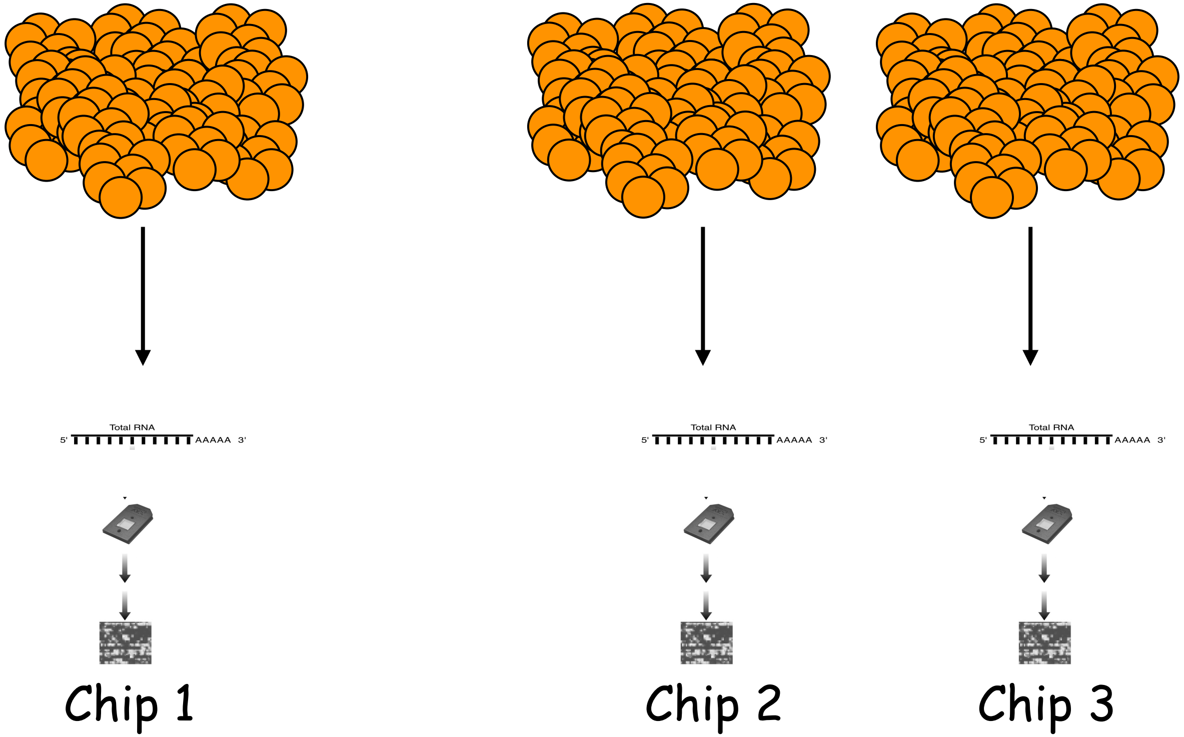
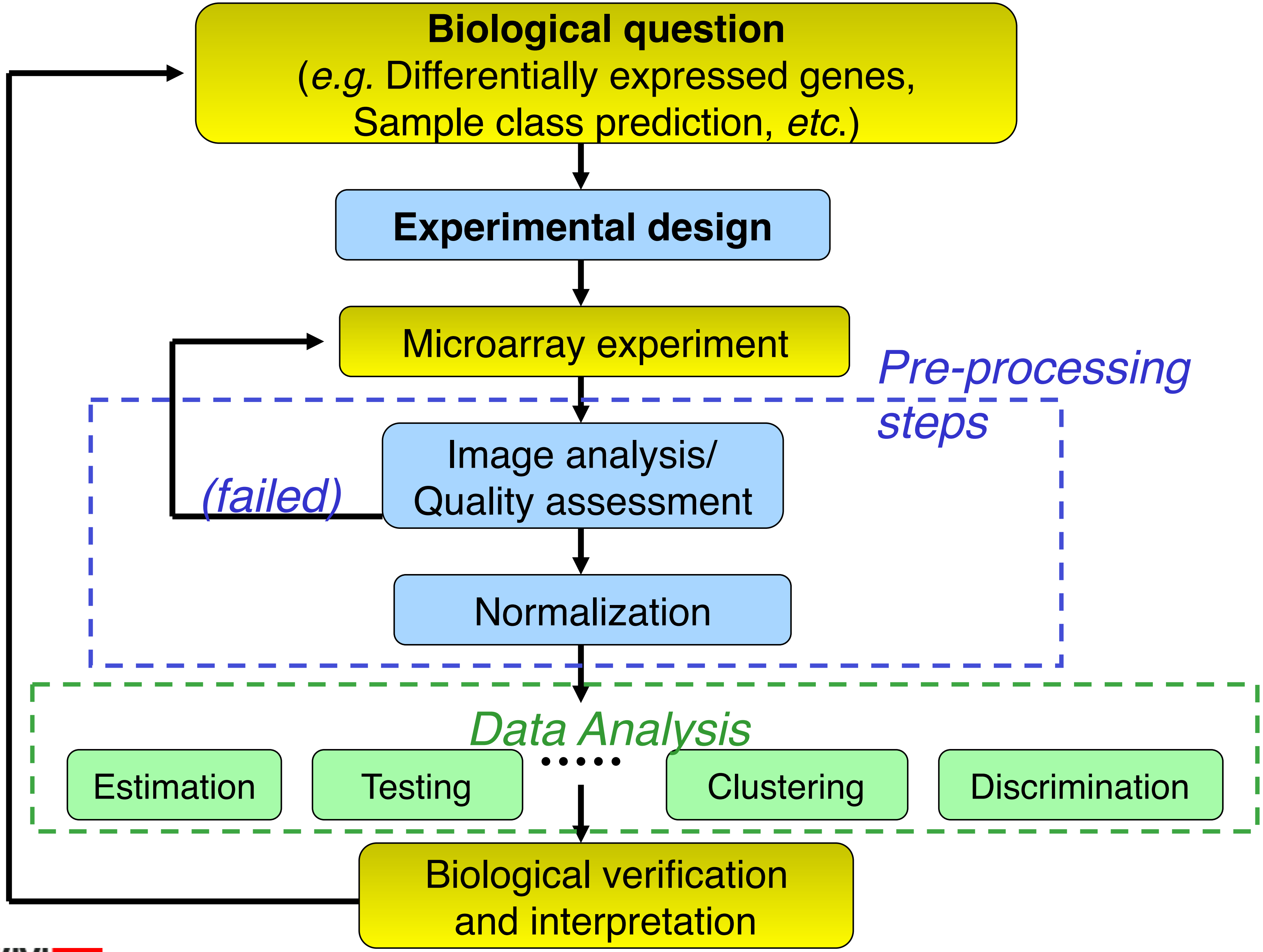
Statistics for Genomic Data Analysis

Experimental design; Linear models



Chip 1 Chip 2 Chip 3


Biological question
(*e.g.* Differentially expressed genes,
Sample class prediction, *etc.*)
Experimental design
Microarray experiment
Pre-processing
steps
Image analysis/
Quality assessment
(failed)
Normalization
Data Analysis
Estimation
Testing
……
Clustering
Discrimination
Biological verification
and interpretation

Replication, Randomization, Blocking

- These are the ‘big three’ of experimental design
- *Replication* – to reduce random variation of the test statistic, increases generalizability
- *Randomization* – to remove bias
- *Blocking* – to reduce unwanted variation
- Idea here is that units within a block are similar to each other, but different between blocks
- ‘Block what you can, randomize what you cannot’

Some Considerations for Microarray Experiments (I)

Scientific (Aims of the experiment)

- Specific questions and priorities
- How will the experiments answer the questions

Practical (Logistic)

- Types of mRNA samples: reference, control, treatment, mutant, etc
- Source, amount of material (tissues, cell lines)
- *Number of slides available (amount of money!)*

Some Considerations for Microarray Experiments (II)

Other Information

- Experimental process prior to hybridization sample isolation, mRNA extraction, amplification, labelling,…
- Controls planned: positive, negative, ratio, *etc.*
- Verification method: Northern, RT-PCR, in situ hybridization, etc.

What is a pilot study?

- A pilot study is a *small scale version* of a full, larger experiment
- Usually, the *pilot sample size is much smaller* than for the full experiment
- Carried out *before* the full experiment

Pilot studies

- *Small scale version* of an experiment
- Sample size *much smaller* than for full experiment
- Carried out *before* the full experiment to be sure the question makes sense *in the system you will be studying*
- To be sure the *techniques work*
  - Practice, standardize techniques
  - identify *problems* and look for *solutions*
- To obtain *preliminary data*
  - practice for statistical analyses
  - see if planned experiment size sufficient

More reasons to do a pilot study

- Gives a relatively *low-cost, quick indication* of the likely outcome of the full experiment
- Determining what *resources* (finance, staff) are needed for the planned study
- Further development or refinement of *research questions* and *research plan*
- *Training* researcher/experimentalist in as many elements of the process as possible
- *Convincing funding bodies*, other research colleagues that the main study is feasible and worth funding

Pilot Study – limitations

- Possibility of making *inaccurate predictions or assumptions* on the basis of pilot data
  - successful pilot does not guarantee success in the full study
  - pilot based on small sample size
- Might *not find all potential difficulties*
- Problems arising from '*contamination*'
  - data from the pilot study are included in the main results, OR
  - pilot participants included in the main study, but new data are collected from them

Replication

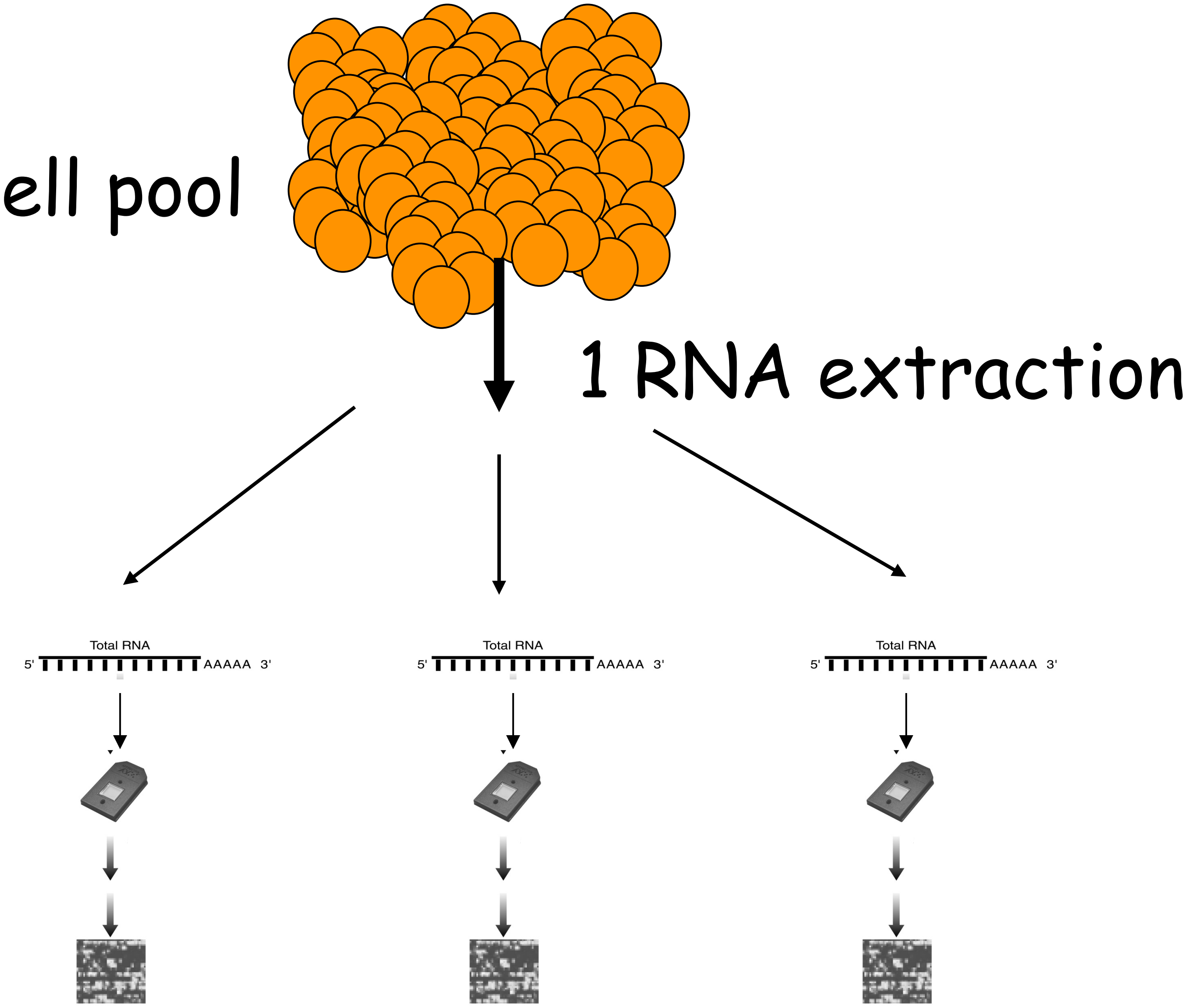
- Why?
  - To reduce variability
  - To increase generalizability

- What is it?
  - Duplicate spots/probes
  - Duplicate slides
    - *Technical replicates* – usually less desirable
    - *Biological replicates*

Biological and Technical Replicates

- Biological replication:
  - multiple cases per group are studied
  - is ***ESSENTIAL***
- Technical replication:
  - RNA sample from one case hybridized to multiple arrays
  - provides information about variability of the labeling, hybridization and quantification processes

Triplicates preparation:

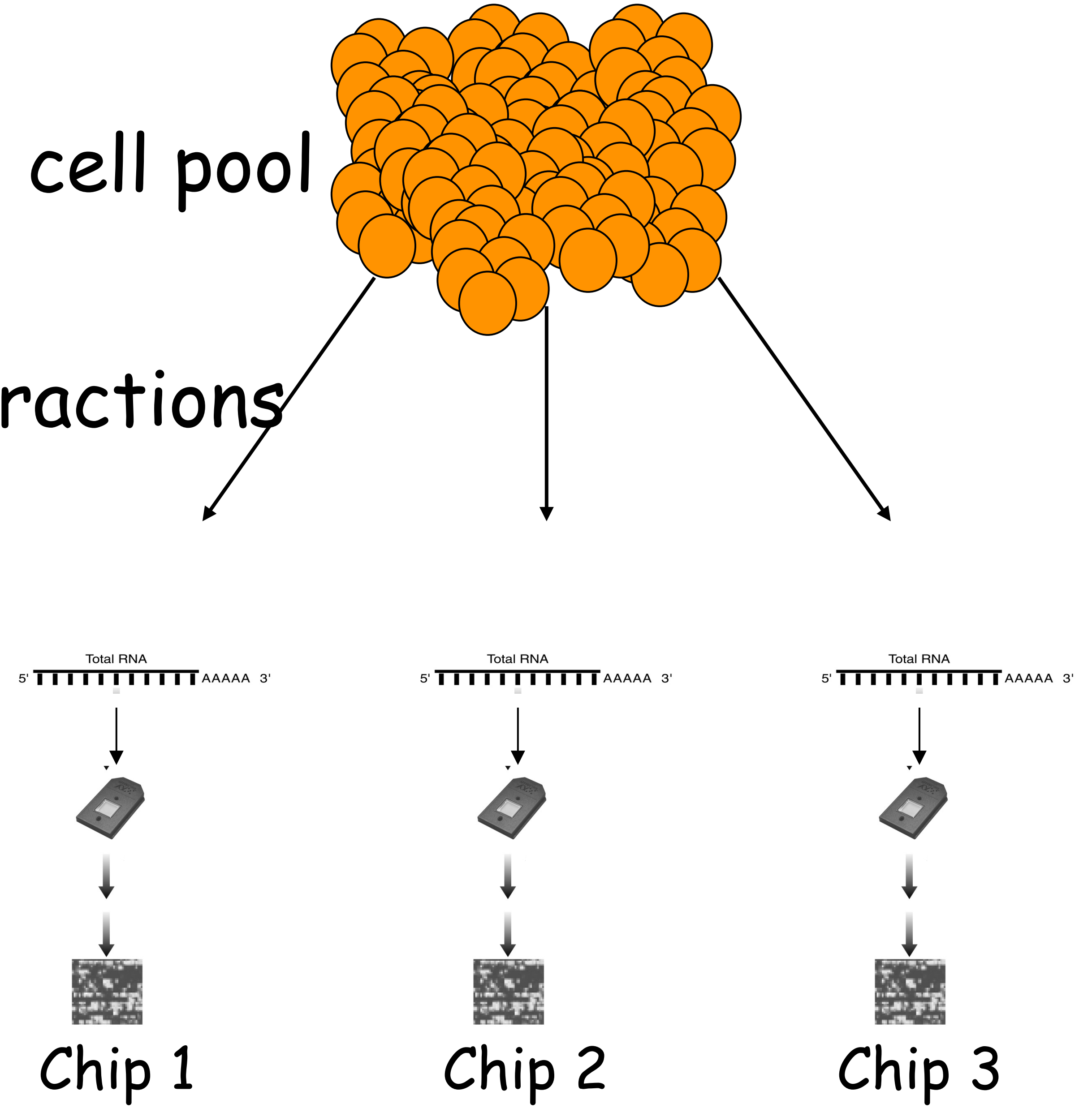


1 cell pool

1 RNA extraction

Chip 1 Chip 2 Chip 3

Triplicates preparation:



1 cell pool

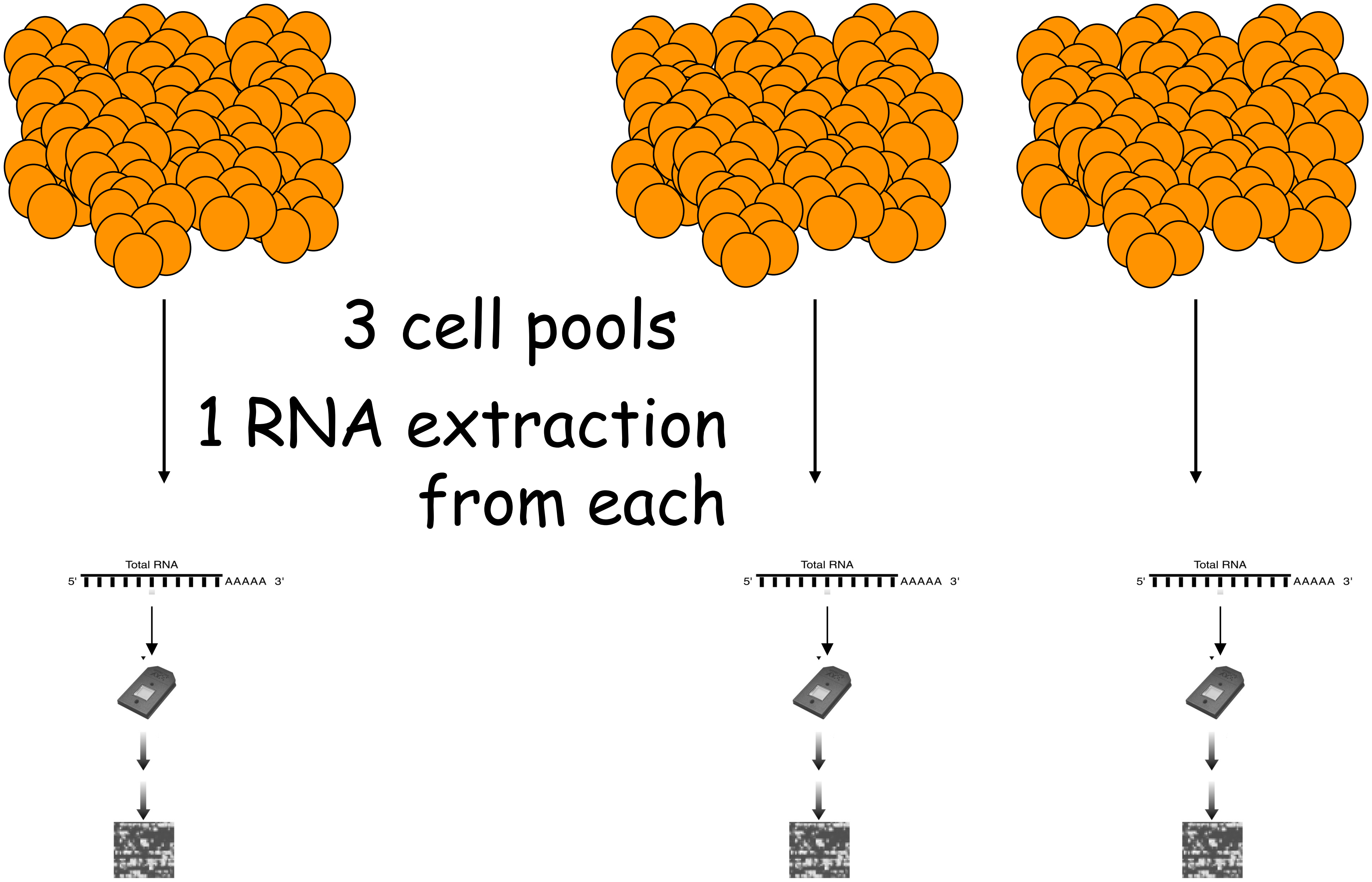
3 RNA extractions

Chip 1

Chip 2

Chip 3

Triplicates preparation:



3 cell pools

1 RNA extraction from each

Chip 1

Chip 2

Chip 3

Replication – Sample size

Statistical considerations:

- *Variance* of individual measurements
- *Effect size(s)* to be detected
- Acceptable *false positive rate*
- Desired *power* (probability of detecting an effect of at least the specified size)

Practical considerations:

- Cost
- Difficulty of obtaining samples
- More difficult than usual, as there are 1,000s of possible changes, each with its own SD

Bottom line: *As many as you can get! (within reason)*

Replication vs. pooling

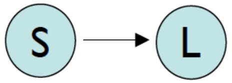
- mRNA from *different samples* are sometimes combined to form a *pooled sample* (or *pool*)
  - If each sample doesn't yield enough mRNA
  - To compensate an excess of variability
- Pooling may be OK if properly done:
  - Combine several samples in each pool
  - Use several pools from different samples
- Do *NOT* use pools when individual information is important (*e.g.* paired designs, classification)
- Never substitute sampling by pooling:
  - A pool of 3 individuals ≠ 3 individual samples !!

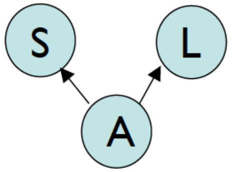
Examples of pooling

- Study with 12 patients : 12 chips = Expensive
- Option 1:
  - Group A: 6 individuals -> 1 pool of 6 -> 1 chip
  - Group B: 6 individuals -> 1 pool of 6 -> 1 chip
- Option 2:
  - Group A: 12 individuals -> 4 pools of 3 -> 4 chips
  - Group B: 12 individuals -> 4 pools of 3 -> 4 chips
- Option 2 *may* have similar precision to full expt.
- (But cannot know for certain without info about variability between individuals and within pools)

Confounding

- Ideally, both the treatment and control groups are exactly alike in all respects (except for group membership)
- A *confounding factor* (or *confounder*) is associated with *both* the group membership and the response
- Reduce/remove effects of confounders through randomization and blocking
- Example: shoe size + literacy



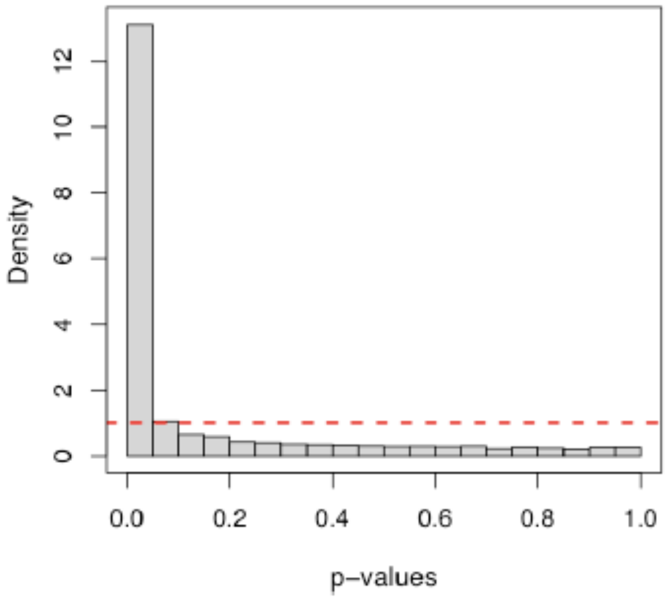


Confounding – genomic example

- Nature Genetics 39, 226 - 231 (2007)

Common genetic variants account for differences in gene expression among ethnic groups

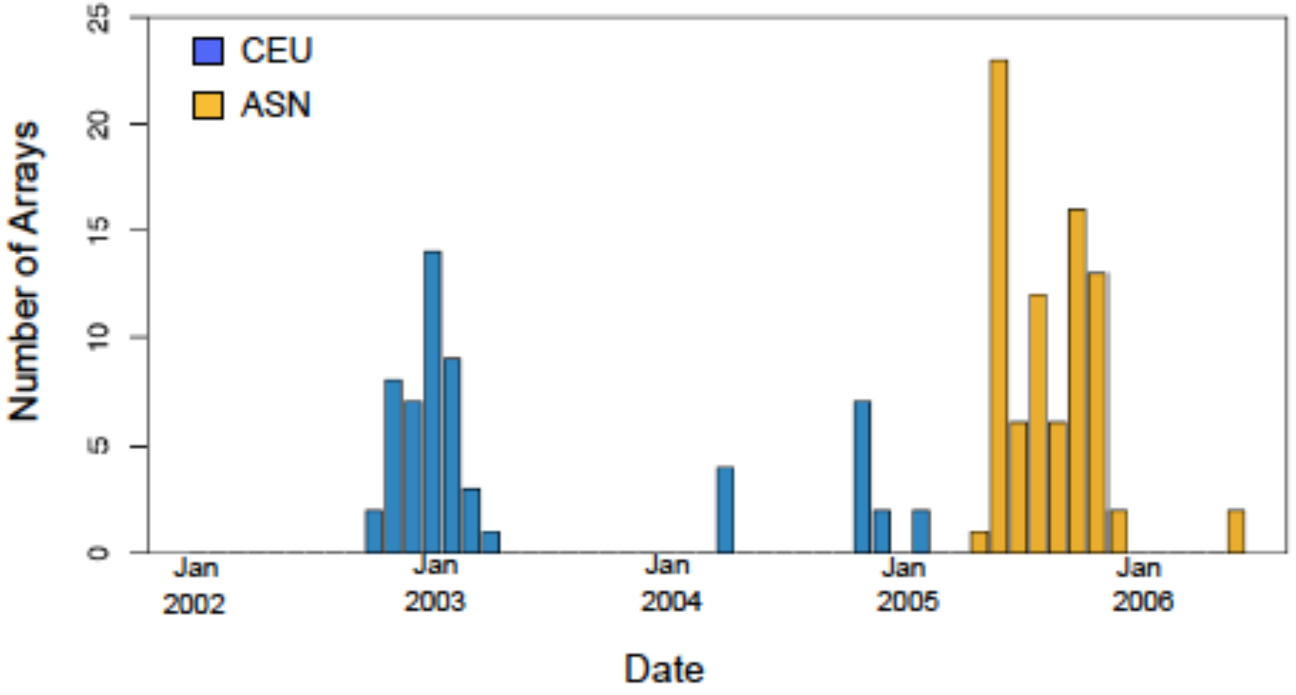
Richard S Spielman[1], Laurel A Bastone[2], Joshua T Burdick[3], Michael Morley[3], Warren J Ewens[4] & Vivian G Cheung[1,3,5]



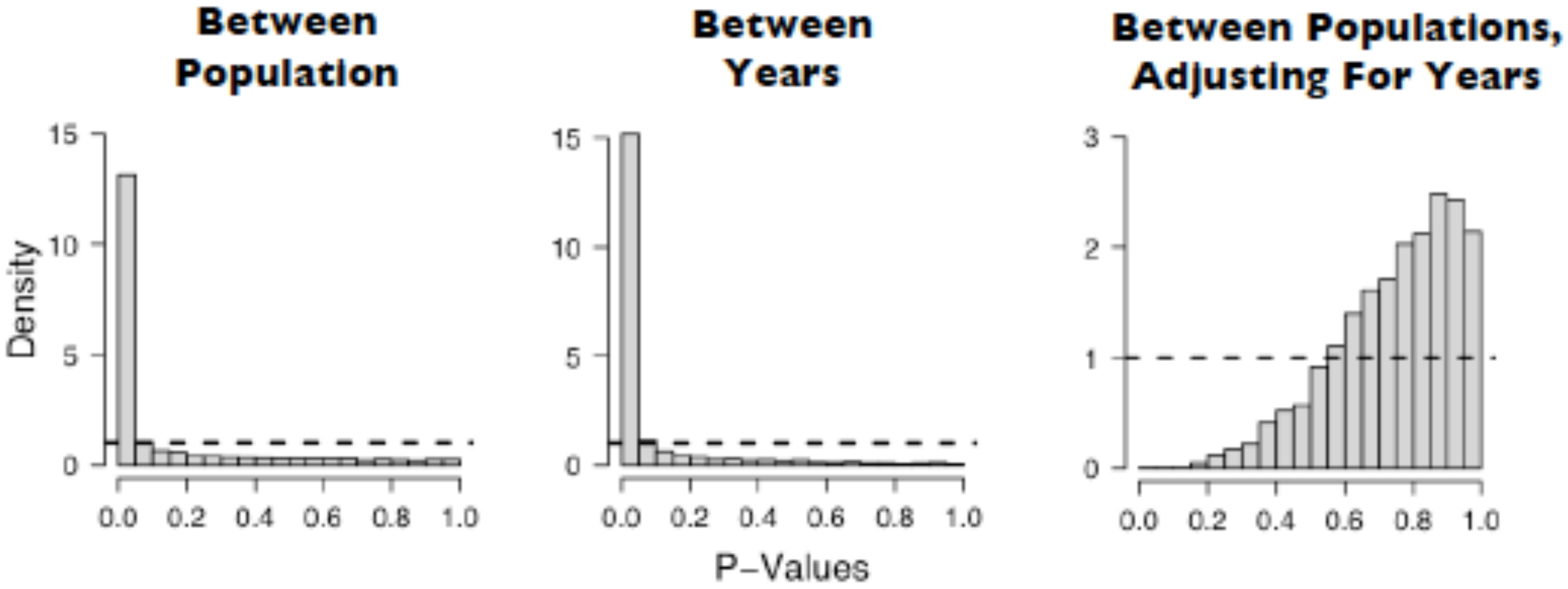
78% of genes 'differentially expressed'

Confounding factor: time

- Time of hybridization confounded with population membership:



Re-analysis - *NO* DE (!!)



78% of genes estimated to be differentially

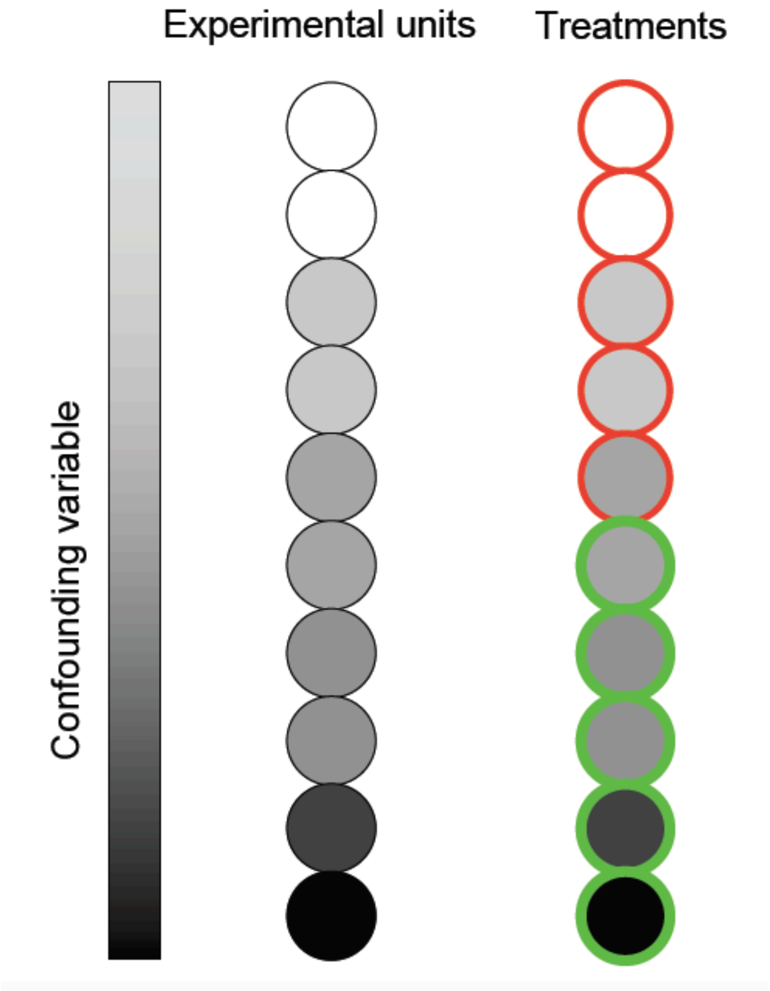
96% of genes estimated to be differentially

0% of genes estimated to be differentially

Randomization

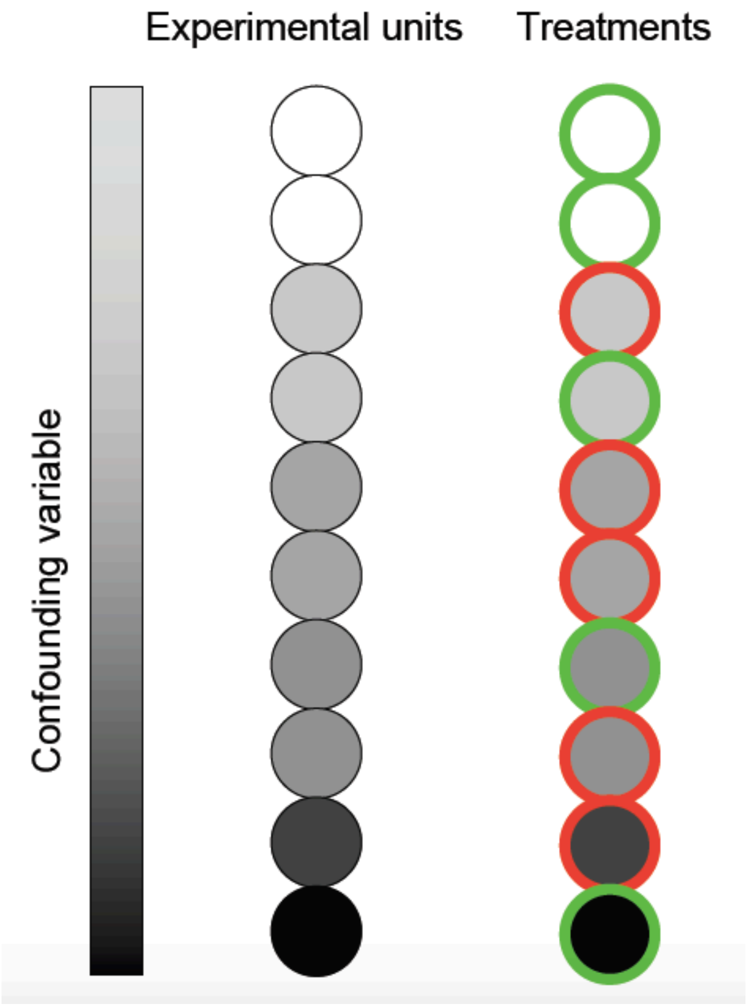
- *Especially important* in larger experiments
  - *e.g.* many samples, different techs, long time, ...
- *Randomization* – to remove bias
  - Would like to 'even out' confounders between groups
  - Do *NOT* process all your control samples on one day and all the treatments on another

Without randomization



Without randomization, confounding variable *differs* among treatments

With randomization



With randomization, confounding variable *does not differ* among treatments

(BREAK)

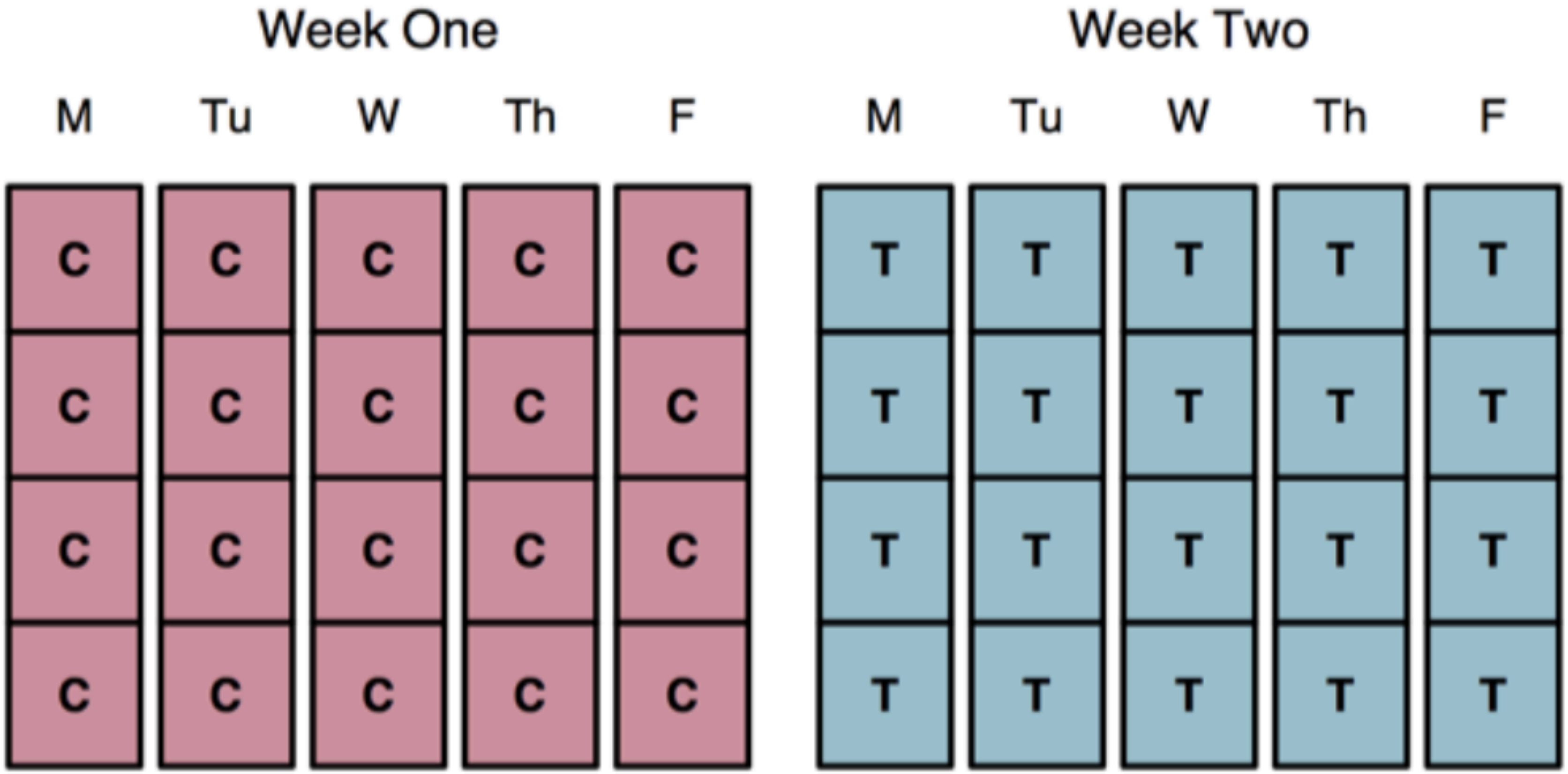
Blocking (local control)

- *Blocking* consists of grouping *similar* individuals (experimental units)
- The idea is that individuals *within* a block are more similar than are individuals *between* blocks
  - e.g., drug treatment given to men and women
  - randomize *separately* within blocks
- Reduce *unwanted variation* and gain *precision*
  - *Example:* using chips from the same batch
- Must know the blocking factor(s) *in advance*
- 'Block what you can, randomize what you cannot'

Example - blocking

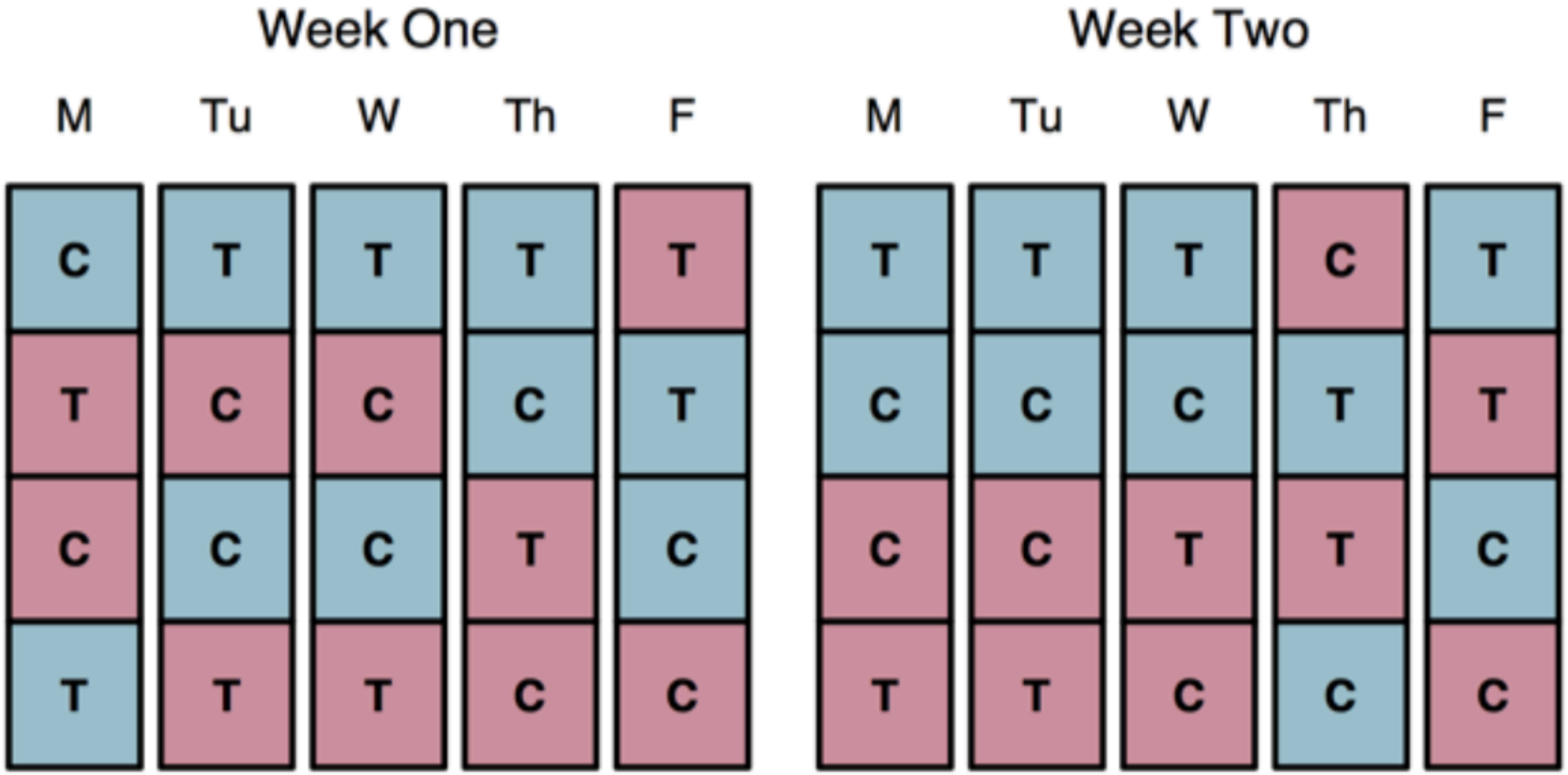
- 20 males, 20 females
- Half to be treated, half left untreated
- Can only work on 4 individuals per day
- *Question:*
  - How to assign individuals to treatment groups and to days?

A poor design *(why??)*



Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T

A better design *(why??)*

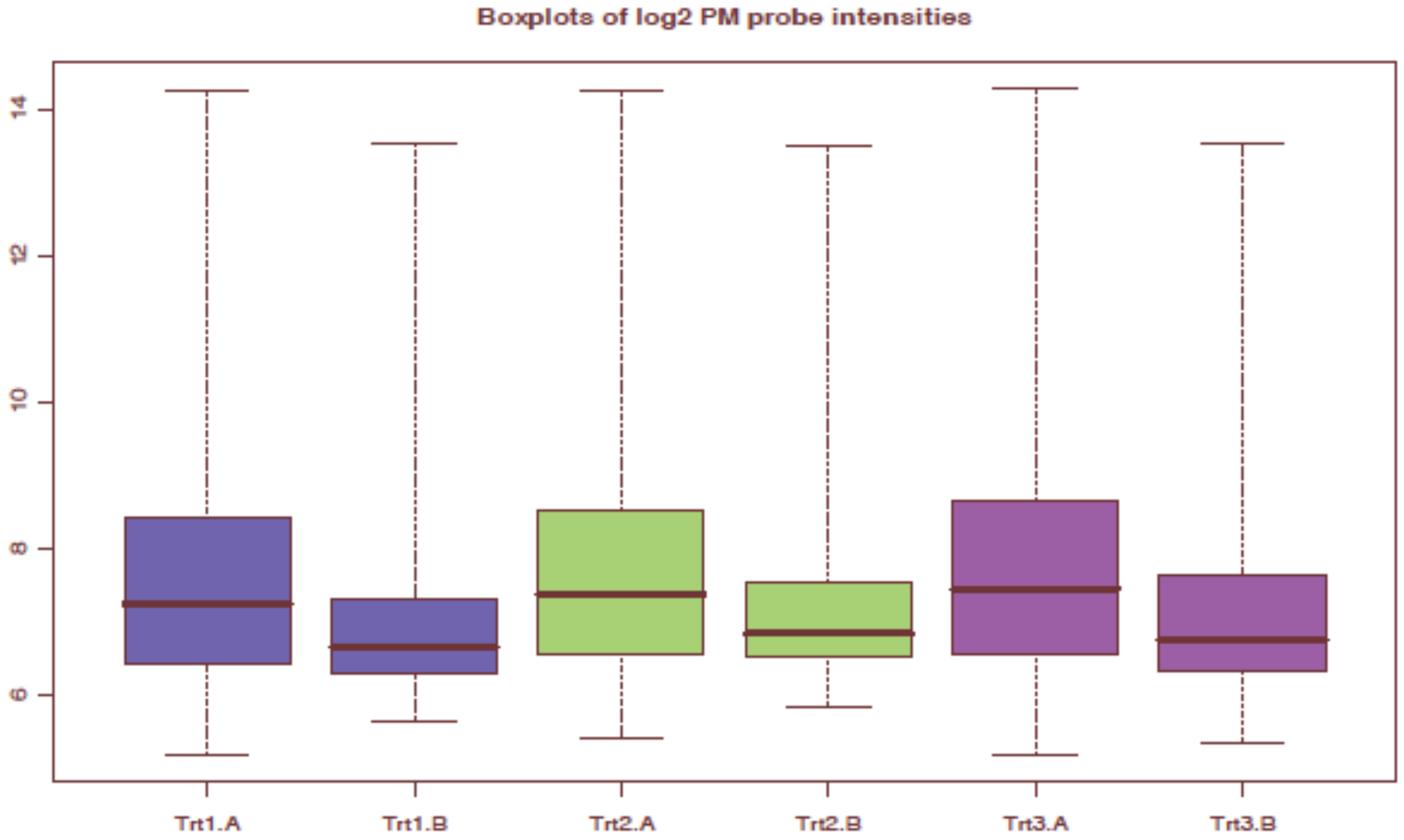


Week One					Week Two				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	T	T	T	T	T	T	T	C	T
T	C	C	C	T	C	C	C	T	T
C	C	C	T	C	C	C	T	T	C
T	T	T	C	C	T	T	C	C	C

Reducing technical variability and avoiding confounding

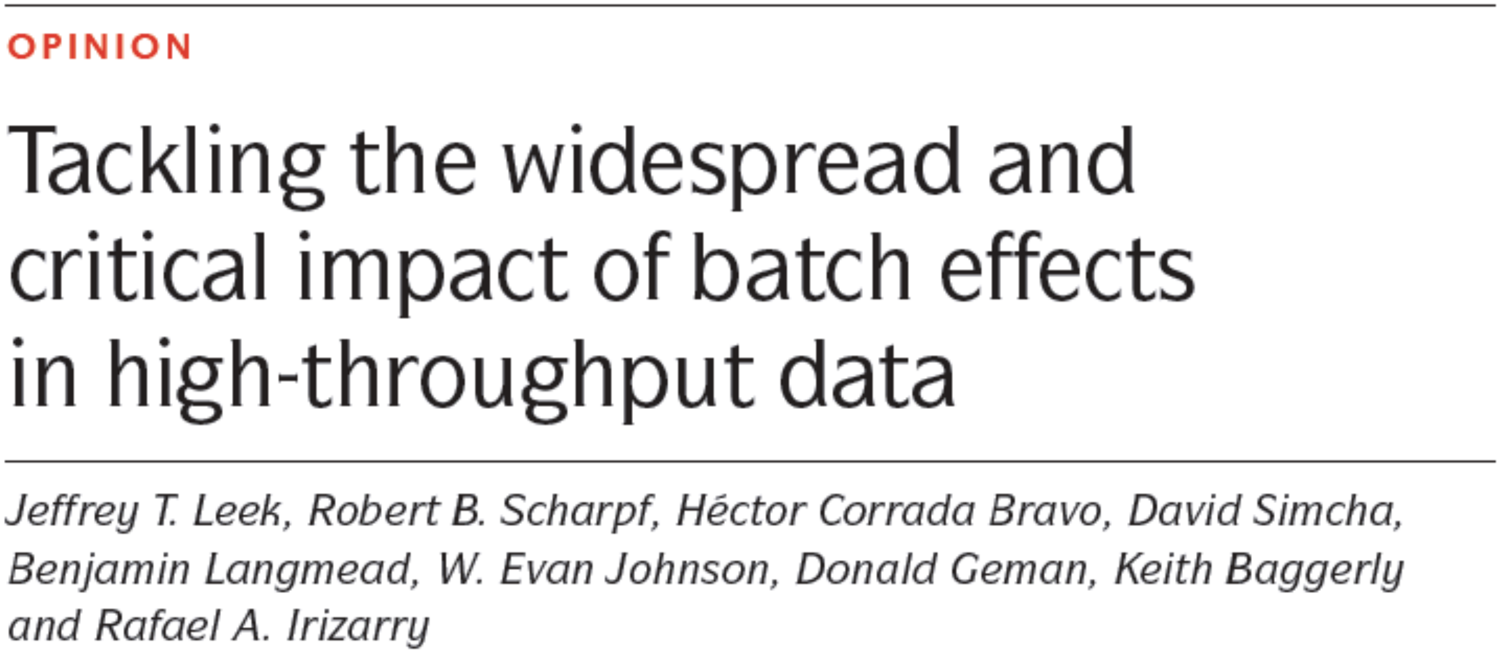
- Attempt to *reduce technical variability* and *avoid confounding* in a study
- If possible, sample collection, RNA extraction and labeling of all samples should be performed by the same individual at the same time of day using the same protocol and reagents
- If samples become available at different times, *consider freezing then processing together*
- If possible, arrays should be used from *a single manufacturing batch* and processed by *one technician* on the *same day*

Typical example of batch effect - completely replicated experiment



Dealing with batch effects and other technical artifacts

- Nature Reviews Genetics 11, 733-739



OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

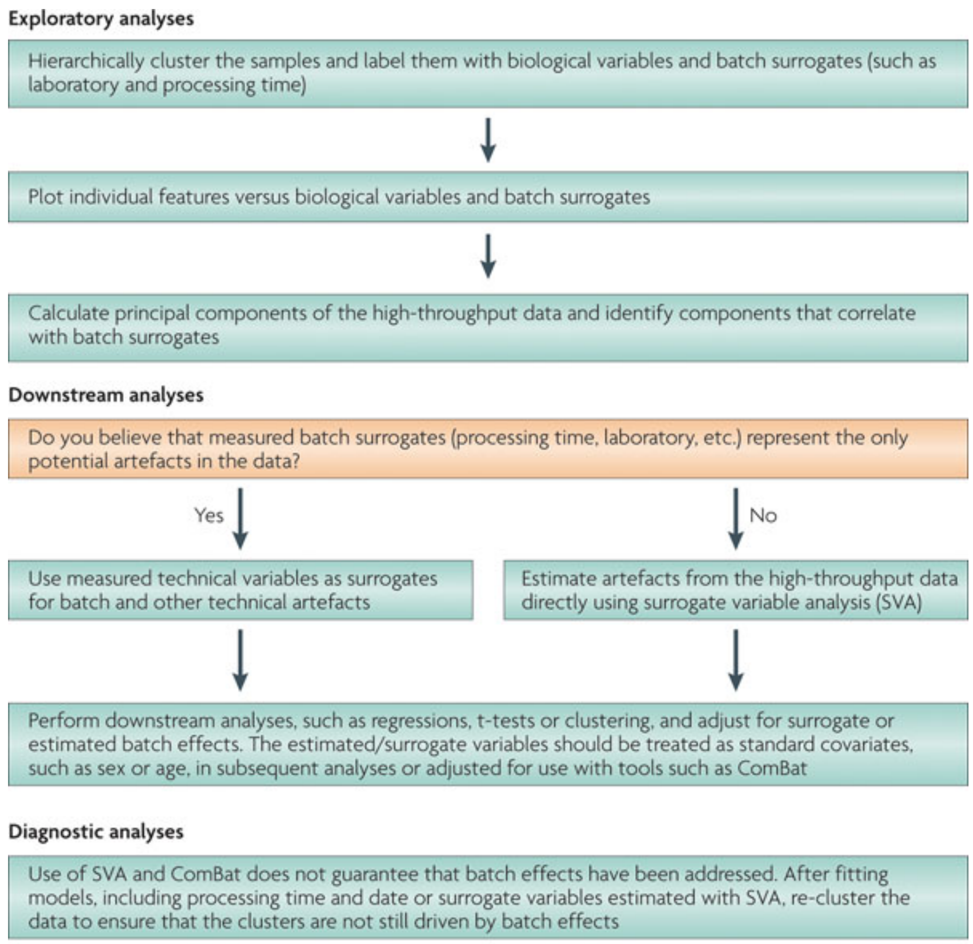
Experimental design solutions

- Careful study design
  - *distribute* batches and other potential *sources of experimental variation* across biological groups
  - *record information* about personnel, reagents, sample storage and labs
- Large experiments/experiments carried out over a long time period most susceptible (but smaller studies not immune)

Statistical solutions

- *Exploratory analyses* to identify and quantify batch effects (and other technical artifacts)
- *Adjust* later ('downstream') statistical analyses to account for these unwanted effects
- Carry out *diagnostic analyses* – did the adjustment work?

Dealing with batch effects - summary



Exploratory analyses

Hierarchically cluster the samples and label them with biological variables and batch surrogates (such as laboratory and processing time)

Plot individual features versus biological variables and batch surrogates

Calculate principal components of the high-throughput data and identify components that correlate with batch surrogates

Downstream analyses

Do you believe that measured batch surrogates (processing time, laboratory, etc.) represent the only potential artefacts in the data?

Yes

Use measured technical variables as surrogates for batch and other technical artefacts

No

Estimate artefacts from the high-throughput data directly using surrogate variable analysis (SVA)

Perform downstream analyses, such as regressions, t-tests or clustering, and adjust for surrogate or estimated batch effects. The estimated/surrogate variables should be treated as standard covariates, such as sex or age, in subsequent analyses or adjusted for use with tools such as ComBat

Diagnostic analyses

Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects

Nature Reviews | Genetics

Some common experiments

- Comparison of *2 conditions*/types (‘treatment vs. control’)
  - mutant vs. wild type plants
  - liver vs. heart in mouse
- Comparison of *many treatments* to a control
- *Clinical studies* (*e.g.* cancer patients)
- *Time course* – measurements at different times
- *Factorial study* – multiple conditions varied and studied *simultaneously*

Factorial crossing

- Compare 2 (or more) sets of conditions in the *same experiment*
- Designs with factorial treatment structure allow you to measure *interaction* between two (or more) sets of conditions that influence the response
- Factorial designs may be either observational or experimental

Replication in factorial experiment

- One observation per *cell* (combination of levels of factor A and factor B)
  - can estimate full model parameters but no *df* left over for inference
  - can assume no interaction – assess graphically
- More than one observation per cell
  - when all $n_i = n$ (*balanced design*) the design is *orthogonal*
  - orthogonality can also occur if row/column cell numbers are *proportional*
  - orthogonality is good – most precise estimation and easiest to interpret parameters
- *Bottom line*: design with *equal replicates* usually best

Balanced vs. Unbalanced Experimental Designs

- *Balanced design:* Cell sample sizes are proportional (maybe equal)
- Explanatory variables have *zero relationship* to one another
- Numerator SS in ANOVA are *independent* => order of variables in model doesn't matter
- Most experimental studies are designed this way – analysis is *most simple*
- As soon as somebody drops a test tube, it's no longer (exactly) true!

Analysis of unbalanced data

- When explanatory variables are related, there is potential *ambiguity*
  - A is related to Y, B is related to Y, and A is related to B
  - Which variable gets credit for the portion of variation in Y that could be explained by either A or B?
- *Order of variables* in model fitting makes a difference
- Analysis more complicated, messy

Gene expression data

Data on *G* genes for *n* samples:

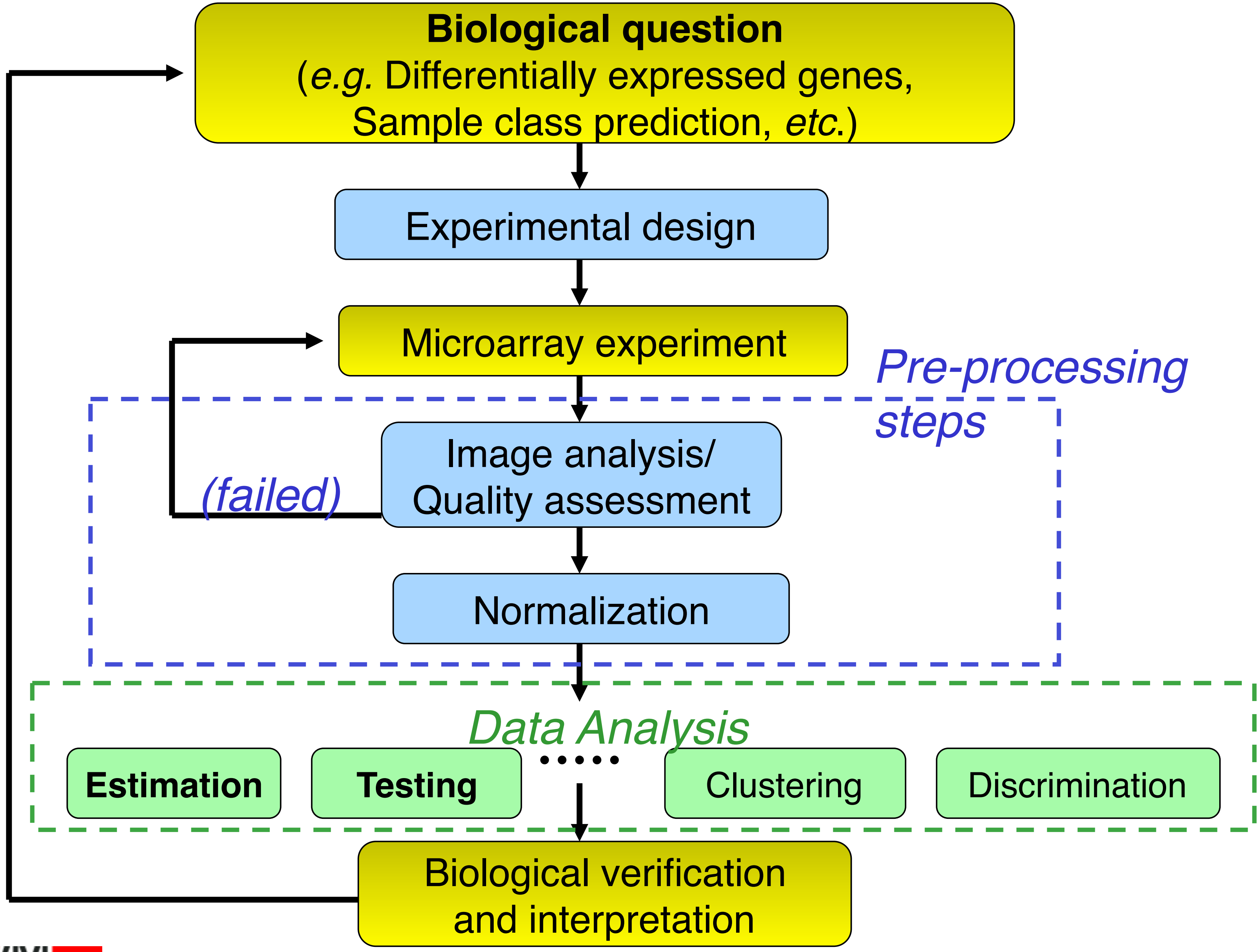
mRNA samples

Genes

	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene *i* in mRNA sample *j*

= (normalized) Log_2(Red intensity / Green intensity)
or: *RMA* value



Biological question
(*e.g.* Differentially expressed genes,
Sample class prediction, *etc.*)
Experimental design
Microarray experiment
Pre-processing steps
Image analysis/
Quality assessment
(failed)
Normalization
Data Analysis
Estimation
Testing
……
Clustering
Discrimination
Biological verification
and interpretation

Linear models

- In statistics, a 'linear model' refers to a model that is *linear in the parameters*
- Which are linear models?
  1. $Y = \beta_0 + \beta_1 x + \varepsilon$
  2. $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  3. $Y = \beta_0 + \beta_1 e^x + \varepsilon$
  4. $Y = \alpha + e^{\beta x} + \varepsilon$
  5. $Y = \alpha e^{\beta x} \varepsilon$

Linear models

- Simplest version: comparing single treatment (T) to single control (C)

  $y_C = \mu + \varepsilon_C$; $\hat{u} = y_C$

  $y_T = \mu + \alpha + \varepsilon_T$; $\hat{a} = y_T - y_C$

- With multiple observations, the estimates are averages (or differences of averages)
- Readily extends to *more than 2 conditions*
- Matrix notation

Linear modeling

- Simple regression model:

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$

$$\text{response variable} = \underbrace{\underbrace{\substack{\text{population}\\ \text{intercept}}}_{\text{intercept term}} + \underbrace{\substack{\text{population}\\ \text{slope}} \times \substack{\text{predictor}\\ \text{variable}}}_{\text{slope term}}}_{\text{model}} + \text{error}$$

- Multiple regression model:

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \ldots + \varepsilon_i$$

- Anova model:

$$y_{ij} = \mu + \beta_1 (dummy_1)_{ij} + \beta_2 (dummy_2)_{ij} + \ldots. + \varepsilon_{ij}$$

Effects model

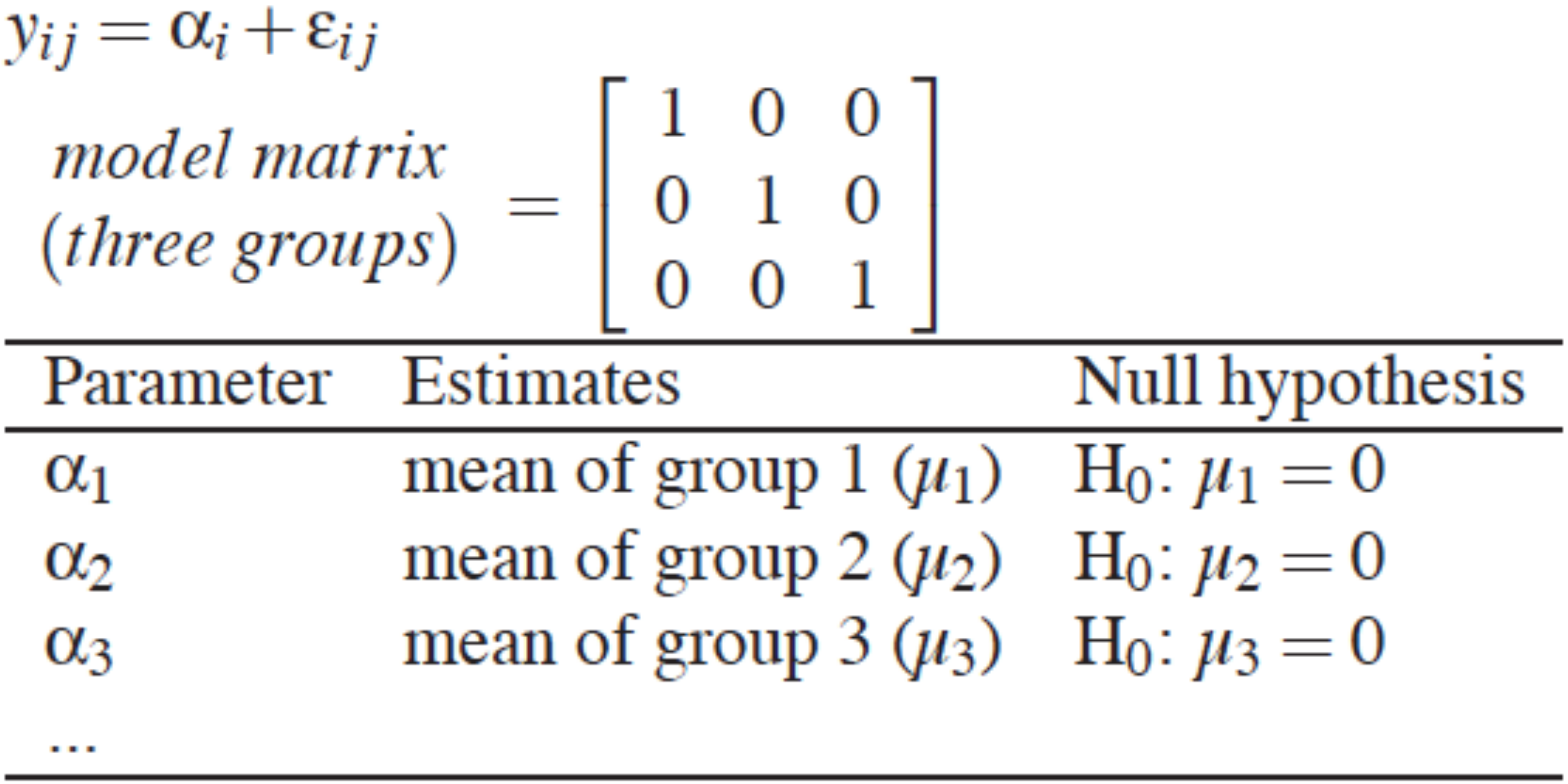
- Anova model more typically expressed as an *effects model* :

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Y	$dummy_1$	$dummy_2$	$dummy_3$
2	1	0	0
3	1	0	0
4	1	0	0
6	0	1	0
7	0	1	0
8	0	1	0
10	0	0	1
11	0	0	1
12	0	0	1

$$design\ matrix = \begin{array}{cccc} \mu & \alpha_1 & \alpha_2 & \alpha_3 \\ \hline 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{array}$$

Set μ to zero



$$y_{ij} = \alpha_i + \varepsilon_{ij}$$

$$\begin{array}{c} \textit{model matrix} \\ \textit{(three groups)} \end{array} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Parameter	Estimates	Null hypothesis
α_1	mean of group 1 (μ_1)	H_0: $\mu_1 = 0$
α_2	mean of group 2 (μ_2)	H_0: $\mu_2 = 0$
α_3	mean of group 3 (μ_3)	H_0: $\mu_3 = 0$
...		

Treatment contrasts

over-parameterized design matrix contrast matrix model matrix

$$\begin{array}{cccc} \text{Intercept} & \alpha_1 & \alpha_2 & \alpha_3 \\ (\mu) & (G1) & (G2) & (G3) \\ \left[\begin{matrix}1\\1\\1\end{matrix}\right. & \begin{matrix}1\\0\\0\end{matrix} & \begin{matrix}0\\1\\0\end{matrix} & \left.\begin{matrix}0\\0\\1\end{matrix}\right] \end{array} * \begin{array}{cc} & \begin{array}{cc}\alpha_2^* & \alpha_3^*\end{array} \\ \begin{array}{c}G1\\G2\\G3\end{array} & \begin{bmatrix}0 & 0\\1 & 0\\0 & 1\end{bmatrix} \end{array} \Rightarrow \begin{array}{ccc} \text{Intercept} & \alpha_2^* & \alpha_3^* \\ \left[\begin{matrix}1\\1\\1\end{matrix}\right. & \begin{matrix}0\\1\\0\end{matrix} & \left.\begin{matrix}0\\0\\1\end{matrix}\right] \end{array}$$

Parameter	Estimates	Null hypothesis
Intercept	mean of ‘control’ group (μ_1)	H_0: $\mu = \mu_1 = 0$
α_2^*	mean of group 2 minus mean of ‘control’ group ($\mu_2 - \mu_1$)	H_0: $\alpha_2^* = \mu_2 - \mu_1 = 0$
α_3^*	mean of group 3 minus mean of ‘control’ group ($\mu_3 - \mu_1$)	H_0: $\alpha_3^* = \mu_3 - \mu_1 = 0$
...		

Sum to zero contrasts

over-parameterized design matrix | contrast matrix | model matrix

$$\begin{array}{cccc} \text{Intercept} & \alpha_1 & \alpha_2 & \alpha_3 \\ (\mu) & (G1) & (G2) & (G3) \\ \hline 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array} * \begin{array}{c|cc} & \alpha_1^* & \alpha_2^* \\ \hline G1 & 1 & 0 \\ G2 & 0 & 1 \\ G3 & -1 & -1 \end{array} \Rightarrow \begin{array}{ccc} \text{Intercept} & \alpha_1^* & \alpha_2^* \\ \hline 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{array}$$

Parameter	Estimates	Null hypothesis
Intercept	mean of group means (μ_{i*}/p)	H$_0$: $\mu = \mu_q/p = 0$
α_1^*	mean of group 1 minus mean of group means ($\mu_1 - (\mu_q/p)$)	H$_0$: $\alpha_1 = \mu_1 - (\mu_q/p) = 0$
α_2^*	mean of group 2 minus mean of group means ($\mu_2 - (\mu_q/p)$)	H$_0$: $\alpha_2 = \mu_2 - (\mu_q/p) = 0$
...		

Typical analysis using `limma`

- Read in data
- Create design matrix
- Create contrast matrix (if needed)
- Fit model
- Make comparisons
- Output interesting results

Design Matrix and Contrasts

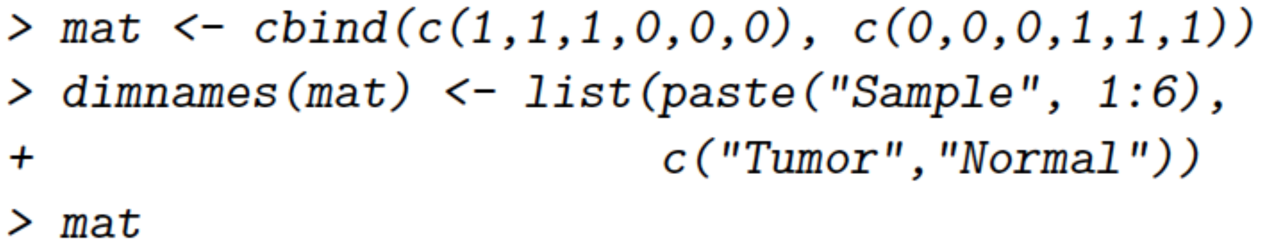
- The *design matrix* indicates the hybs (which RNA hybridized to each array)
- The *contrasts* are the comparisons of interest
- Making the design matrix for *common reference* or *single color arrays* is the same as for ordinary regression/anova
- (*more involved* for (2-color) direct designs)

Design matrix for 2 group comparison

- Predictors are (unordered) factors
  - tumor/normal
  - experimental/control
  - mutant/wild type
- Decide on model, *THEN* create design matrix
  - Do *NOT* create design matrix and then figure out what the model is (!!)
  - Design model to reflect hypotheses of interest
- *Tip* : when straightforward, parameterize the model in terms of comparisons of interest

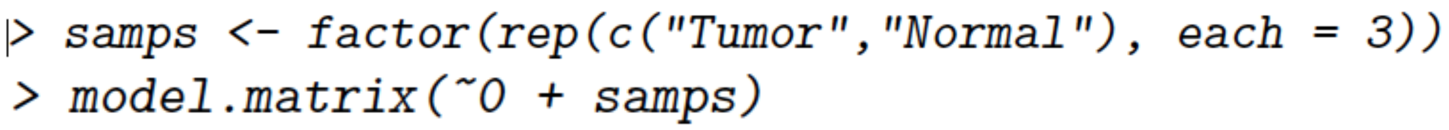
Example: 3 tumor/3 normal samples

- Parameterization:
  - Y = tumor (1_tumor) + normal(1_normal)
- Design matrix:
  - by hand



```
> mat <- cbind(c(1,1,1,0,0,0), c(0,0,0,1,1,1))
> dimnames(mat) <- list(paste("Sample", 1:6),
+                          c("Tumor","Normal"))
> mat
```

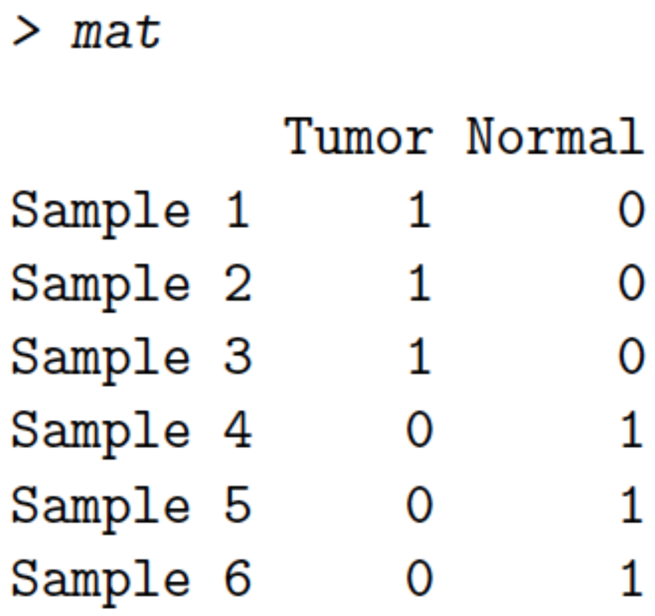
  - Using **model.matrix**



```
|> samps <- factor(rep(c("Tumor","Normal"), each = 3))
> model.matrix(~0 + samps)
```

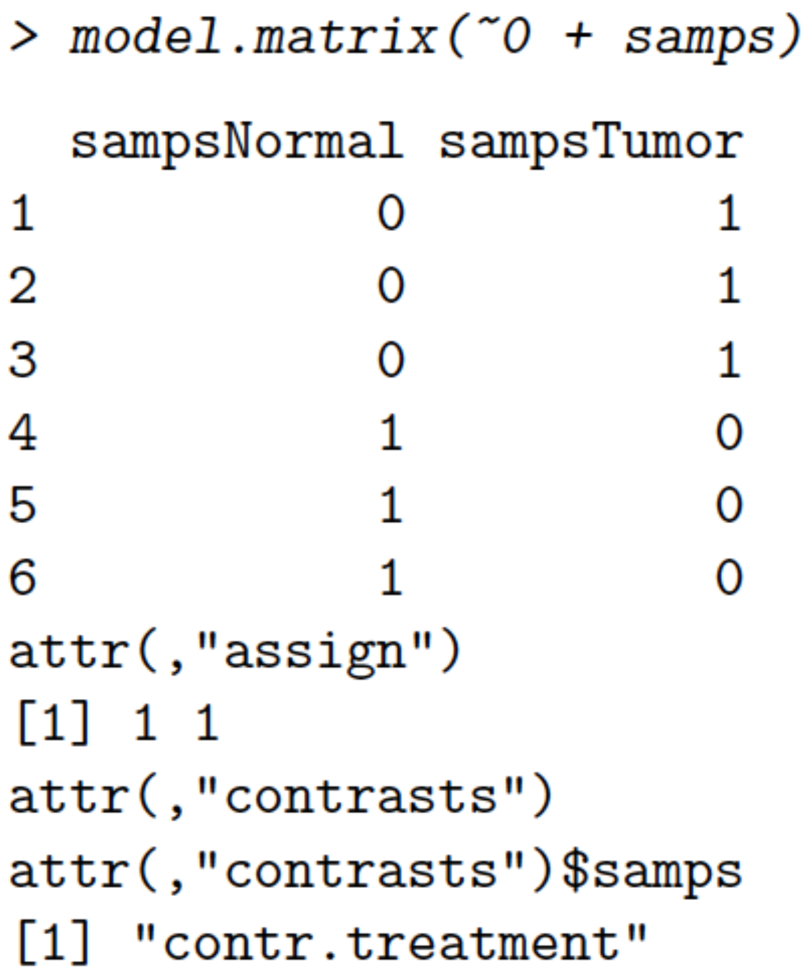
Design matrix for the parameterization

explicitly remove intercept



```
> mat

          Tumor Normal
Sample 1      1      0
Sample 2      1      0
Sample 3      1      0
Sample 4      0      1
Sample 5      0      1
Sample 6      0      1
```

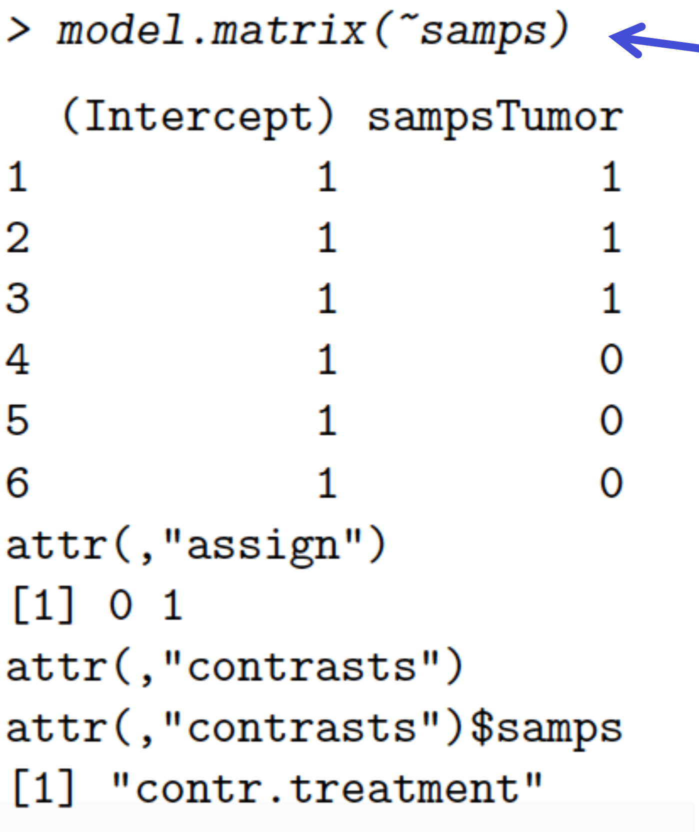


```
> model.matrix(~0 + samps)

  sampsNormal sampsTumor
1           0          1
2           0          1
3           0          1
4           1          0
5           1          0
6           1          0
attr(,"assign")
[1] 1 1
attr(,"contrasts")
attr(,"contrasts")$samps
[1] "contr.treatment"
```

Different parameterization

- Parameterization:
  - Y = intercept + (tum-norm)(1_tumor)



```
> model.matrix(~samps)
  (Intercept) sampsTumor
1           1          1
2           1          1
3           1          1
4           1          0
5           1          0
6           1          0
attr(,"assign")
[1] 0 1
attr(,"contrasts")
attr(,"contrasts")$samps
[1] "contr.treatment"
```

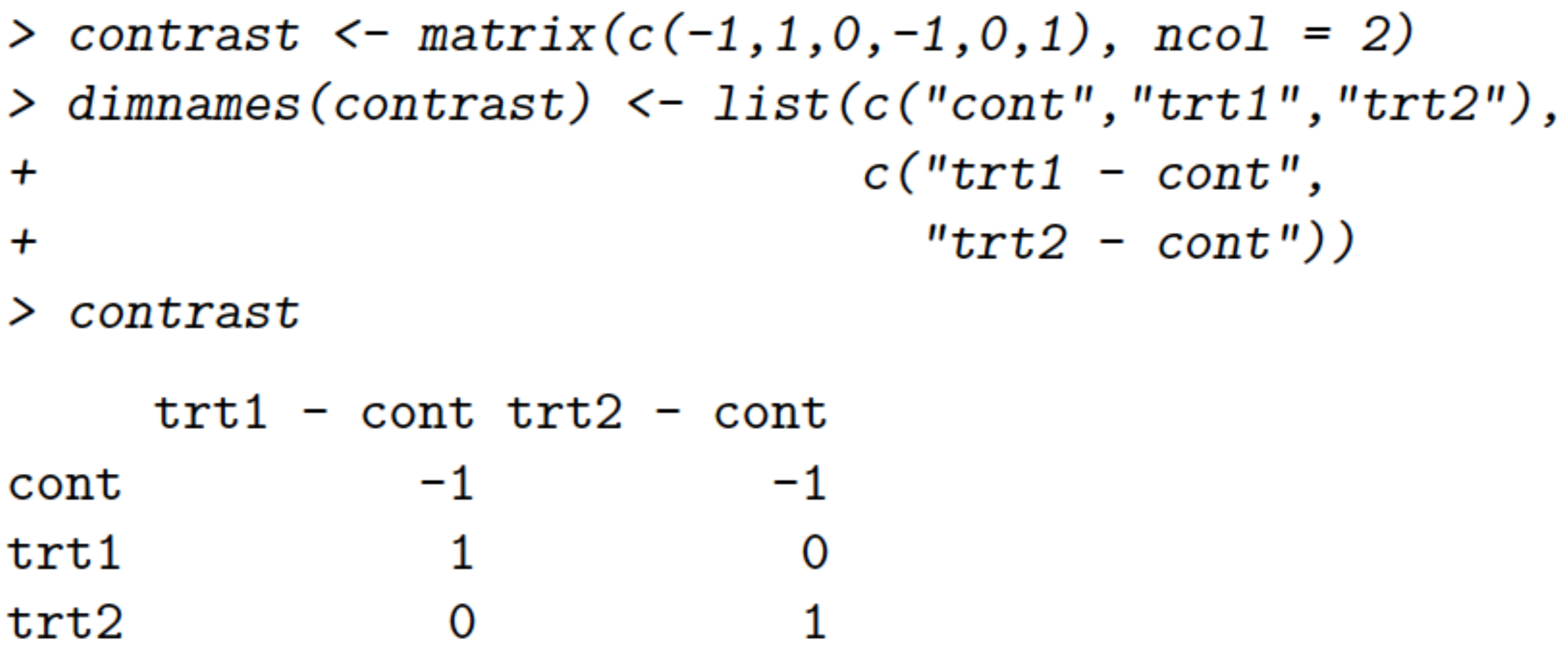
intercept included by default

Contrasts

- Linear combination of parameters
- Coefficients *sum to zero*
- Allows for *comparison* of different treatments
- Number of testable contrasts (rows in contrast matrix) equals number of parameters
- Need contrast matrix when comparison of interest is not a model parameter

Example: 3 groups

- Control/treatment 1/treatment 2
- Compare each treatment to control

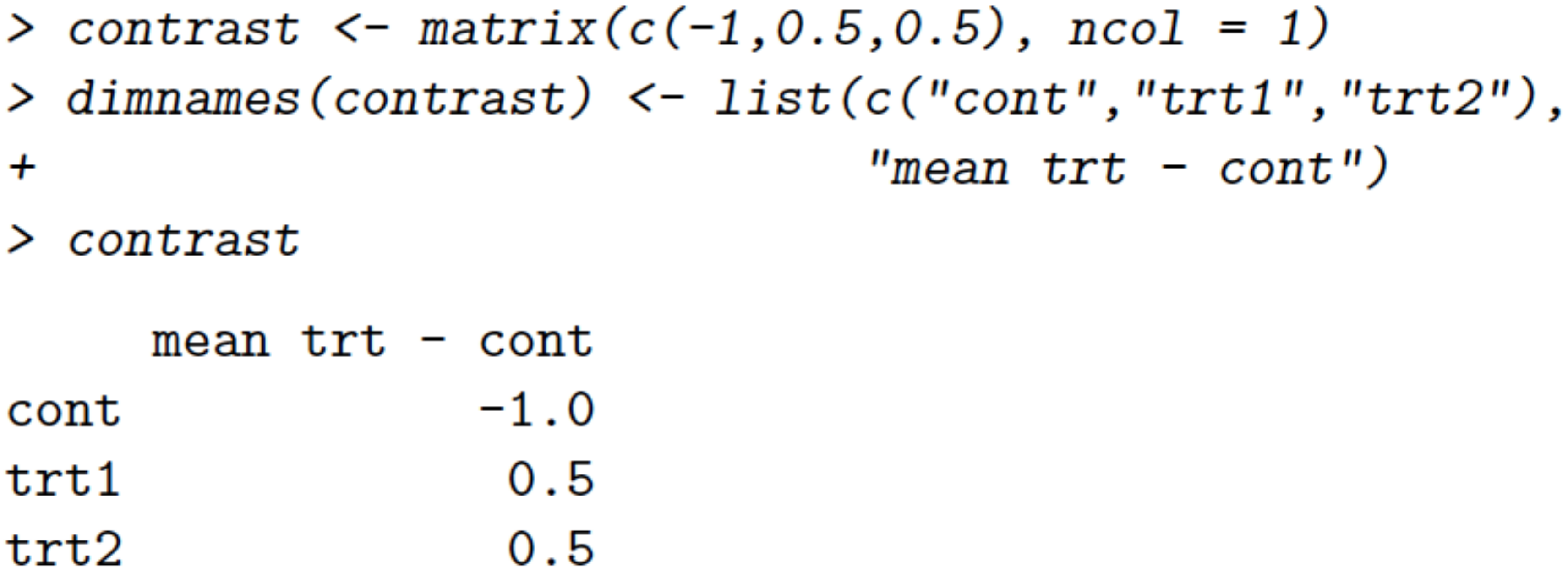


```
> contrast <- matrix(c(-1,1,0,-1,0,1), ncol = 2)
> dimnames(contrast) <- list(c("cont","trt1","trt2"),
+                            c("trt1 - cont",
+                              "trt2 - cont"))
> contrast

     trt1 - cont trt2 - cont
cont          -1          -1
trt1           1           0
trt2           0           1
```

Example: 3 groups

- Control/treatment 1/treatment 2
- Compare treatment mean to control



```
> contrast <- matrix(c(-1,0.5,0.5), ncol = 1)
> dimnames(contrast) <- list(c("cont","trt1","trt2"),
+                                "mean trt - cont")
> contrast

     mean trt - cont
cont            -1.0
trt1             0.5
trt2             0.5
```

Linear models for microarray data

- Specify linear model by design matrix
  - Rows correspond to arrays
  - Columns correspond to coefficient describing RNA sources
- Single channel (*e.g.* Affy chips) or common reference design: need one coefficient for each source type
- *Fit model* for each gene singly (**`lmFit`**)
- *Borrow information* across genes (**`eBayes`**)
- *DE genes* (**`topTable`**)