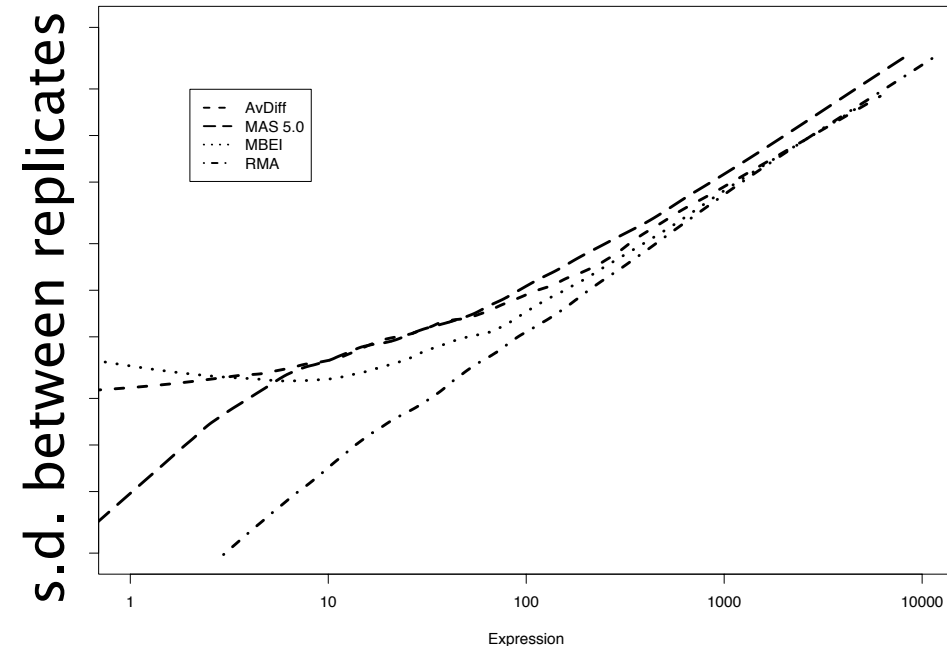
TERENCE P. SPEED

Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia. Department of Statistics, University of California at Berkeley

Biostatistics 2003

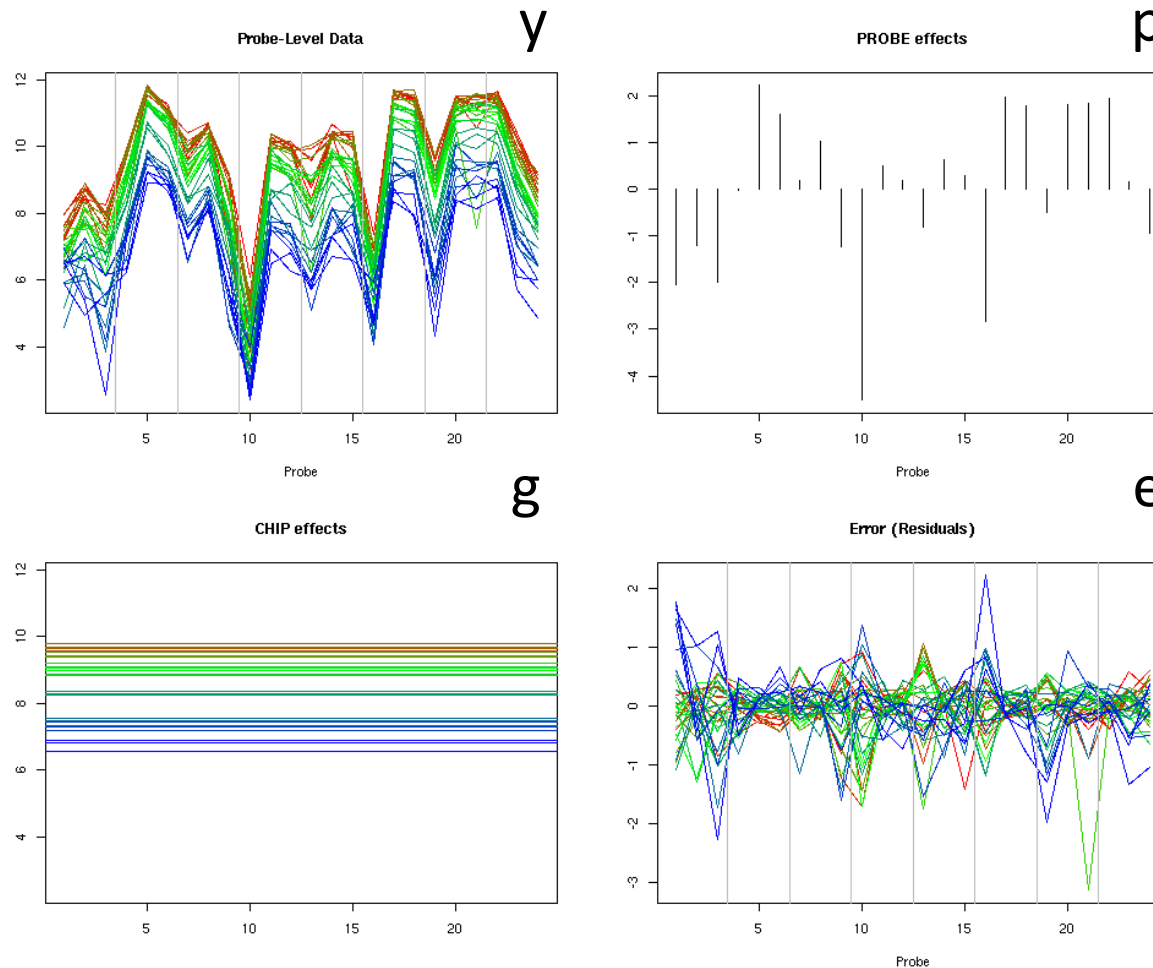
- Encompasses 3 steps
- background correction
 - normalization
 - probe level model fit (“summarization”)

b) Standard deviation vs. average expression





Linear model decomposes the probe-level data into **PROBE** effects and **CHIP** effects



Linear model:

$$y_{ik} = g_i + p_k + e_{ik}$$

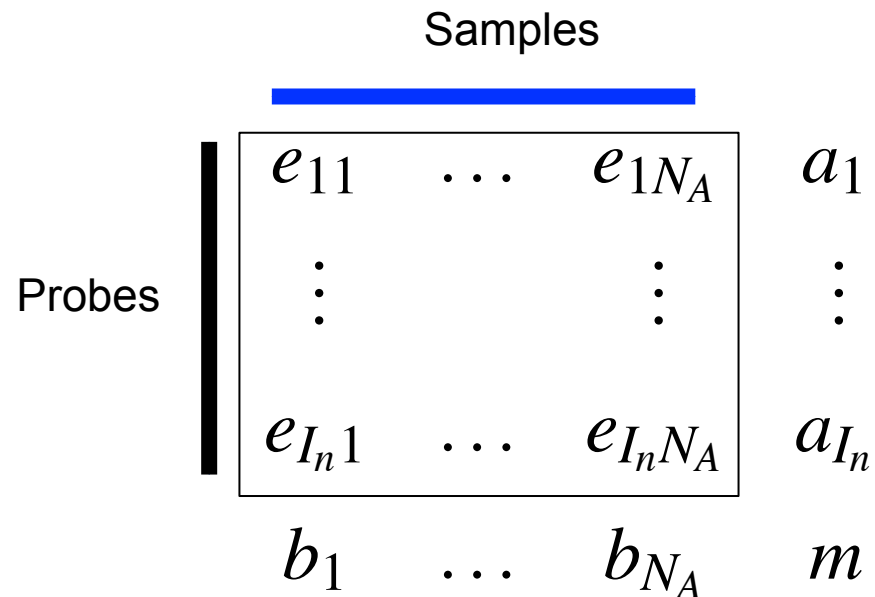
Robust Multichip Analysis (RMA) uses this model.

Irizarry et al. 2003, Biostatistics

Parameters are estimated **robustly**, meaning a small number of outliers have minimal effect



Fitting the model – median polish



```

pe <- rnorm(11)
ce <- rnorm(8)+8
z <- outer(pe,ce,"+") +
      rnorm(length(pe)*length(ce),sd=.5)
e <- z
m <- a <- b <- 0
niter <- 3

for(i in 1:niter) {
  rm <- rowMedians(e) # calc row medians
  e <- sweep(e,1,rm)  # subtract row medians
  a <- a + rm         # add row medians to a
  mb <- median(b)
  b <- b-mb
  m <- m+mb

  cm <- colMedians(e) # calc col medians
  e <- sweep(e,2,cm)  # subtract col medians
  b <- b + cm         # add col medians to b
  ma <- median(a)
  a <- a-ma
  m <- m+ma
}

# a - "probe effects"
# m+b - "chip effects"

```



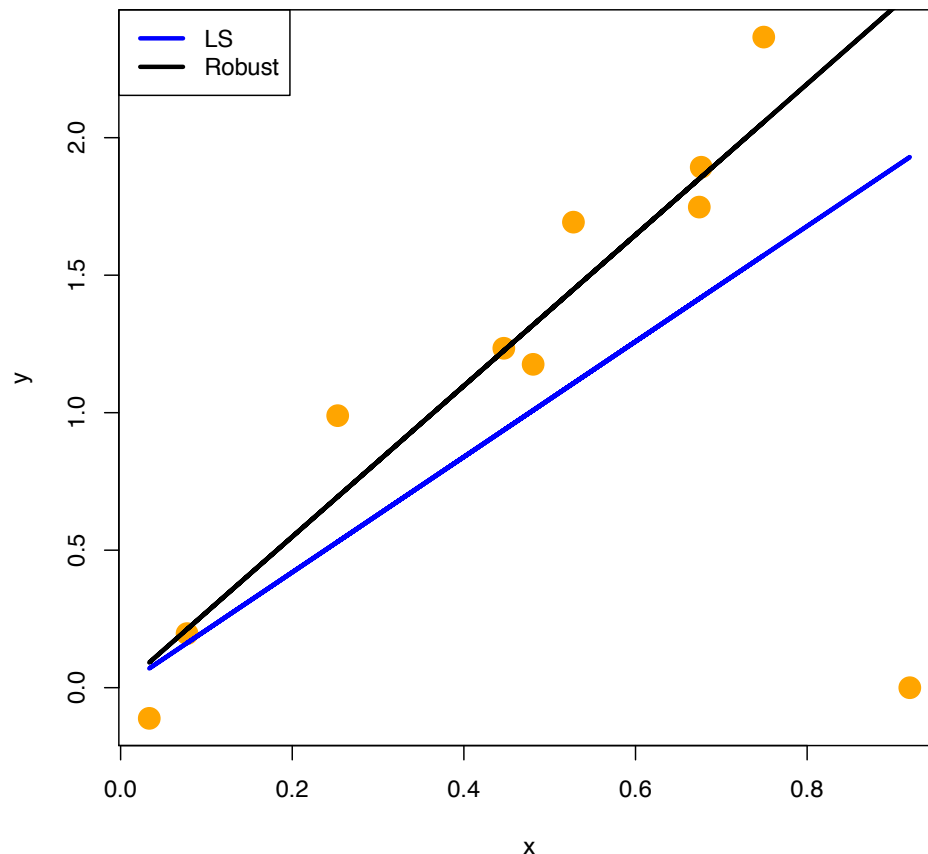
Robust regression – motivating example

```
library(MASS)

n <- 10
x <- runif(n)
y <- 3*x + rnorm(n, sd=.2)
y[which.max(y)] <- 0 # add in outlier

f <- lm(y~0+x)
fr <- rlm(y~0+x)

plot(x,y,pch=19,col="orange",cex=2)
lines(x,predict(fr),lwd=3)
lines(x,predict(f),lwd=3,col="blue")
legend("topleft",c("LS", "Robust"),
      lwd=3,lty=1,col=c("blue", "black"))
```



OLS = ordinary least squares

The OLS estimator is ... optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated ... OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances.

Has good properties, when the data is “nice”.

Replace:

$$\arg \min_{\beta} \sum_{i=1}^n (y_i - f_i(\beta))^2$$

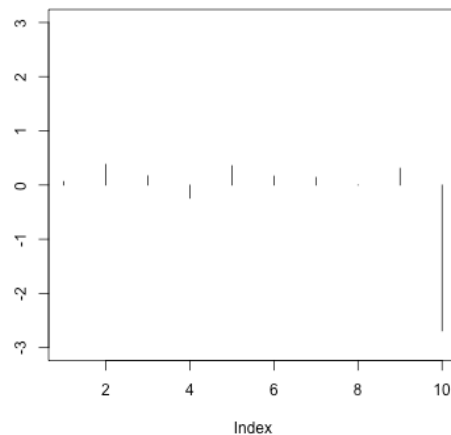
with:

$$\arg \min_{\beta} \sum_{i=1}^n w_i(\beta) (y_i - f_i(\beta))^2$$

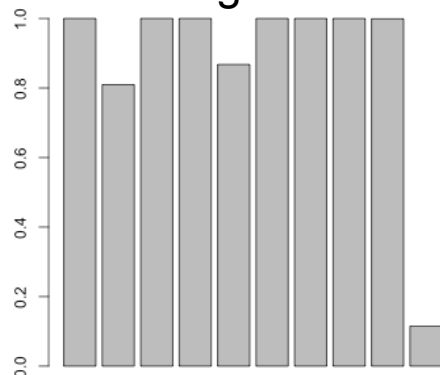


Robust regression – mechanics of iteratively reweighted least squares

Residuals



Weights



Sketch of IRLS:

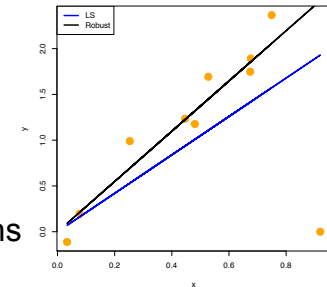
Calculate initial estimates of parameters

Repeat until very little change:

Calculate residuals

Using standardized residuals, weight observations

Re-estimate parameters



```
# this construction only works for the
# 1-parameter no-intercept linear model
tukey <- function(r,k=1.345) {
  abs(r) < k + k/abs(r)*(abs(r)>k)
}
```

```
w <- 1
niter <- 2
b <- sum(w*y*x)/sum(w*x^2)
```

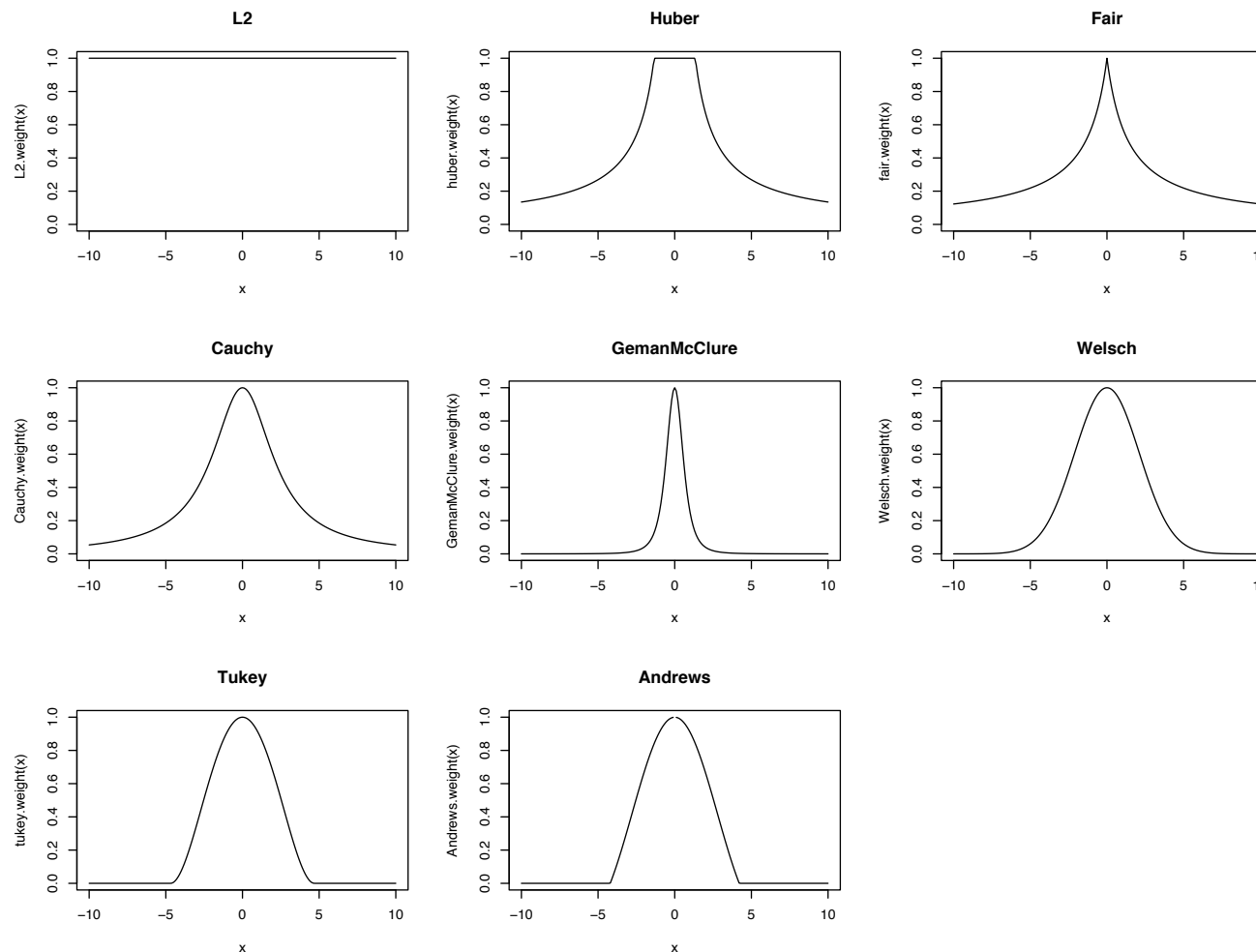
```
for(i in 1:niter) {
  r <- y-b*x
  w <- tukey( r/mad(r) )
  b <- sum(w*y*x)/sum(w*x^2)
}
```

← mad = median
absolute deviation

```
par(mfrow=c(2,1))
plot(r,type="h",ylim=c(-3,3))
barplot(w)
```



More details – weight functions (as function of standardized residuals)





More details – weight functions (of normalized residuals) Concept: influence / bounded influence

The estimated standard error for our estimators is thus given by

$$\text{SE}(\hat{\beta}_j^{(n)}) = \frac{1}{\sqrt{I_n}} \sqrt{\frac{\sum_{i=1}^{I_n} \psi\left(\frac{\log_2(y_{ij}^{(n)}) - \hat{\beta}_j^{(n)}}{s}\right)^2 / I_n}{\left(\sum_{i=1}^{I_n} \psi'\left(\frac{\log_2(y_{ij}^{(n)}) - \hat{\beta}_j^{(n)}}{s}\right) / I_n\right)^2}}.$$

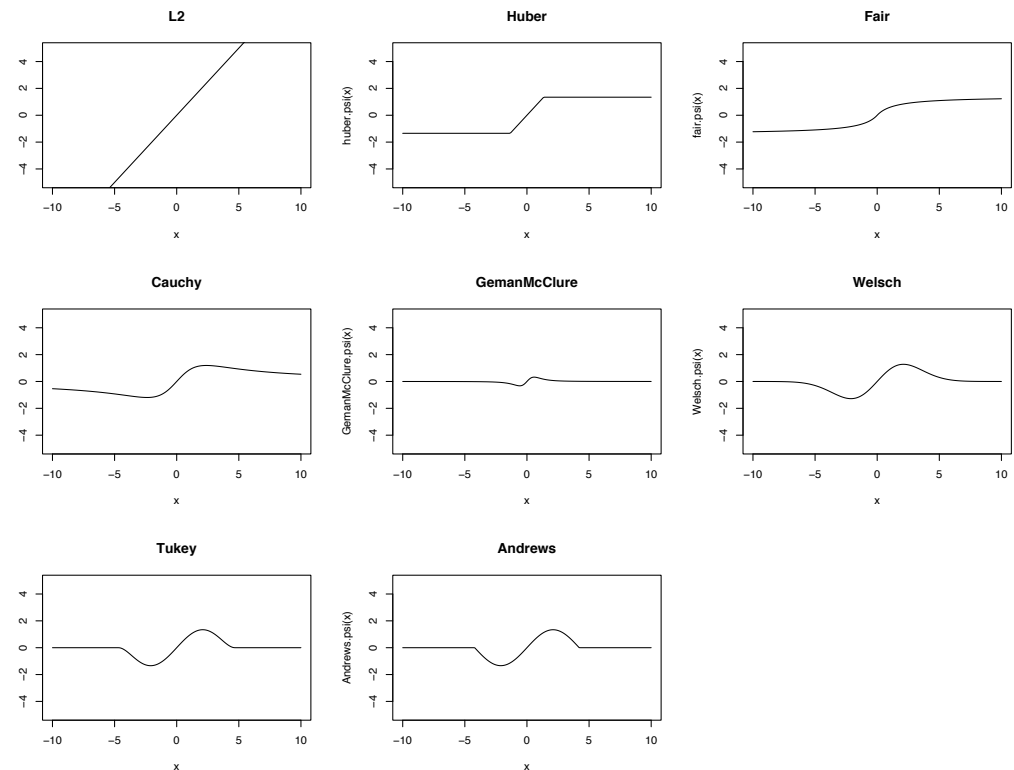
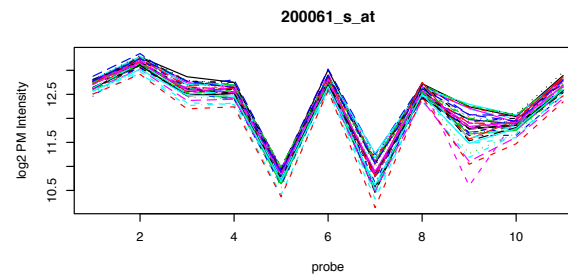
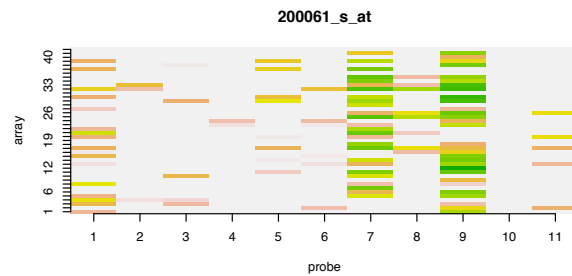


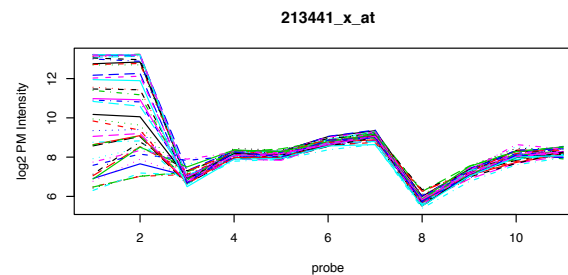
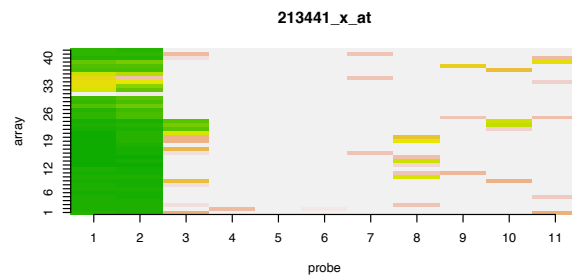
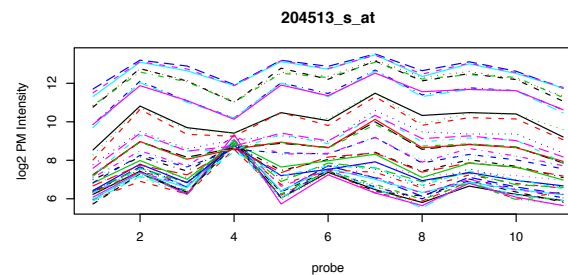
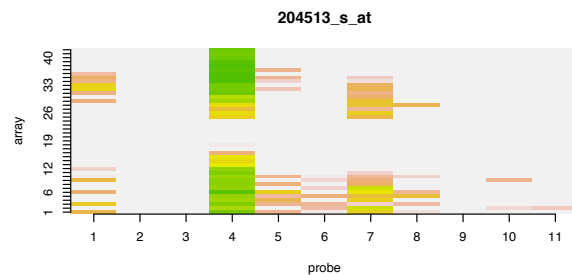
Figure 4.2: The ψ functions for some common M-estimators.



Robust regression leads to various quality assessment metrics

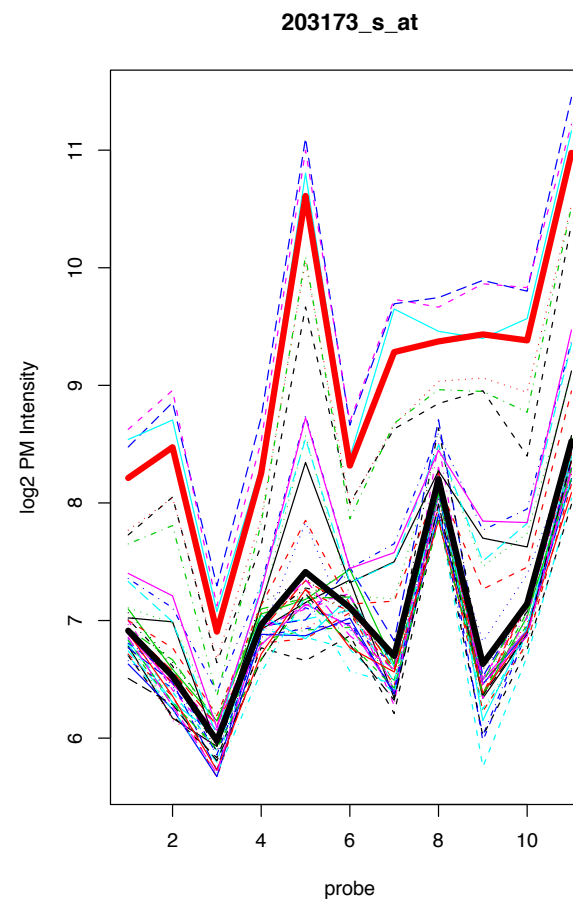
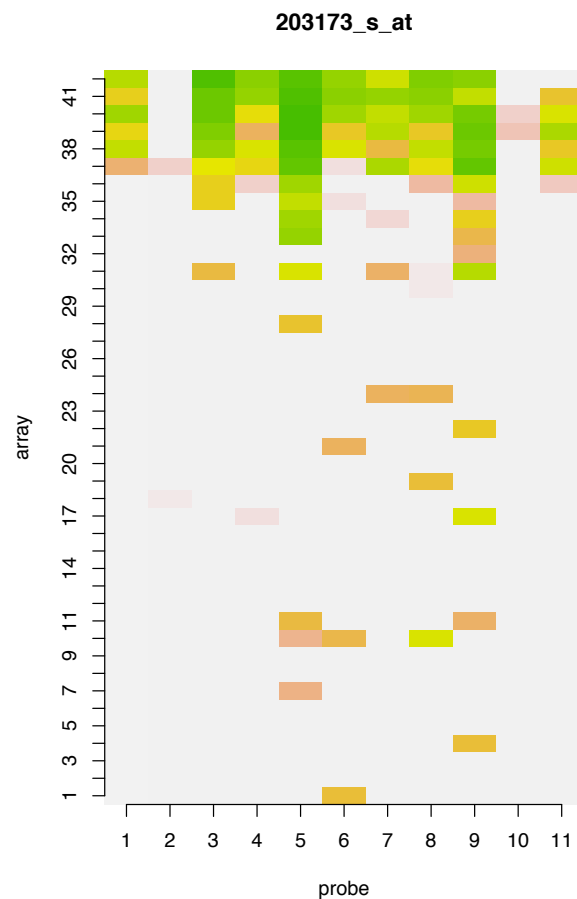


Identifies poor performing probes





Robust regression leads to various quality assessment metrics



Identifies poor performing samples



Relate to limma objects

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$$E[y_1]=E[y_2]=\alpha_1$$

$$E[y_3]=E[y_4]=\alpha_2$$

$$E[y_5]=E[y_6]=\alpha_3$$

$$\beta = C\alpha = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_2 \end{bmatrix}$$

```
> design
  alpha1 alpha2 alpha3
1      1      0      0
2      1      0      0
3      0      1      0
4      0      1      0
5      0      0      1
6      0      0      1
> cont.matrix <- makeContrasts(beta1="alpha2-alpha1",
                                beta2="alpha3-alpha2",levels=design)
```

```
> cont.matrix
      Contrasts
Levels beta1 beta2
alpha1    -1     0
alpha2     1    -1
alpha3     0     1
```

```
fit <- lmFit(y,design)
```

```
fit.c <- contrasts.fit(fit, cont.matrix)
fit.c <- eBayes(fit.c)
```

```
> head(round(y,2),3)
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] -1.62  1.49  2.50  1.57 -0.71  0.38
[2,] -4.50 -4.95 -3.66 -7.83 -1.59  6.94
[3,] -10.17 -21.90 14.03  3.66 -12.21 -15.26
```

```
> head(round(fit$coef,2),3)
      alpha1 alpha2 alpha3
[1,] -0.07   2.03  -0.16
[2,] -4.73  -5.75   2.67
[3,] -16.04   8.85 -13.74
```

```
> head(round(fit.c$coef,2),3)
      Contrasts
      beta1 beta2
[1,]  2.10 -2.20
[2,] -1.02  8.42
[3,] 24.89 -22.59
```