Preprocessing and normalization

The path from colored specks to priceless data

Intensities are not just mRNA concentrations

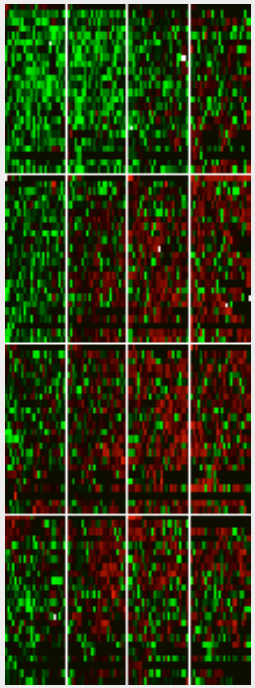
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

- Tiss
- Clo
- Ima
- RN
- PC
conta
- Sig
- Am
- Spatial effects

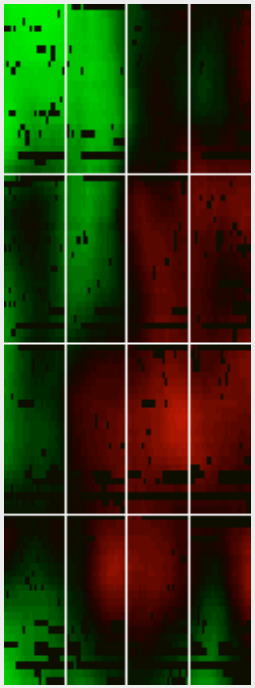
ction
on
ng
ency
d to

Example of spatial effects on microarrays

In theory, the spatial location of a given spot should matter little, since the locations were randomly selected.



Raw data



Spatial bias estimate

But in reality, things like the distribution of solvent over the array surface and the quality of washing, have their say on the matter.

Two degrees of variation

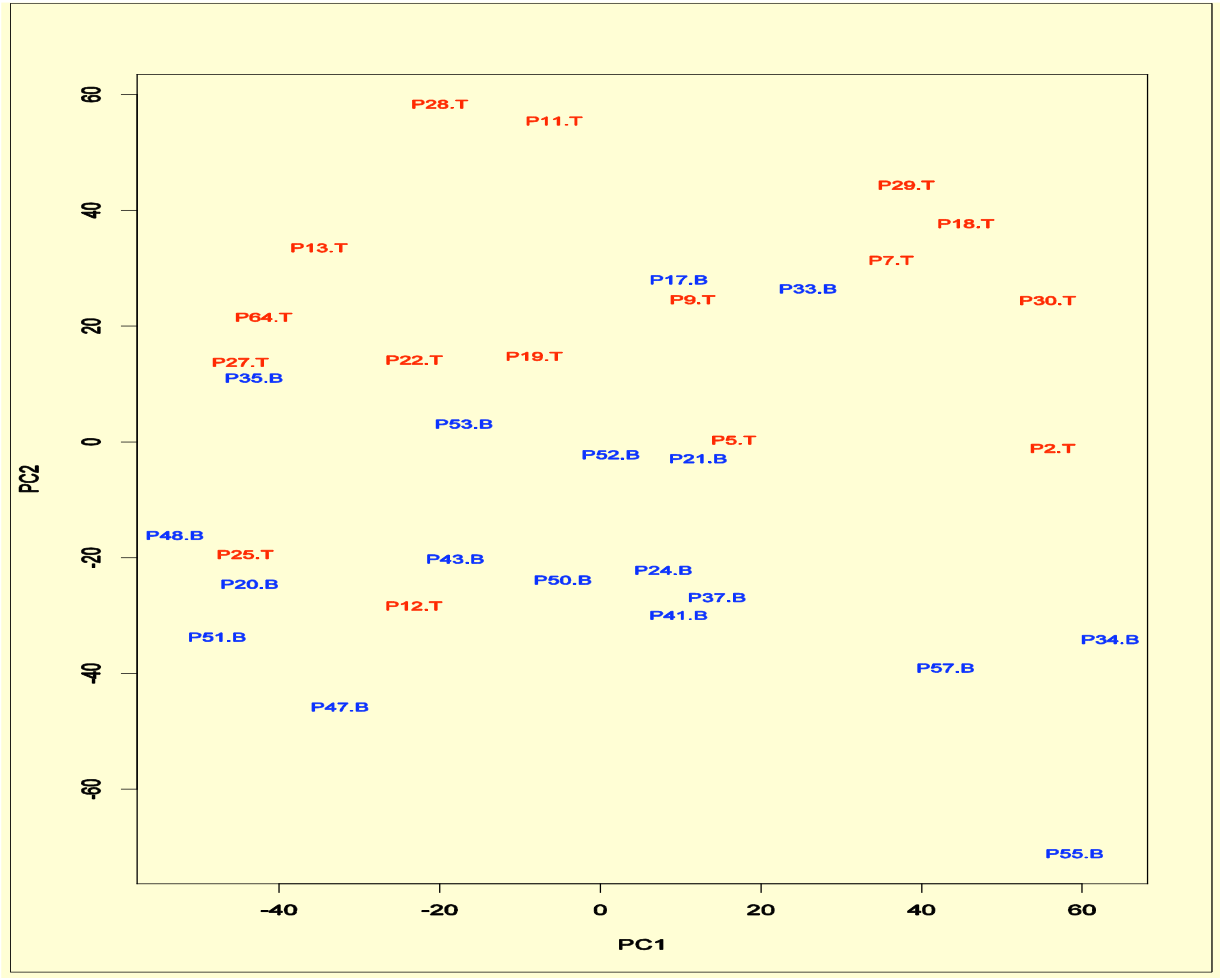
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Array-specific variation:	Gene-specific variation:
Amount of RNA in the biopsy	PCR yield / DNA quality
Efficiencies of: – RNA extraction – Reverse transcription – Labeling – Photodetection	Spotting efficiency, – Spot size
	Cross-/unspecific hybridization
Systematic	**Stochastic**

Stochastic noise can be dealt with by a t-test...

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

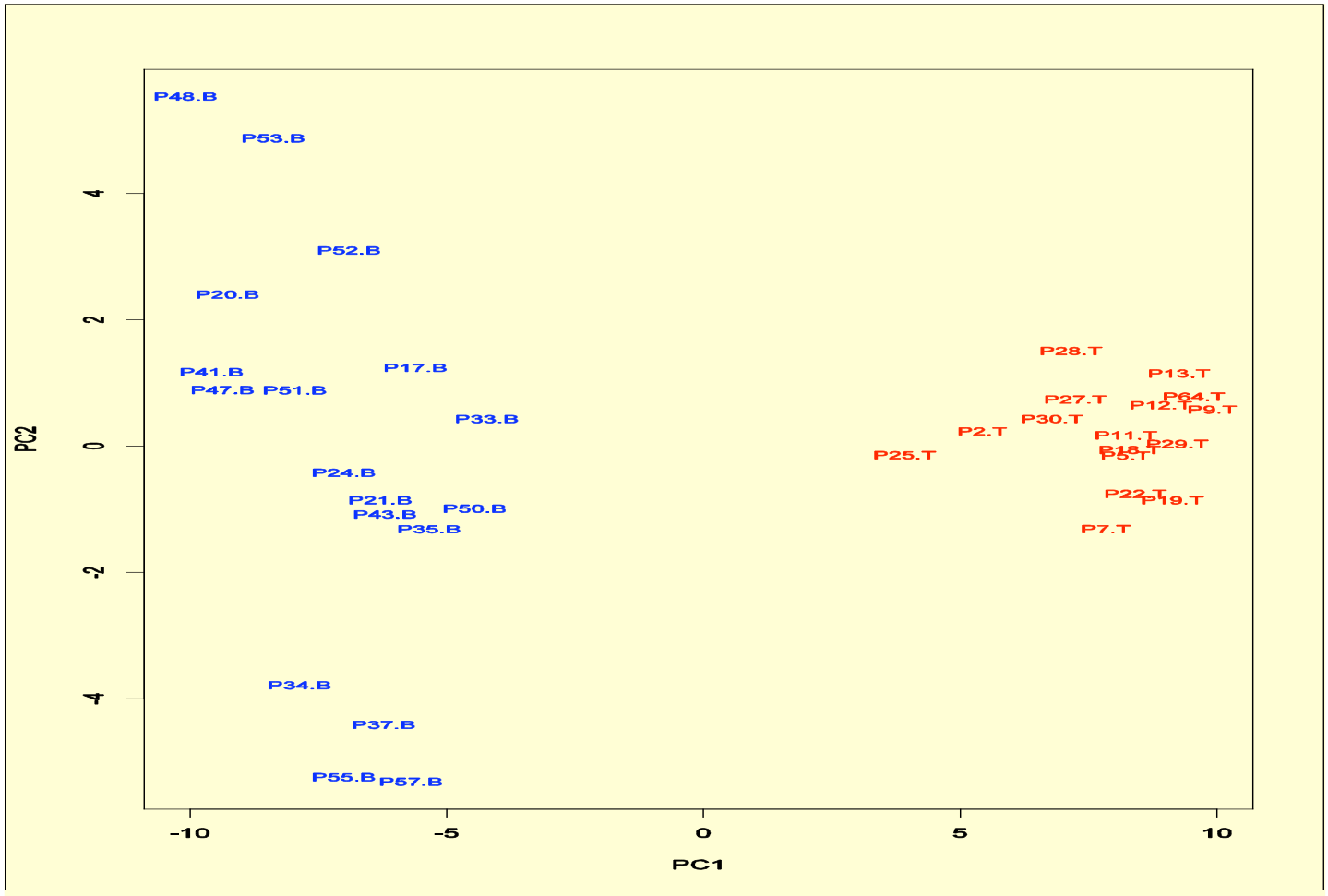
PCA Plot of 34 patients, 8973 dimensions (genes) reduced to 2



...like we will see later today

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

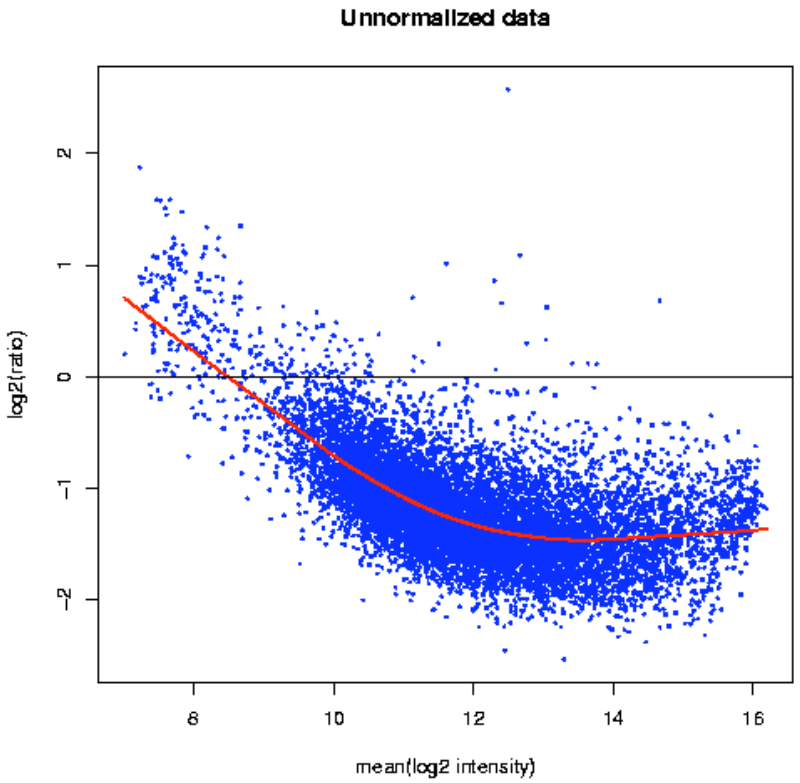
PCA for 100 most significant genes reduced to 2 dimensions

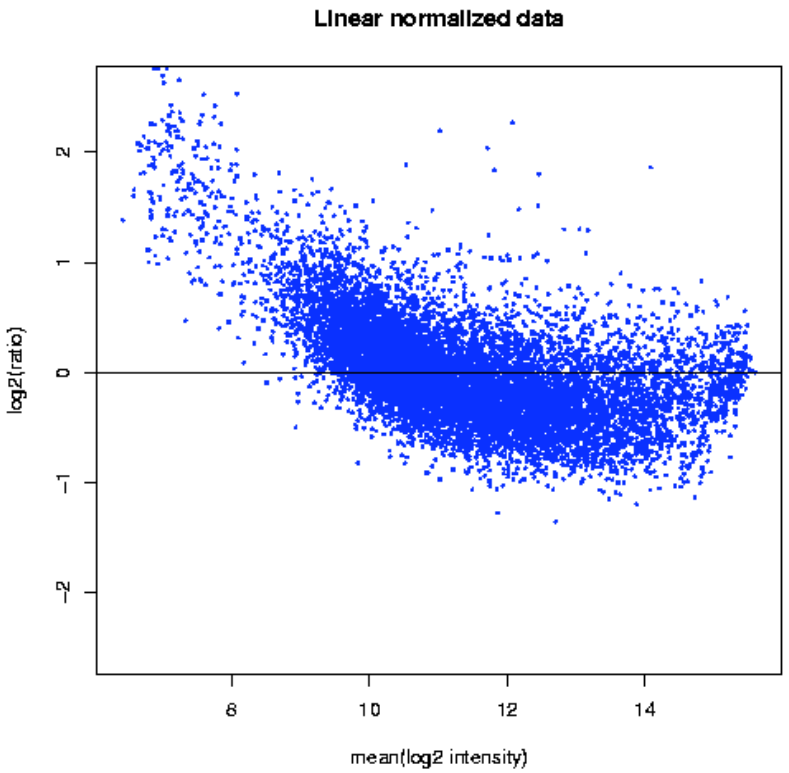


Sources of variation

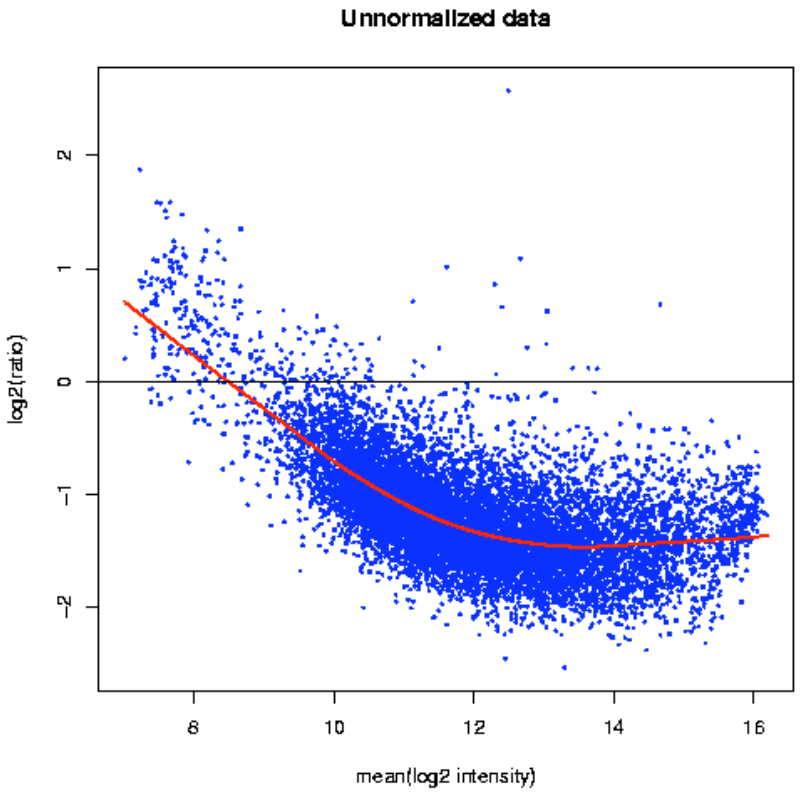
Array-specific variation:	Gene-specific variation:
Systematic	**Stochastic**
• Similar effect on many measurements • Corrections can be estimated from data	• Too random to be explicitly accounted for • “noise”
Normalization	**Statistical testing**

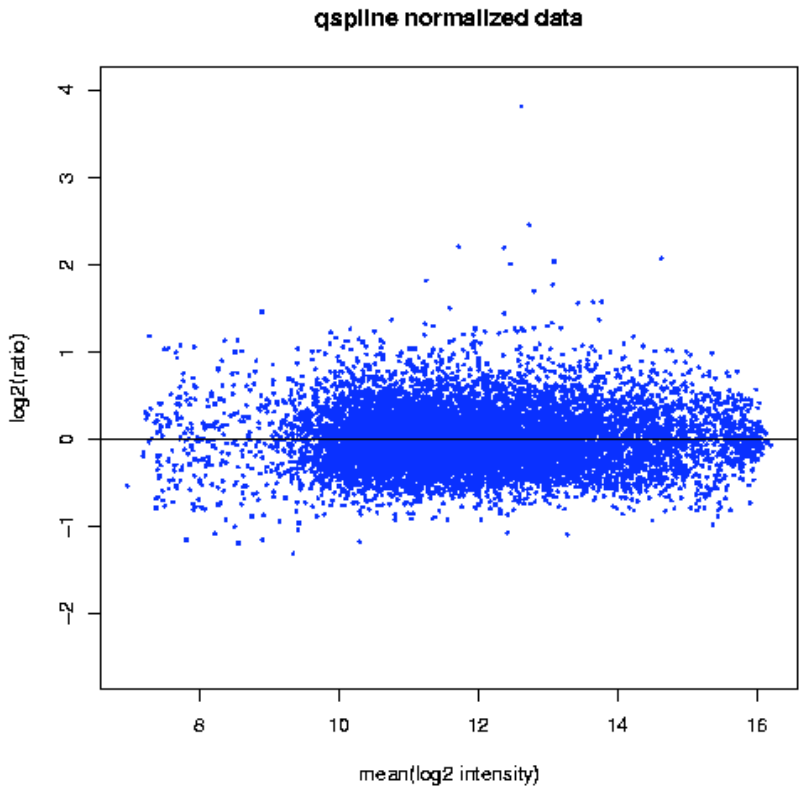
Calibration = Normalization = Scaling



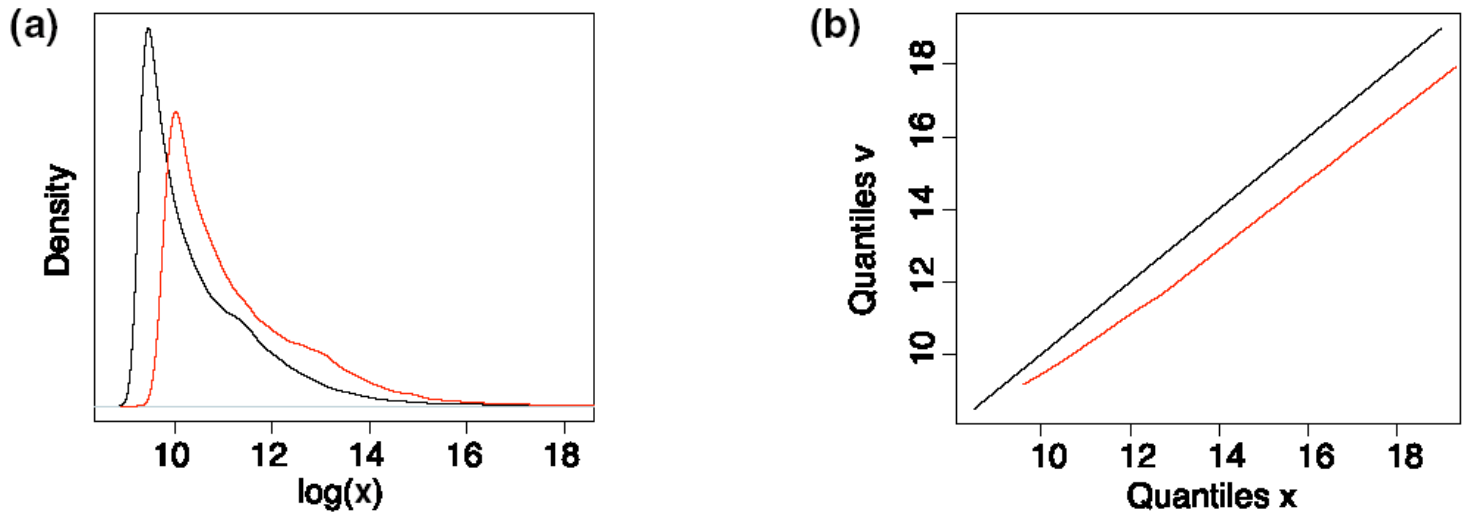


Nonlinear normalization





The Qspline method

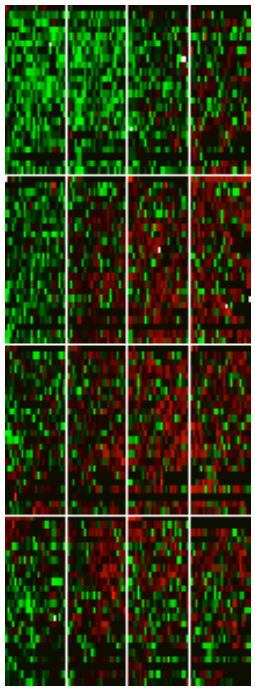


From the empirical distribution, a number of quantiles are calculated for each of the channels to be normalized (one channel shown in red) and for the reference distribution (shown in black)

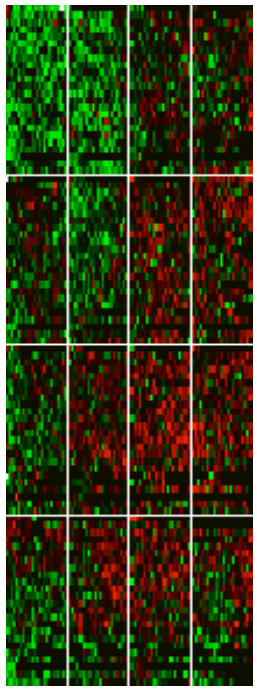
A QQ-plot is made and a normalization curve is constructed by **fitting a cubic spline function**

As **reference** one can use an **artificial "median array"** for a set of arrays or use **a log-normal distribution**, which is a good approximation.

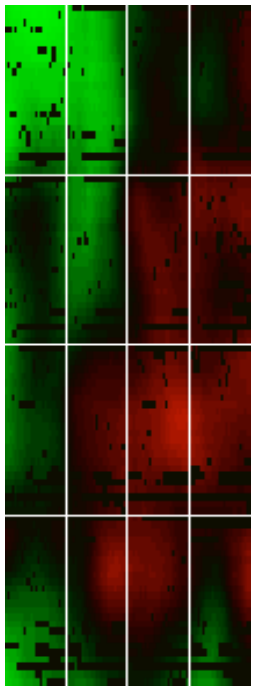
Non-linear normalization



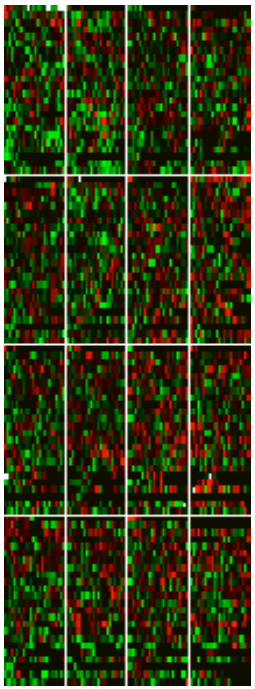
Raw data



After intensity normalization



Spatial bias estimate



After spatial normalization

The really cool thing about R...

...is all the nice libraries out there

The BioConductor packages encompasses many very useful methods for microarray analysis

- Including the *qspline* method, and other normalization algorithms

Check out the www.bioconductor.org website!

Exercise in normalization

- Download the normalization exercise and open the pdf document
  - Please consider that you learn more if you read the commands thoroughly before you copy-and-paste