

Extending the Loop Design for Microarray Experiments

Naomi S. Altman,
Pennsylvania State University),
naomi@stat.psu.edu
Interface Meetings May 04

Expt Design and Microarrays

- Microarrays are
 - Expensive
 - Noisy
- A perfect situation for optimal design

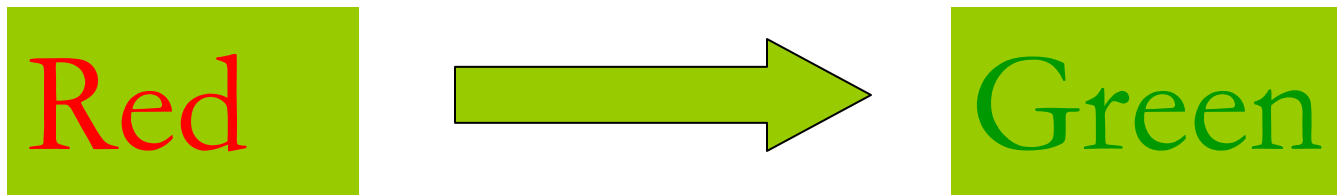
Outline

- Reference Design
- Loop Designs
- Replication
- Optimal Design/Analysis
- Incorporating Multiple Factors and Blocks

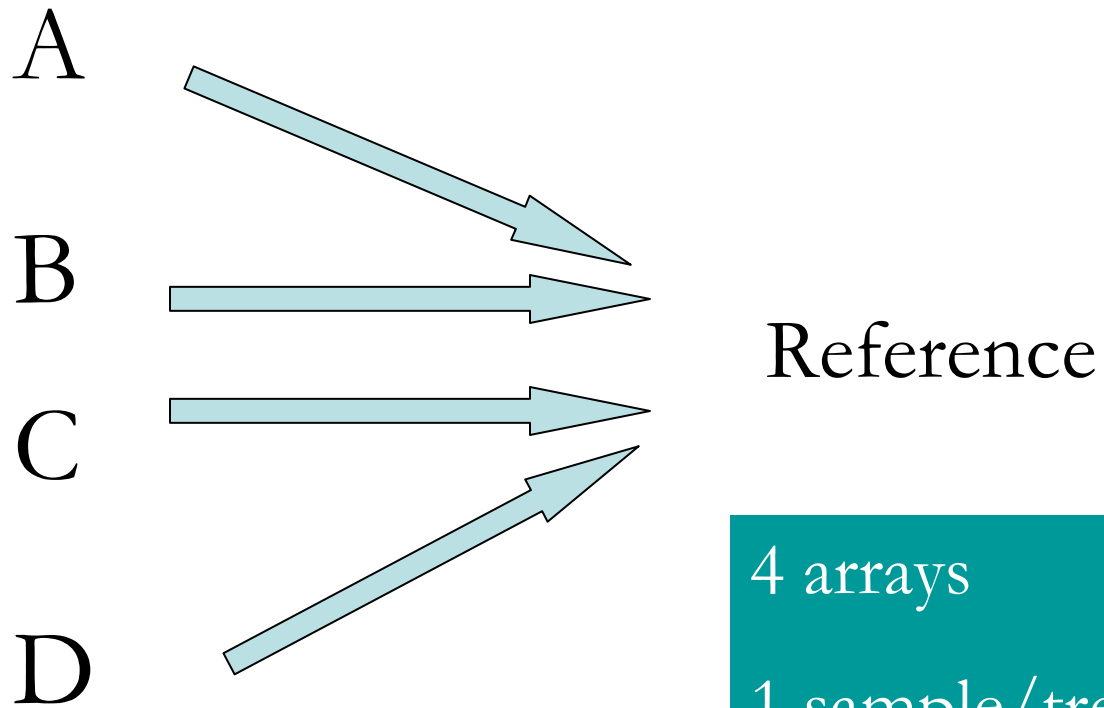
Arrow Notation

Introduced by Kerr and Churchill (2001)

Each array is represented by an arrow.



Reference Design



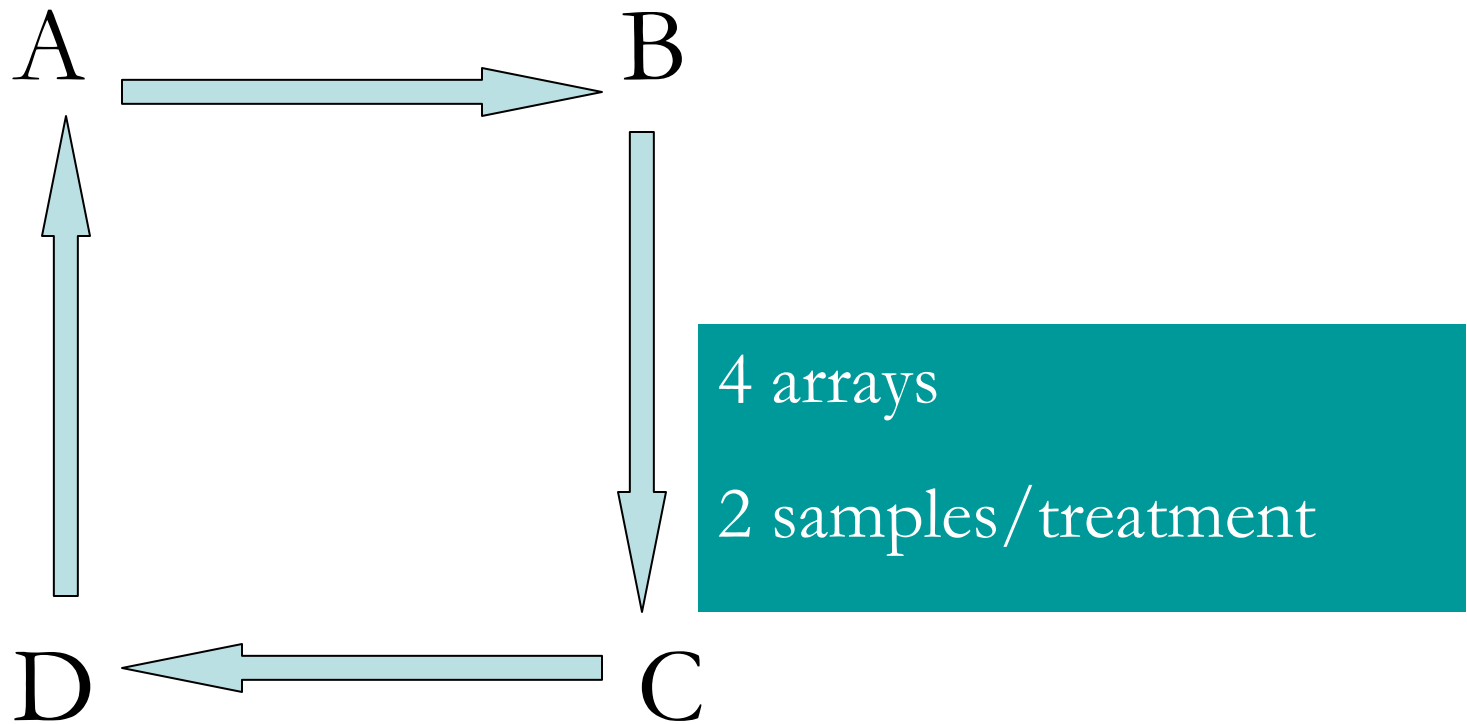
4 arrays

1 sample/treatment

4 reference samples

Loop Design

(Kerr and Churchill 2001)



Replication

Often there is confusion among:

Biological replicates

Technical replicates

- repeated samples

- split sample and relabel

- spot replication

In this presentation: We consider only

- one spot/gene/array

- any technical replicates are averaged

- each sample is an independent biological replicate

Linear Mixed Model for Microarray Data

$$Y_{ijk} = \mu + \tau_i + \delta_j + \alpha_k + \varepsilon_{ijk}$$

Y_{ijk}

- is the response of the gene in one channel

μ

- is the mean response of the gene over all treatments, channels, arrays

τ_i

- is the effect of treatment i

δ_j

- the effect of dye j

α_k

- is the effect of the array k (or spot on the array)

ε_{ijk}

- is the random deviation from the other effects and includes biological variation, technical variation and random error

Linear Mixed Model for Microarray Data

$$Y_{ijk} = \mu + \tau_i + \delta_j + \alpha_k + \varepsilon_{ijk}$$

The 2 channels on a single spot are correlated

→ array should be treated as a random effect

Differencing Channels on an Array

Often the difference between samples on a single array is the unit of analysis:

$$\Delta_{(ir).(kt)} = Y_{iRk} - Y_{rGk}$$

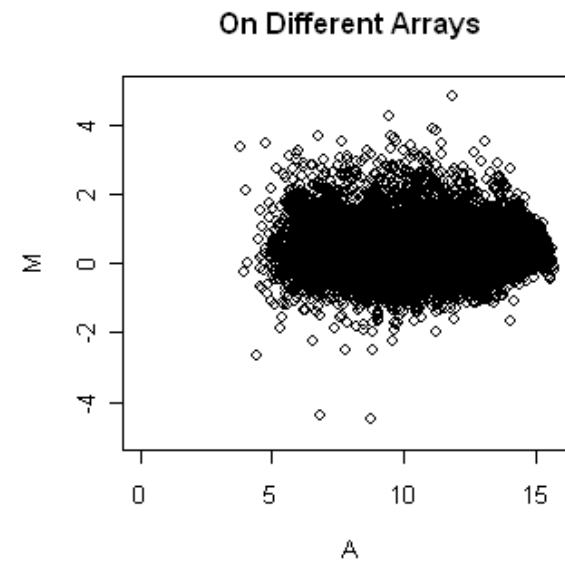
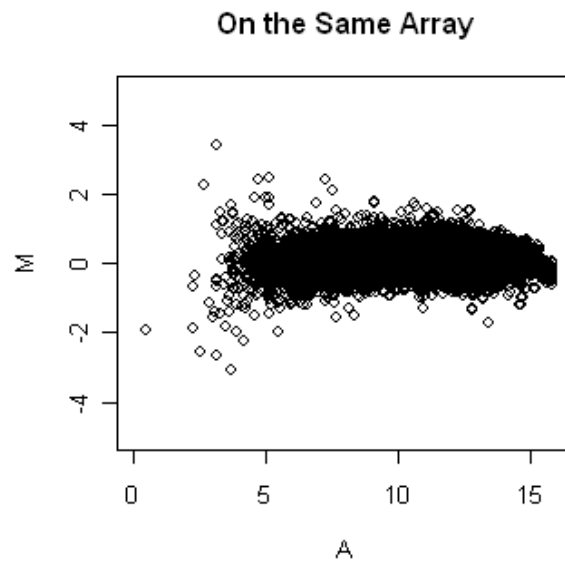
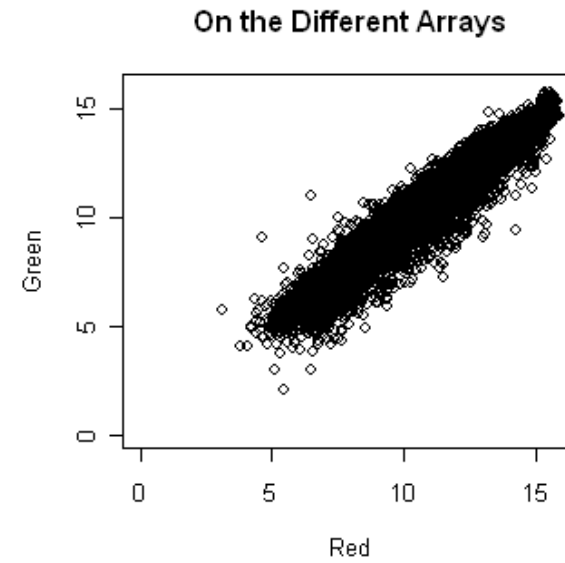
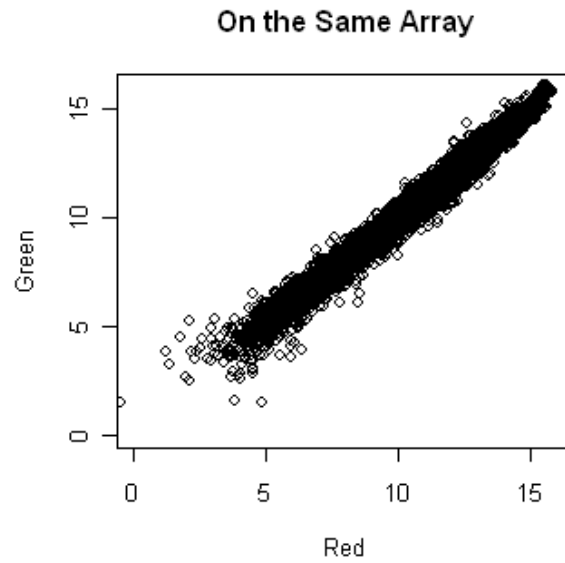
Normalization is almost always done on this quantity.

In a reference design, the difference between treatments A and B can be estimated from 2 arrays by

$$\hat{A} - \hat{B} = \Delta_{(Ar).(kt)} - \Delta_{(Br).(lu)}$$

But there can be a large loss of information.

$$\Delta_{(Ar),(kt)}$$



Drosophila arrays
courtesy of
Bryce MacIver,
PSU

Var(M)=0.126

Var(M)=0.453

Reference Design

The reference sample is the same biological material on every array

T treatments, **k replicates**, kT arrays

If there are technical dye-swaps, these are averaged to form 1 replicate.

If all comparisons are between treatments, there is no need to dye-swap. If there are dye-swaps, these should be balanced by treatment.

Reference Design – Usual Analysis

Usually the analysis is done on Δ .

E.g. $\hat{A} - \hat{B} = \bar{\Delta}_{(Ar).(\cdot)} - \bar{\Delta}_{(Br).(\cdot)}$

Using the linear mixed model, we see that the variance of one pair is

$$4\sigma_{\varepsilon}^2$$

and with k replicates, the variance of the estimated difference is

$$4\sigma_{\varepsilon}^2 / k$$

Reference Design – Optimal Weights

Consider using $\Delta^w_{(ir).(kt)} = Y_{iRk} - wY_{rGk}$

Then $\hat{A} - \hat{B} = \bar{\Delta}^w_{(Ar).(kt)} - \bar{\Delta}^w_{(Br).(lu)}$

The optimal w is $\sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$

The resulting variance for a single replicate is $4\sigma_\varepsilon^2 - 2\sigma_\varepsilon^4 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$

and with k replicates, the variance of the estimated difference is $4\sigma_\varepsilon^2 / k - 2\sigma_\varepsilon^4 / k(\sigma_\alpha^2 + \sigma_\varepsilon^2)$

$$w_{opt} = \sigma_{\alpha}^2 / (\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

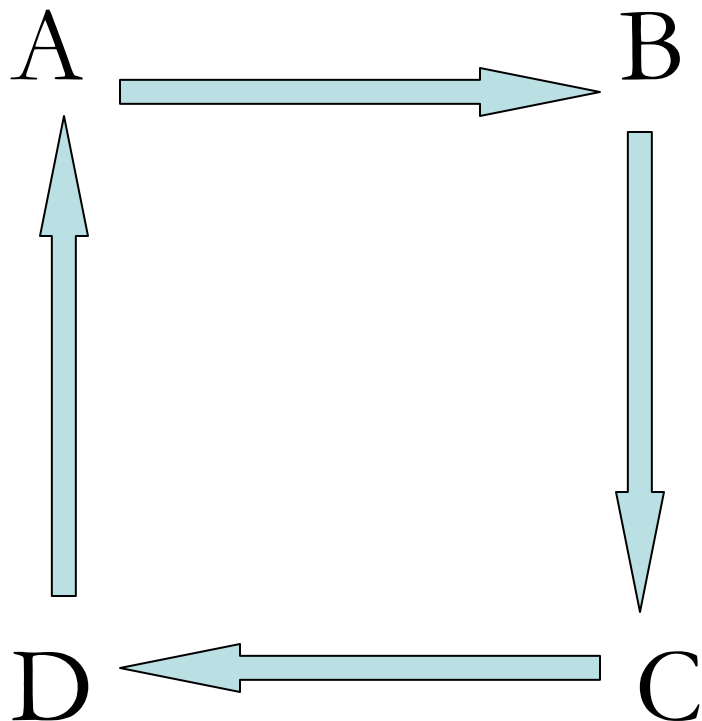
$$Var_{min} = 4\sigma_{\varepsilon}^2 - 2\sigma_{\varepsilon}^4 / (\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2) = 2\sigma_{\varepsilon}^2 + 2\sigma_{\alpha}^2\sigma_{\varepsilon}^2 / (\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

Reference Design – Optimal Weights

We do not know the optimal weights but

if we use mixed model ANOVA such as those available in SAS, Splus or R, the weights are approximated from the data – leading to more efficient computations.

Loop Designs



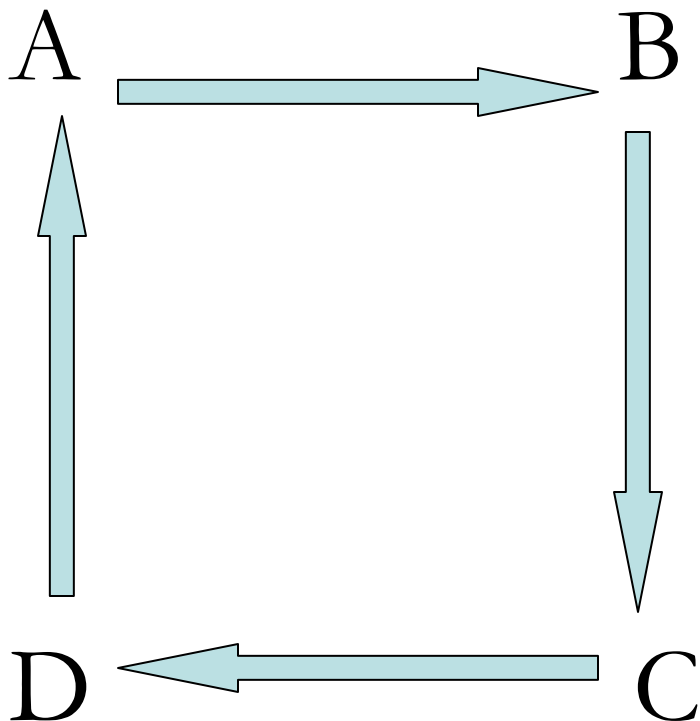
A loop is balanced for dye effects and has two replicates at each node.

T treatments, **2k replicates**,
Tk arrays

Recall: for a reference design we get only k replicates on Tk arrays

Loop Designs T=4, 4 arrays

Using optimal weighting



$$\text{Var}(A-B) = \text{Var}(A-D)$$

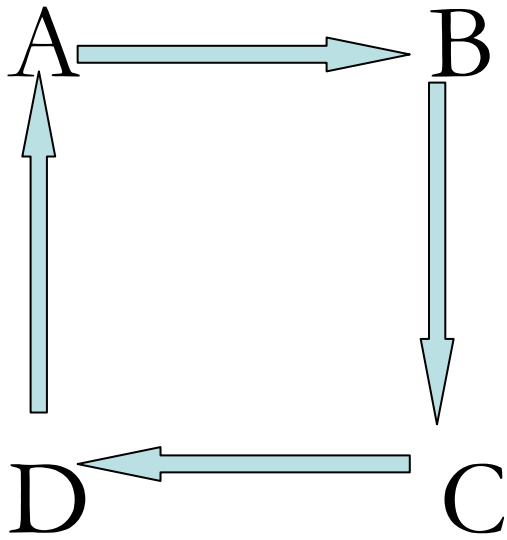
$$= \sigma_{\varepsilon}^2 + \sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / 2(\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

$$\text{Var}(A-C) = \sigma_{\varepsilon}^2 + \sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / (\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

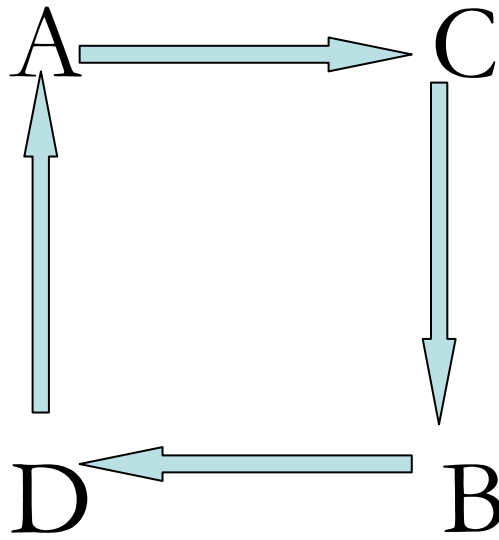
Both are smaller than the variance of the reference design with 4 arrays

$$2\sigma_{\varepsilon}^2 + 2\sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / (\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

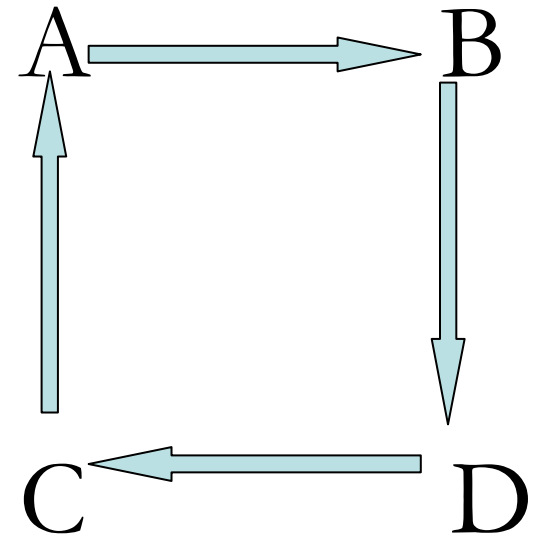
Loop Designs T=4



Design L4C



Design L4B



Design L4D

T=4, 12 arrays

Reference Design – 3 replicates/treatment

$$\text{Var}(\text{difference}) = 2\sigma_{\varepsilon}^2 / 3 + 2\sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / 3(\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

Loop Design – 3 loops = 6 replicates/treatments

3* L4C $\text{Var}(A-B) = \sigma_{\varepsilon}^2 / 3 + \sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / 6(\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$

$$\text{Var}(A-C) = \sigma_{\varepsilon}^2 / 3 + \sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / 3(\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2)$$

L4B+L4C+L4D

$$\text{Var}(\text{difference}) = \sigma_{\varepsilon}^2 / 3 + 2\sigma_{\alpha}^2 \sigma_{\varepsilon}^2 / 3(4\sigma_{\alpha}^2 + 3\sigma_{\varepsilon}^2)$$

T=4, 12 arrays

Assuming

Reference Design – 3 replicates/treatment

$$\text{Var}(\text{difference}) = 0.83$$

Loop Design – 3 loops = 6 replicates/treatments

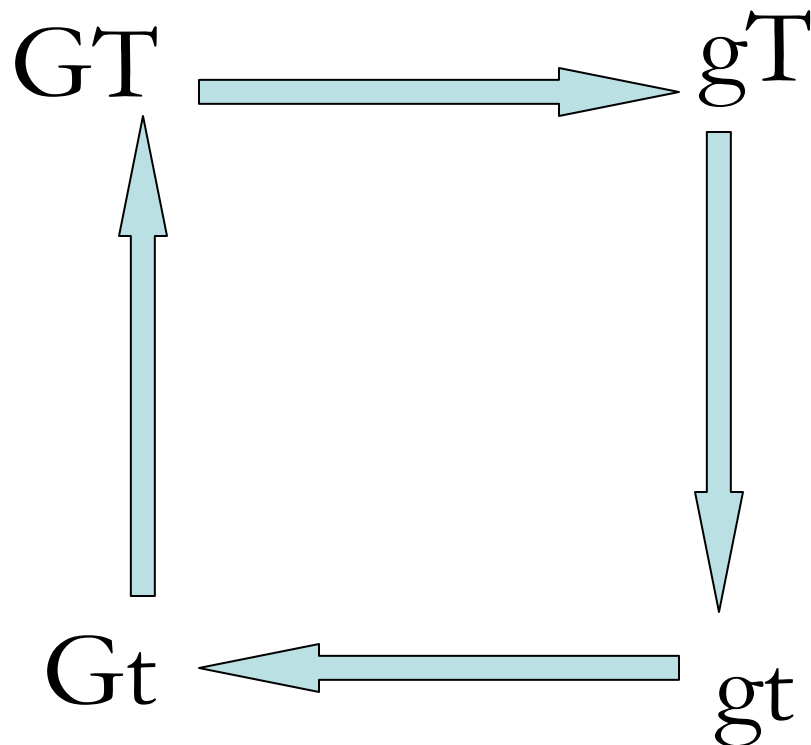
$$3^* \text{L4C} \quad \text{Var}(A-B) = 0.46$$

$$\text{Var}(A-C) = 0.58$$

L4B+L4C+L4D

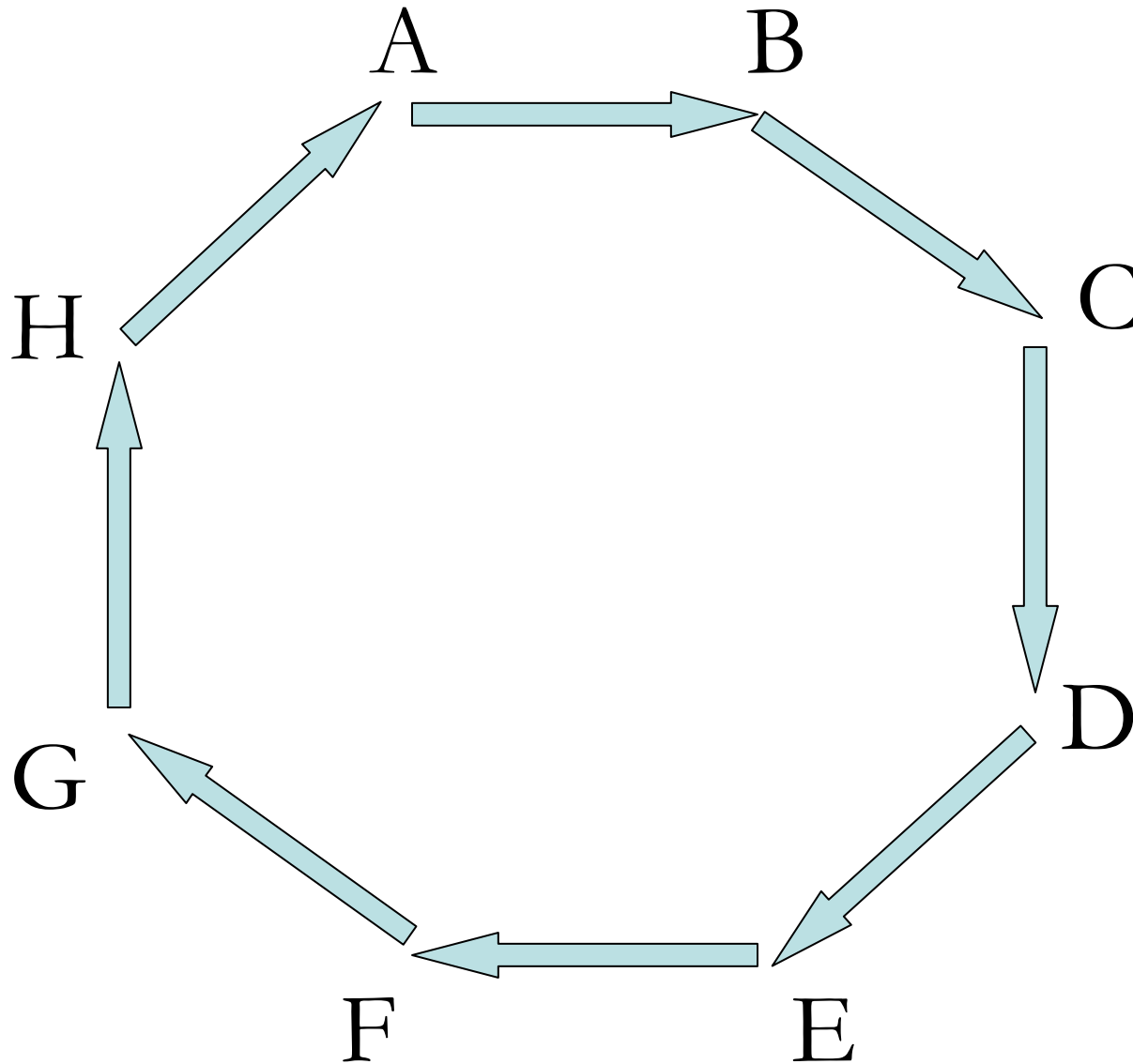
$$\text{Var}(\text{difference}) = 0.47$$

Incorporating 2x2 Factorial in a Loop

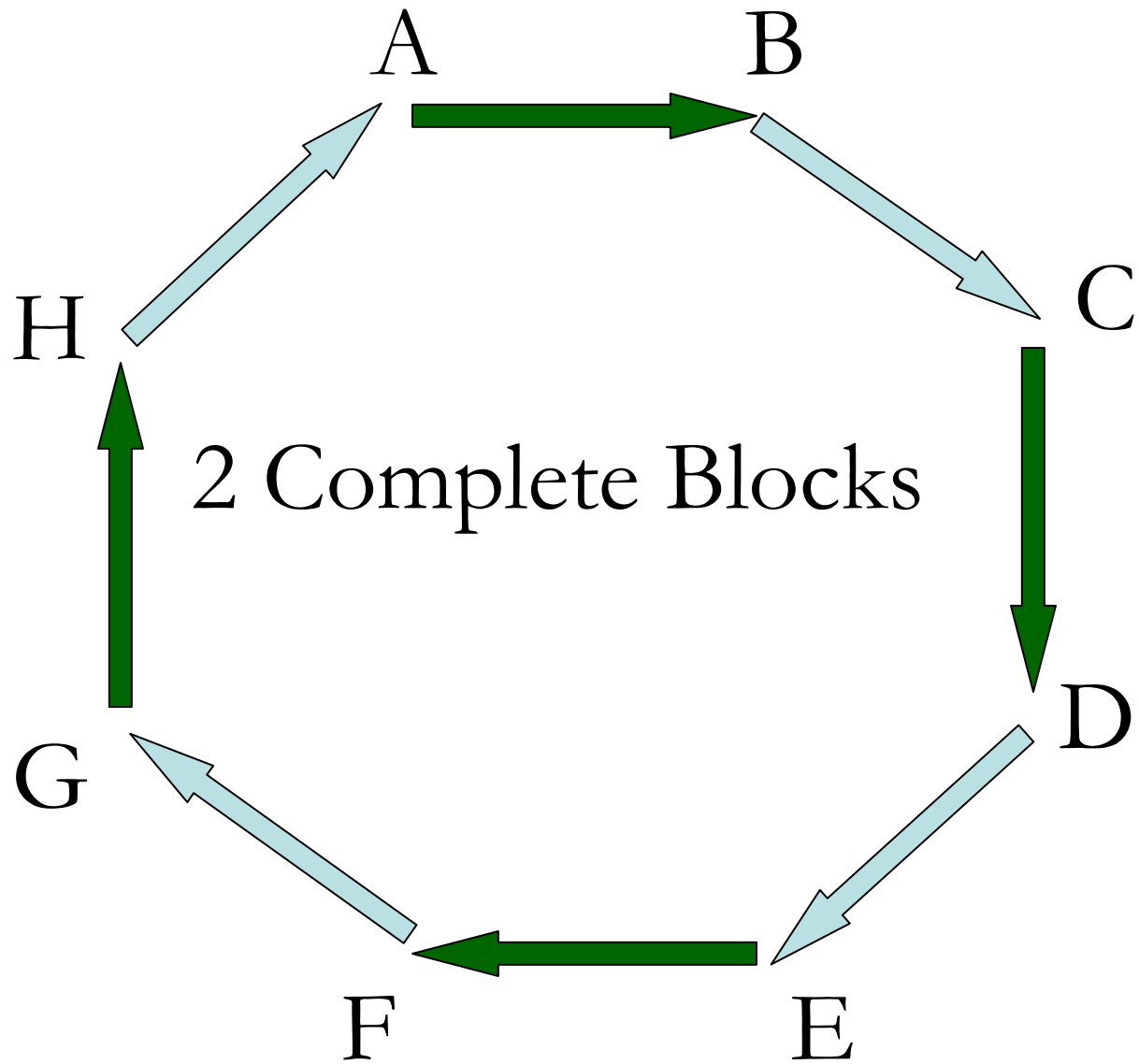


The design is 2
genotypes G, g and 2
tissues T, t
Only within genotype
and within tissue
comparisons are of
interest

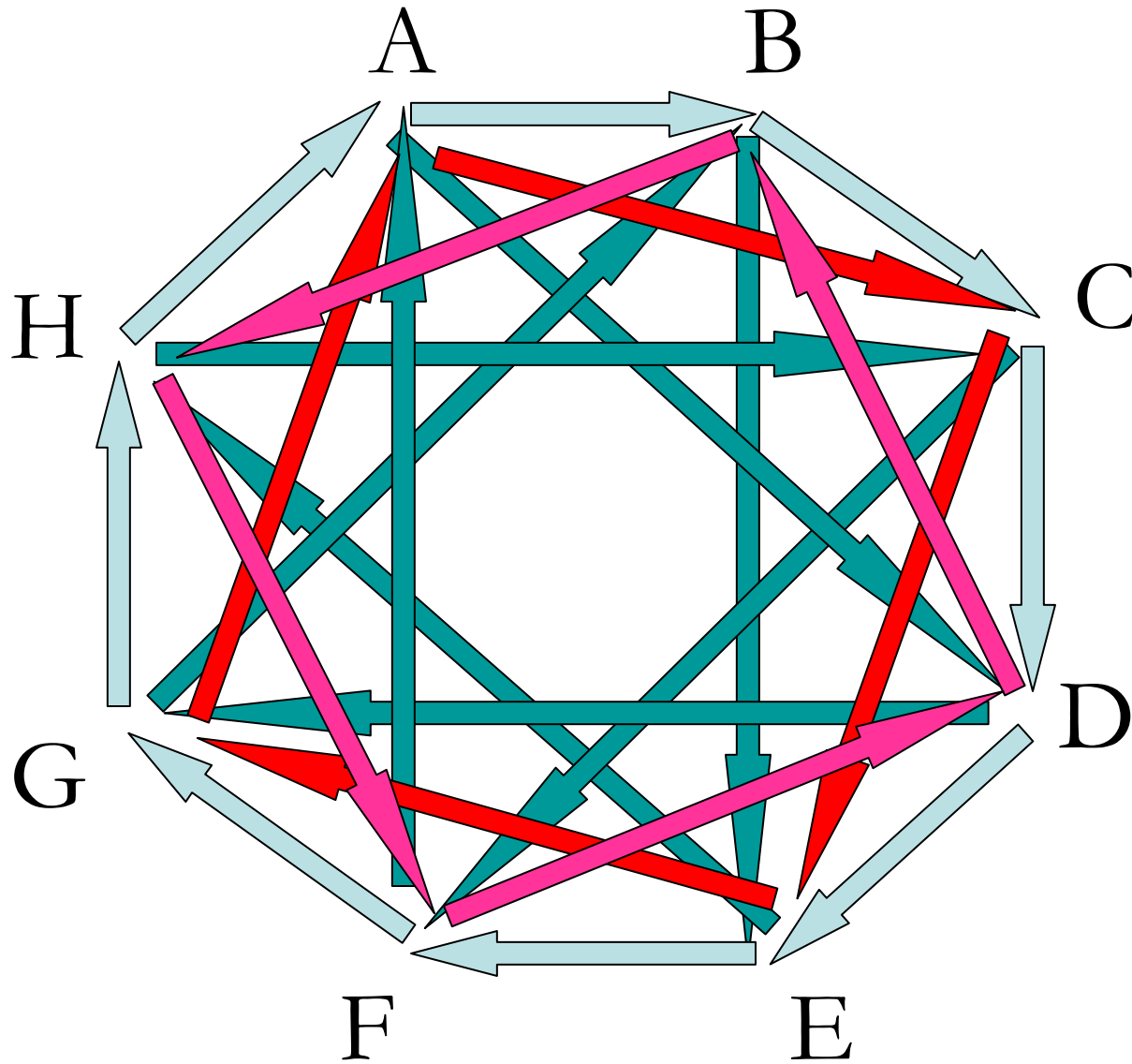
An 8 Treatment Example



An 8 Treatment Example



An 8 Treatment Example



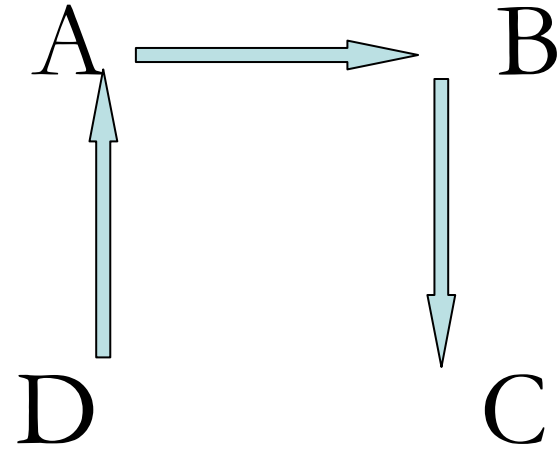
Replication:

Yellow loop?

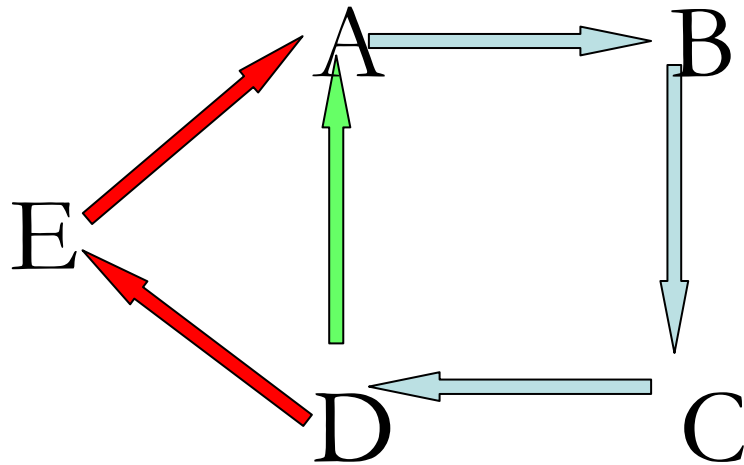
Red “loop”?

And now for the rest of the story

Missing arrays –
not fatal but
reduce efficiency

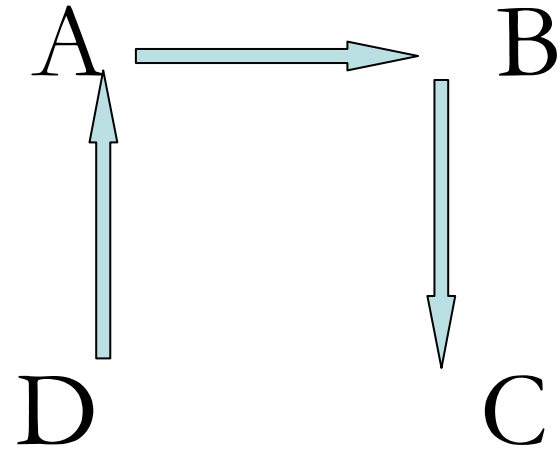


Added treatments

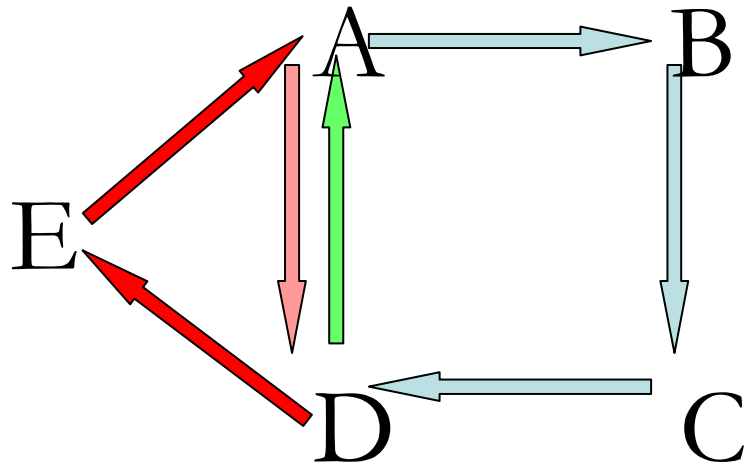


And now for the rest of the story

Missing arrays –
not fatal but
reduce efficiency



Added treatments



The Moral of the Story

- Loop designs are very efficient
 - Can incorporate factorial arrangements
 - Can incorporate blocks
 - Can be replicated in various ways to improve efficiency
- Optimal design can help determine which (generalized) loop design to use
- ANOVA-type analyses on the individual channels – not differencing

