Robust microarrary experiments by design: a multiphase framework

Chris Brien

Phenomics & Bioinformatics Research Centre,
University of South Australia

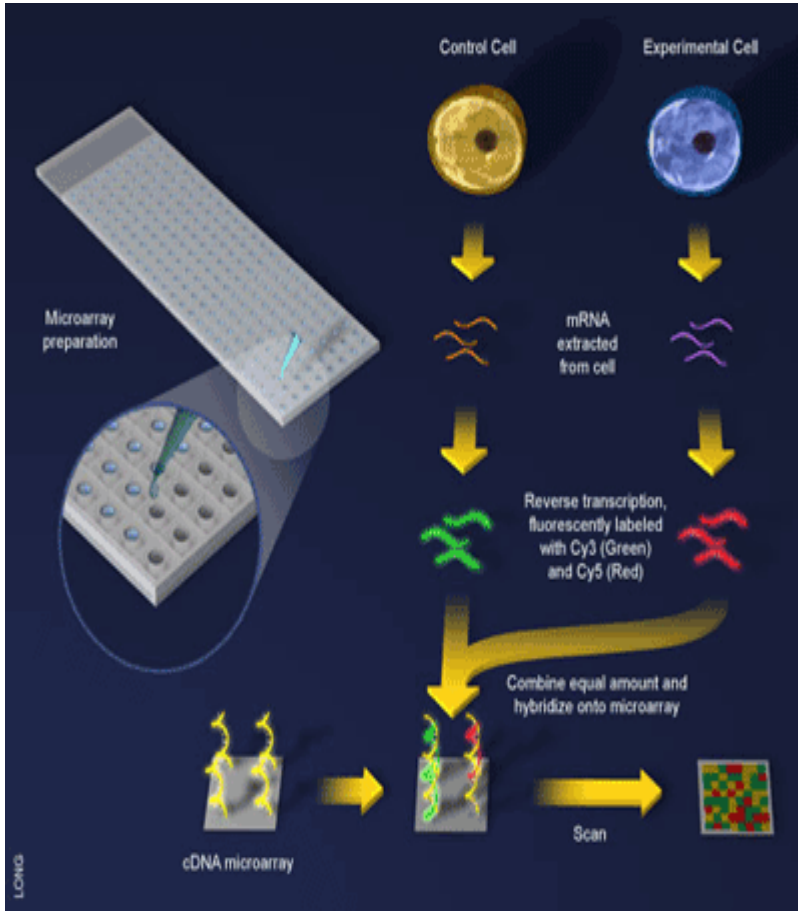
http://chris.brien.name/multitier

Chris.brien@unisa.edu.au

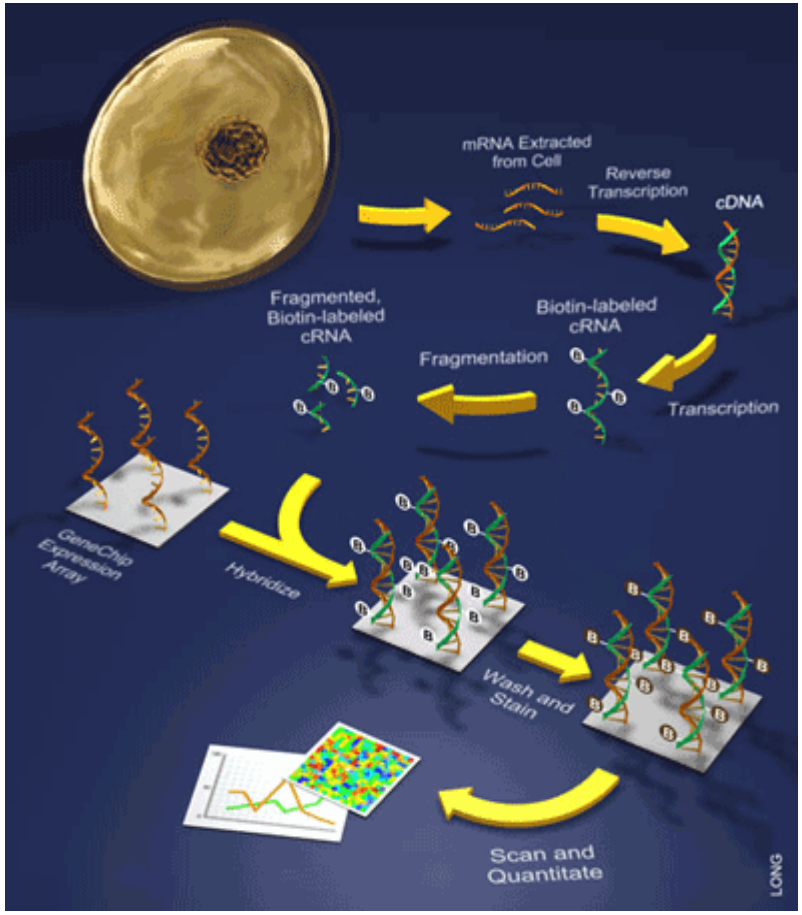
Outline

1. Phases in microarray technologies.
2. Designing and analysing multiphase microarray experiments:
   a) Avoiding bias;
   b) Minimizing variability.
3. Conclusions.

1. Phases in microarray technologies



Two-channel spotted



Single-channel oligo

- Similar multi-step processes.

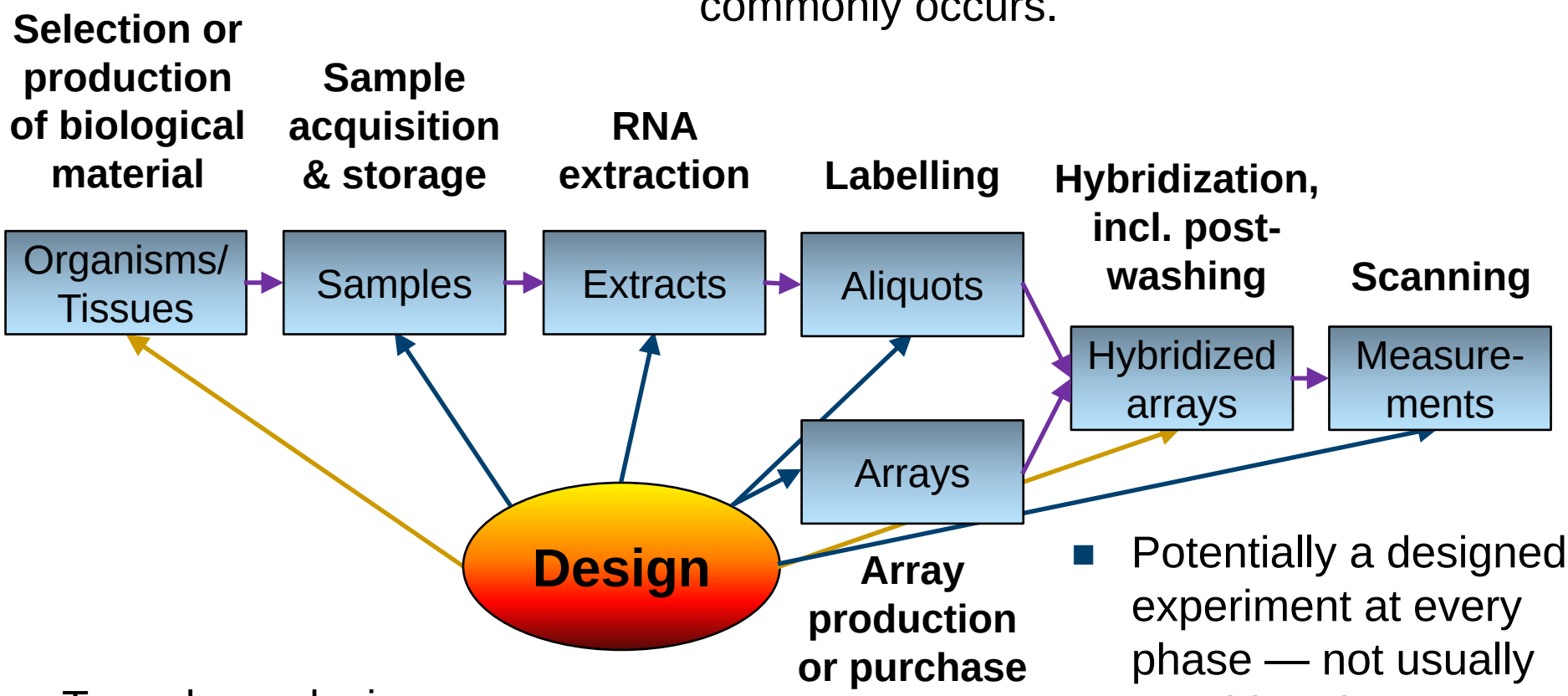
Figures by Jiang Long (from The Science Creative Quarterly, URL: http://www.scq.ubc.ca/image-bank/).3

Taking up Speed et al.'s (INI 2008) challenge: a multiphase framework

- The multiphase framework is based on Brien et al. (2011).
- We define a **phase** to be the period of time during which a set of units are engaged in producing a particular outcome.
  - The outcome can be material for processing in the next phase, or values for response variables, or both.
  - Only the final phase need have a response variable.
  - Also, one phase might overlap another phase.
- Then, **multiphase experiments** consist of two or more such phases.
- Generally, multiphase experiments randomize (randomly allocate) the outcomes from one phase to the next phase.
  - For microarray experiments, while the need for randomization is often mentioned in general terms, how to actually deploy it is little discussed.

Physical phases in a microarray experiment

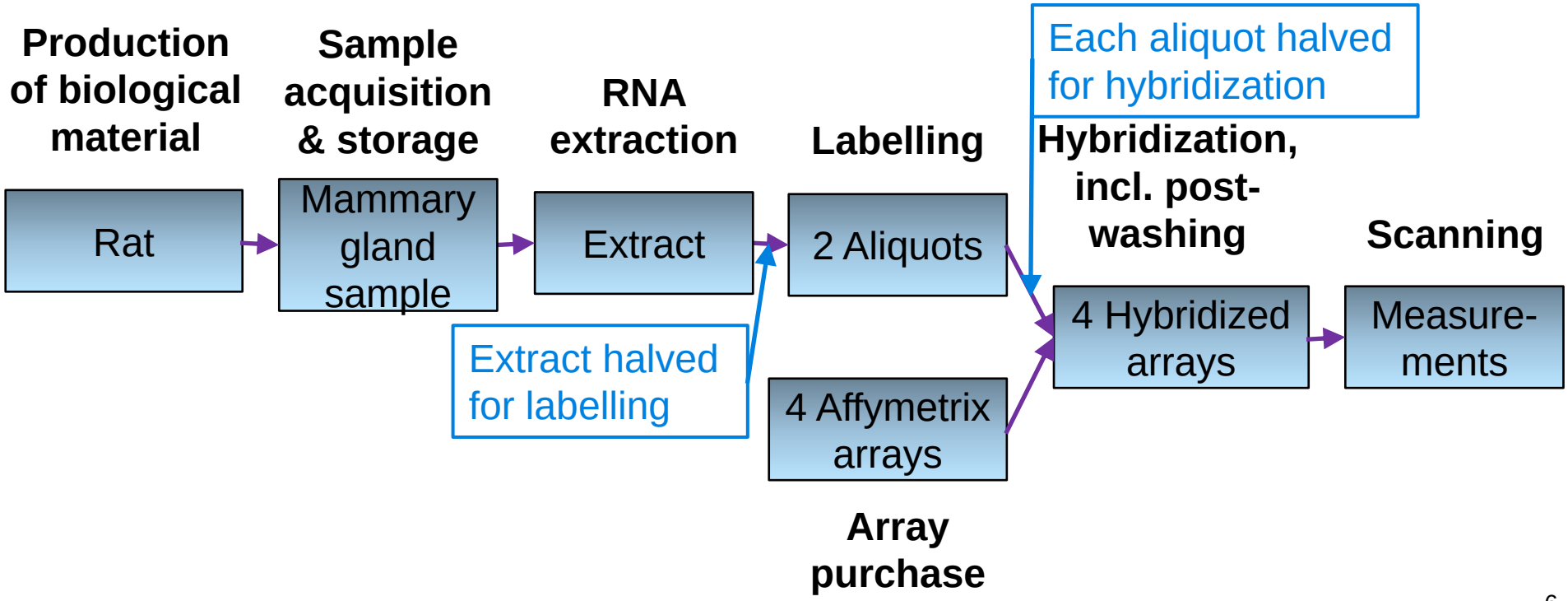
- One list of phases, and their outcomes, that commonly occurs.



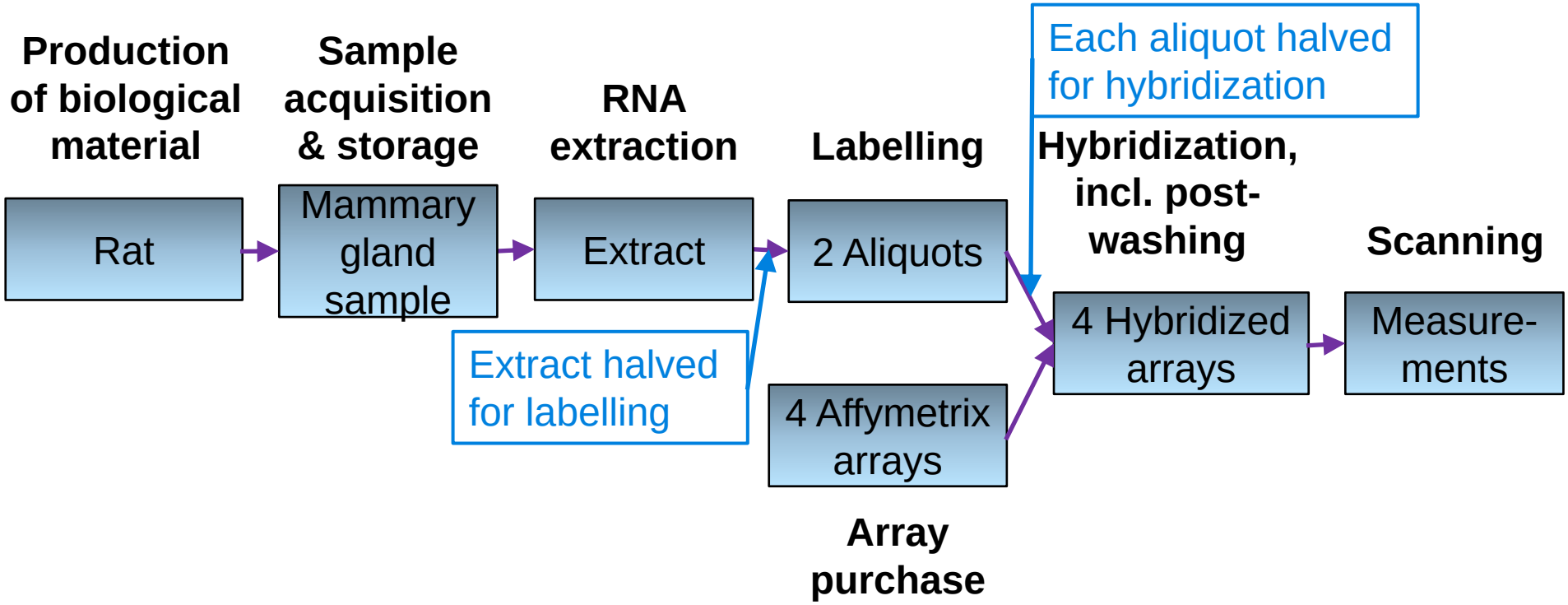
- Two-phase design (McIntyre, 1955; Kerr, 2003):
  - 1st phase can be an experiment or an observational (epidemiological) study;
  - 2nd phase is a laboratory phase (Brien et al., 2011).
- Potentially a designed experiment at every phase — not usually considered.
- Multiple randomizations (Brien & Bailey, 2006) in a seven-phase process.

A sources-of-variability Affymetrix microarray experiment

- A modified version of a study to examine the variability of labelling and hybridization described by Zakharkin et al. (2005):
  - It involved 8 rats;
  - The phases are:



Analysis of experiment



- Zakharkin et al. (2005) proposed a model like this:
  - $\mathbf{Y} = \mathbf{X}\mu + \mathbf{Z}_R\mathbf{u}_R + \mathbf{Z}_{L[R]}\mathbf{u}_{L[R]} + \mathbf{e}$ **OR**
  - Grand mean | Rats + Labellings[Rats] + Error,
    where terms to the left of '|' are fixed and those to the right are random.
  - The model is equivalent to that for an ANOVA for a doubly-nested study:
    2 Hybridizations within 2 Labellings within 8 Rats.

2. Designing and analysing multiphase microarray experiments:

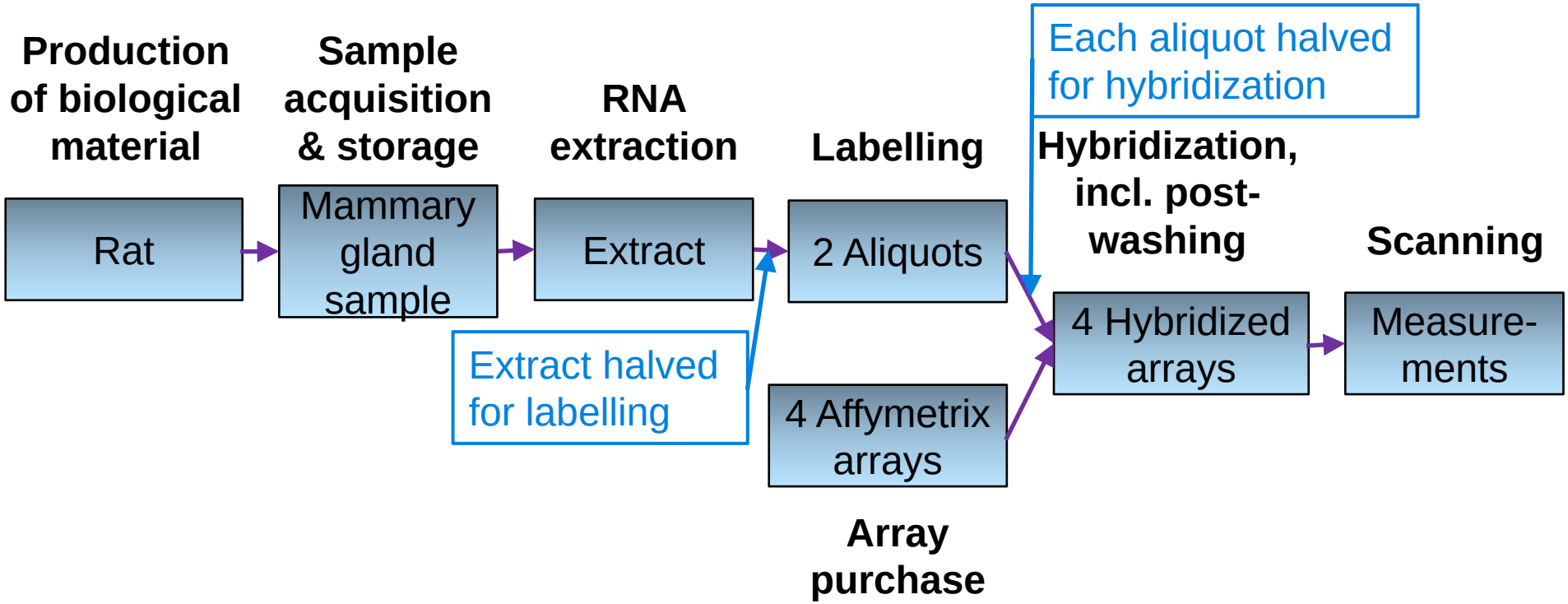
Approach to be taken:

I. Identify the set of phases involved.
II. Consider how the outcome of one phase is to be assigned to the units in the next phase.
III. Produce factor-allocation description (Brien et al., 2011).
IV. Formulate the full mixed model.
V. Derive the ANOVA table and use it to
   - investigate design; and
   - obtain a mixed model of convenience.

(a) avoiding bias

Illustrate using the example experiment

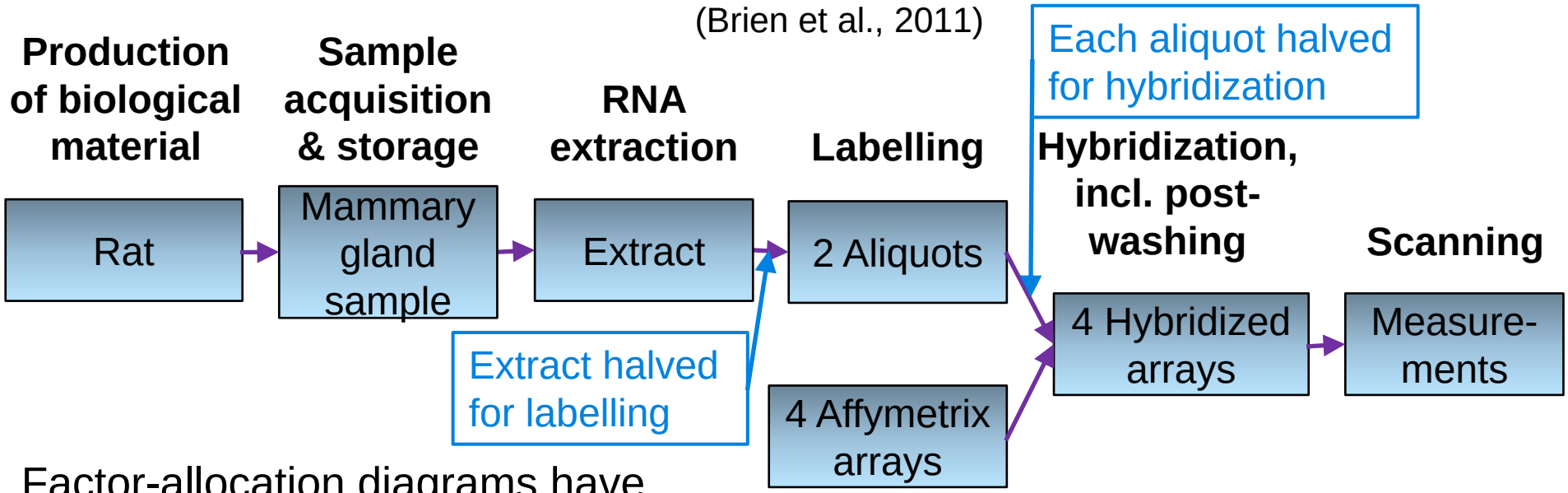
- We have the set of phases involved.



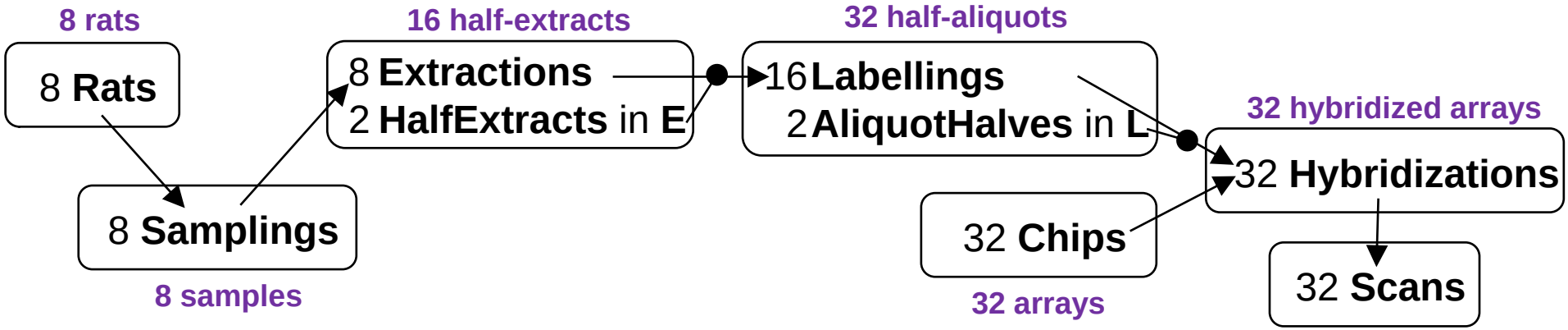
- But, how is the outcome of one phase to be assigned to the units in the next phase?
  - Suppose we take the simplest option: completely randomize every phase.

Factor-allocation description of experiment

(Brien et al., 2011)

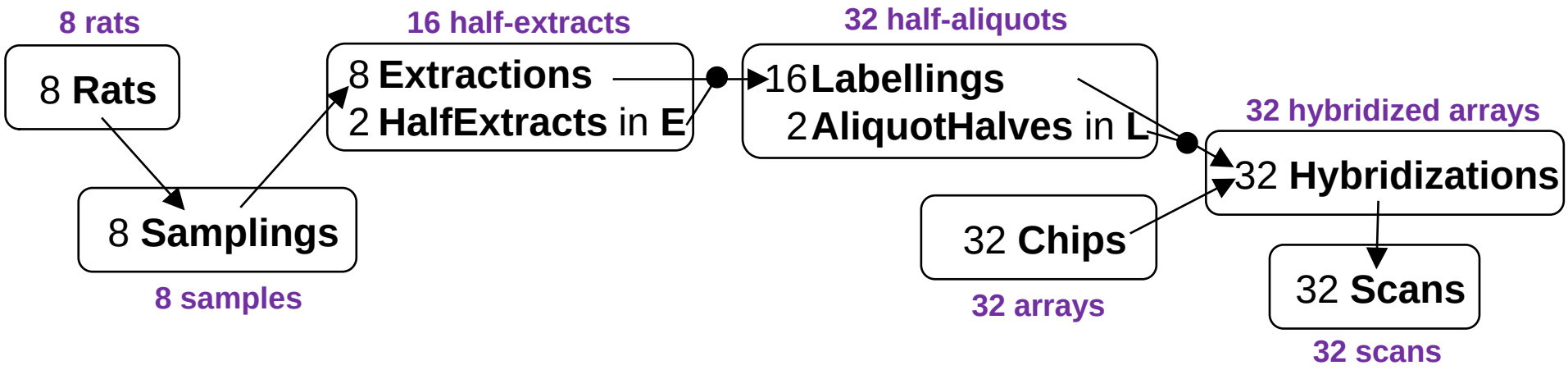


Factor-allocation diagrams have
a panel for a set of objects
(here the set of outcomes of a phase);
a panel lists the factors indexing a set.



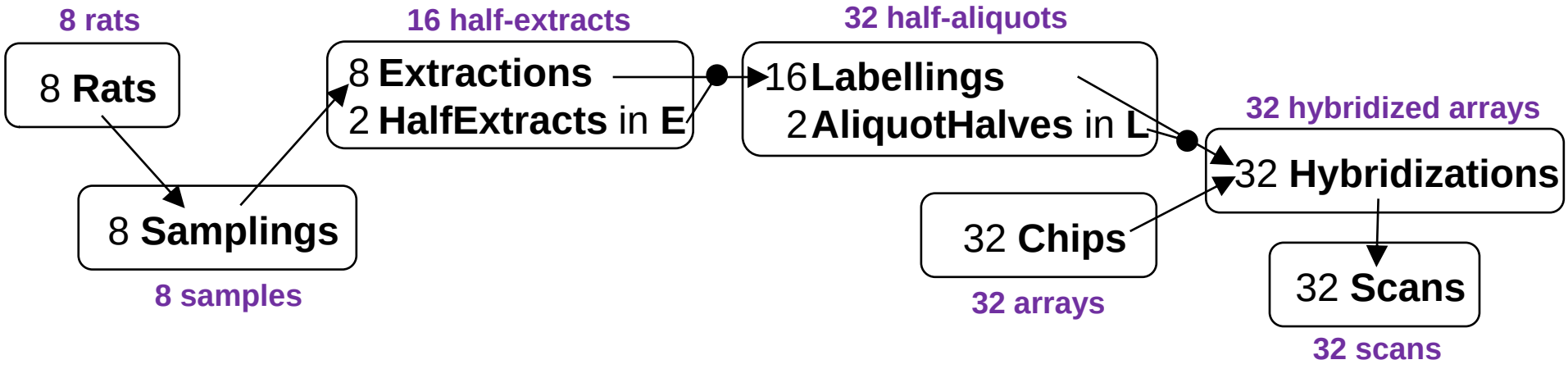
Arrow indicates randomization

Mixed model for the experiment



- To get mixed model use Brien & Demétrio's (2009) method:
  - In each panel, form terms as all combinations of the factors, subject to nesting restrictions;
  - For each term from each panel, add to either fixed or random model.
- Mixed model is:
  - Grand mean | Rats + Samplings +
    Extractions + Extractions∧HalfExtracts +
    Labellings + Labellings∧AliquotHalves +
    Chips + Hybridizations + Scans.
- Not all terms are estimable — use ANOVA to show this.

ANOVA for the experiment



scan, hybridized arrays & arrays tiers		half-aliquots tier		half-extracts tier		samples tier		rats tier	
source	df	source	df	source	df	source	df	source	df
Sc, H, C	31	Labellings	15	Extractions	7	Samplings	7	Rats	7
				HalfExtracts [E]	8				
		AliquotHalves[L]	16						

ANOVA for the experiment (cont'd)

scan, hybridized arrays & arrays tiers		**half-aliquots tier**		**half-extracts tier**		**samples tier**		**rats tier**	
source	df	source	df	source	df	source	df	source	df
Sc, H, C	31	Labellings	15	Extractions	7	Samplings	7	Rats	7
				HalfExtracts [E]	8				
		AliquotHalves[L]	16						

- Shows can measure variability from:
  - Rats + Samplings + Extractions (biological replication);
  - Extractions∧HalfExtracts + Labellings;
  - Labellings∧AliquotHalves + Chips + Hybridizations + Scans.
- Last referred to as 'Error' by Zakharkin et al. (2005) — prefer more specific description.
- The aim is a model in which all potential variability sources are identified, not one in which all are separately estimable.
- Mixed model of convenience for fitting:
  - Grand mean | Rats + Extractions∧HalfExtracts + Labellings∧AliquotHalves;
  - Model equivalent to that of Zakharkin et al. (2005).

2(b) Minimizing variability

- Desirable to minimize variability so that the variance of treatment estimates is as small as possible.
- As usual, three possibilities for this:
  i. Stringent experimental protocols to reduce variability;
  ii. Experimental design to avoid variability;
  iii. Statistical analysis to adjust for variability.

i) Stringent experimental protocols

- Several authors have investigated variability in the different phases and found some phases, e.g. hybridization phase, are more variable than others (Zakharkin et al., 2005; Tu et al., 2002; Spruill et al., 2002).
- Hence, as Han et al. (2006) advocate, minimizing variability in the hybridization phase will be important.

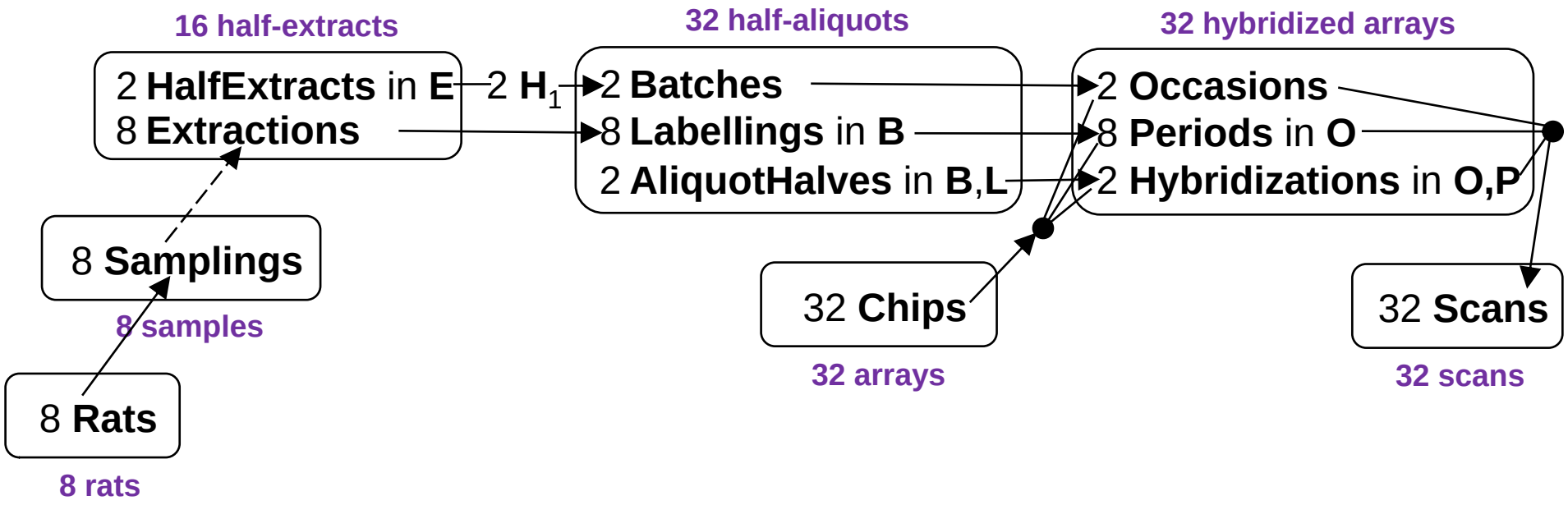
ii) Experimental design

- The need to take account of batches is often stressed.
  - Altman (2009, Table 4.1) lists sources of batch effects in all phases.
  - MAQC_Sample_Processing_Overview_SOP stipulated batching if cannot be done on one day.
- Generally, look for batches built into the processes:
  - e.g. different acquisition times, processing days, operators, batches of reagent or sets of simultaneously-processed specimens.
- When have treatments, will want to block them according to batches.

Design with batches for the sources-of-variability example experiment

- Suppose that:
  - For each rat, the **tissue** is to be obtained and the RNA **extracted** immediately; i.e. RNA-extraction order from samples is not random.
  - In the **labelling phase**, one half-extract from all extracts will be labelled in one batch, and the remaining ones in a second batch.
    - This separation of duplicates will yield a better estimate of the labelling variability; consecutively processed half-extracts are likely to be more similar than those more widely separated in time.
  - This batching will be carried through to the **hybridization phase**, where washing occurs in batches of 16. Further, for practical reasons, 2 half-aliquots from an aliquot are hybridized consecutively.
  - In the **scanning phase**, the arrays will be scanned in a completely random order.

Factor-allocation description



- Mixed model:
  - Grand mean |
    Rats + Samplings + Extractions + Extractions∧HalfExtracts
    + Batches + Batches∧Labellings + Batches∧Labellings∧AliquotHalves
    + Occasions + Occasions∧Periods + Occasions∧Periods∧Hybridizations
    + Chips + Scans.
- Above model will not fit, because of confounding, but this will:
  - Grand mean | Rats + Batches + Batches∧Labellings + Batches∧Labellings∧AliquotHalves.

ANOVA useful for choosing terms

3. Conclusions: Pros and cons of the approach

- Cons
  - Requires extra planning and work to organize — “I randomized the rats at the start. Isn’t this enough?”
  - A lot of needless redundancy — “Such a lot of factors!”
- Pros
  - Encourages consideration of appropriate design in all phases, even if ultimately it is decided not to randomize all phases.
  - As for all experiments, in a microarray experiment,
    - randomizing in a phase makes it robust to systematic biases in that phase.
      - Often processing order not considered, perhaps assuming no systematic change during a phase or processing a batch?
      - But is this tenable? Is it OK to process material from the same rat first in every phase, during operator or equipment warm-up in a phase? Randomization is insurance.
    - Blocking , based on batches, assists in minimizing variability.
  - Promotes the identification of all the sources of variability at play in a microarray experiment, even if not all are estimable.

Overall summary

- Microarray experiments are multiphase:
  - One might employ an experimental design in every phase to randomize and block the processing order in the current phase.
- Factor-allocation description can be used to formulate the analysis for an experiment, this analysis including terms and sources from every phase.
- The multiphase framework is flexible in that it can easily be adapted to another set of phases.

References

- Altman, N. (2009) Batches and blocks, sample pools and subsamples in the design and analysis of gene expression studies, in *Batch effects and noise in microarray experiments : sources and solutions,* A. Scherer, Editor, 34-50. Wiley-Blackwell: Oxford.
- Brien, C.J., and Bailey, R.A. (2006) Multiple randomizations (with discussion). *J. Roy. Statist. Soc., Ser. B*, **68**, 571–609.
- Brien, C.J. and Demétrio, C.G.B. (2009) Formulating mixed models for experiments, including longitudinal experiments. *J. Agr. Biol. Env. Stat.*, **14**, 253-80.
- Brien, C.J., Harch, B.D., Correll, R.L. and Bailey, R.A. (2011) Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *J. Agr. Biol. Env. Stat.*, available online.
- Han, Tao, et al. (2006) Improvement in the Reproducibility and Accuracy of DNAMicroarray Quantification by Optimizing Hybridization Conditions. *BMC Bioinformatics*, **7**, S17-S29.
- Kerr, M. K. (2003) Design Considerations for Efficient and Effective Microarray Studies. *Biometrics,* **59(4)**, 822-828.
- MAQC Consortium (2009) *MAQC Sample Processing Overview SOP,* U.S.F.a.D. Administration.
- McIntyre, G. A. (1955). Design and analysis of two phase experiments. *Biometrics*, **11**, 324-334.

References (cont'd)

- Novak, Jaroslav P., Sladek, Robert, and Hudson, Thomas J. (2002) Characterization of Variability in Large-Scale Gene Expression Data: Implications for Study Design. *Genomics,* **79***,* 104-113.
- Speed, T. P. and Yang, J. Y. H. with Smyth, G. (2008) Experimental design in genomics, proteomics and metabolomics: an overview. *Advanced Topics in Design of Experiments.* Workshop held at INI, Cambridge, U.K.
- Spruill, S. E., et al. (2002) Assessing sources of variability in microarray gene expression data. *Biotechniques,* **33(4)**, 916-20, 922-3.
- Tu, Y., Stolovitzky, G., and Klein, U. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, **99(22)**, 14031-14036.
- Zakharkin, Stanislav O., et al. (2005) Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*, **6**, 214-11.
- Web address for Multitiered experiments site:

http://chris.brien.name/multitier